Modélisation et optimisation numériques

Bijan Mohammadi - Jacques-Hervé Saiac

Faculté des Sciences de Montpellier
&
Conservatoire des Arts et Métiers

Table des matières

Avant-propos

Cet ouvrage est destiné aux étudiants, ingénieurs et chercheurs désireux d'avoir un aperçu des méthodes numériques utilisées aujourd'hui dans les domaines les plus variés ainsi que des recherches les plus récentes. Notre but est de donner une culture relativement large et une vue globale du métier de numéricien à travers la diversité des techniques de calcul et des domaines d'application.

L'objet des approches numériques est d'apporter une aide à la décision et à la conception des systèmes. Le calcul permet de confronter les théories à l'expérience, de passer des concepts à leur réalisation. Les approches numériques sont quasiment les seules à offrir la possibilité d'avoir une idée du fonctionnement d'un système fermé ou inaccessible tels que l'atmosphère d'une planète à des années lumières, l'intérieur d'un moteur à combustion ou le cœur d'une centrale nucléaire. Elles permettent, de plus, de simuler le comportement des systèmes virtuels, et c'est sans doute ce domaine qui ouvrira à la simulation une infinité de nouveaux débouchés, permettant l'emergence de disciplines inexistantes aujourd'hui. Ainsi, les approches numériques sont fortement créatrices de métiers. Pensons à l'explosion des télécommunications, aux jeux informatiques, aux films réalisés en numérique... Ces domaines introduisent à leur tour de nouveaux champs de recherche. La rapidité et la sécurité des transmissions indispensables dans les télécommunications ont motivé les progrès de la compression numérique des données et de la cryptographie.

Il existe un domaine propre au numéricien, au carrefour des mathématiques, de l'informatique, de la physique ou de l'économie. Sa spécificité consiste à assurer la traduction fidèle sur un ordinateur des modèles théoriques représentant la réalité. La difficulté provient de ce que la quantité d'information qui peut être stockée est finie, même si elle est très grande et croît sans cesse. Les conséquences dans le domaine numérique sont les erreurs d'arrondis, les défauts des représentations géométriques et la limitation de la taille des problèmes qu'il est possible de résoudre. D'autres types d'erreurs interviennent pendant la résolution, comme les erreurs liées à l'utilisation de méthodes itératives et dues au critère d'arrêt.

L'objectif sera donc de construire des schémas offrant la plus grande précision, la plus faible complexité (effort et mémoire nécessaire pour le calcul), les moins sensibles aux erreurs d'arrondis et aux déformations de maillage et utilisant, de surcroît, des structures de données simples pour leur mise en œuvre. D'un point de vue informatique, les schémas doivent être facilement adaptables aux différentes architectures d'ordinateurs (scalaire, vectorielle, parallèle gros et petit grain).

Il y a en général incompatibilité entre toutes ces exigences. Par exemple, un schéma précis mais très difficile à mettre en œuvre et sensible aux petites perturbations de maillage aura probablement peu de succès. De plus, on verra que la construction d'un modèle discret exige en pratique de nombreux outils et qu'en réalité une grande partie de la difficulté se trouve dans leur assemblage. A titre d'exemple, on peut citer trois niveaux pour la construction d'un maillage de calcul éléments finis :

1. définition d'un objet par un outil CAO (conception assistée par ordinateur) ;
2. passage de la paramétrisation CAO à la discrétisation ou maillage de surface ;
3. construction de maillage du domaine à partir du maillage de surface.

Ces trois niveaux doivent être correctement interfacés.

Aujourd'hui, le numérique doit faire partie du bagage de tout ingénieur et chercheur. La plupart ne développeront sans doute pas les outils logiciels de base, mais disposeront de plus en plus de techniques encapsulées dans des boîtes noires, qui exigent pour être bien utilisées, une initiation à leur principe. C'est un des objets de cet ouvrage.

Après un ensemble de chapitres introduisant les problèmes liés à la représentation des nombres, les techniques nouvelles de programmation et un exposé des méthodes numériques qui servent de base aux simulations dans tous les domaines d'application, nous présentons les points suivants :

— Les équations aux dérivées partielles modèles, à coefficients constants et variables, linéaires et non-linéaires. Les différents types de conditions aux limites : Dirichlet, Neumann, Fourier, mixte, conditions aux limites cachées. Extension aux conditions aux limites équivalentes et lois de paroi.

— Les méthodes de discrétisation : différences finies, volumes finis, éléments finis.

— Les problèmes d'évolution et leur discrétisation en temps. Stabilité. Relation entre discrétisations spatiale et temporelle.

— Les problèmes d'optimisation non-linéaire. Lagrangien et points-selles. Dualité. Un nouvel algorithme d'optimisation efficace dans le cas de fonctionnelles non-convexes présentant plusieurs minima locaux.

— Le traitement de la géométrie. Résolution à maillage donné. Génération, raffinement et adaptation de maillage. Analyse d'erreur.

— La résolution des EDP stochastiques par des approches statistiques ou bien de type Monte Carlo. Utilisation du filtrage et obtention d'équations pour les quantités moyennes.

Les méthodes présentées permettent d'envisager la simulation d'applications très variées :

— simulation d'écoulements de fluides incompressible et compressible (autour d'une voiture et un avion), prise en compte de la turbulence ;

— simulation du comportement des structures (calcul de crash, ouvrages d'art) ;

— simulation de comportement thermique (chauffage d'une pièce) ;

— simulation de systèmes régis par l'équation des ondes (radar, antenne, micro-onde), onde monochromatique (Helmholtz) ;

— simulation de certains modèles en mathématiques financières utilisés pour la prédiction des prix des options ;

— couplage de plusieurs de ces modèles. On présente quelques exemples de systèmes couplés : couplage fluide-structure, couplage champ électrique-fluide - particules ainsi que les algorithmes de couplage adaptés ;

— simulation pour l'optimisation. Un objectif important de la simulation est la conception et l'optimisation de systèmes dans le but de réduire le nombre de prototypes. De la même manière, on peut s'intéresser à la résolution de problèmes inverses pour l'assimilation des paramètres des modèles physiques. Ceci est un retour de la simulation sur la modélisation. En pratique, on arrive facilement à formuler l'objectif de l'optimisation à travers une expression dans le langage courant : Avoir un avion moins bruyant, le plus portant et opposant le moins de résistance possible à l'air,

avoir un moteur plus propre, un placement plus rentable, une structure plus légère mais solide...

On constate que les critères sont souvent multiples et incompatibles. De plus, la traduction de ces expressions intuitives en problèmes mathématiques bien-posés est souvent difficile. Ceci exige pour le numéricien de passer par une étape inévitable de modélisation, et souvent d'y revenir car les résultats des modèles discrétisés sont rarement satisfaisants du premier coup.

Ainsi, nous verrons comment reformuler certaines des expressions précédentes en termes de problèmes d'optimisation et nous présenterons les outils mathématiques permettant leur résolution. Comme pour la simulation, de nombreux outils supplémentaires seront nécessaires pour la mise en place d'une chaîne d'optimisation et de résolution de problèmes inverses.

Cet ouvrage est organisé de manière à aborder de façon progressive l'ensemble des méthodes et des domaines d'applications. Il commence par les notions de base les plus facilement accessibles, puis sont introduites, en avançant au cours des chapitres, les techniques plus complexes, les recherches récentes et les nouveaux champs d'applications. Les notions mathématiques de base nécessaires sont données en annexes.

Chapitre 1

Représentation discrète et éléments d'algorithmique

1.1 Introduction

Les mathématiques utilisent couramment les notions d'infini et de continu. La solution exacte d'un problème d'équations différentielles ou aux dérivées partielles est définie sur un domaine continu. Les ordinateurs ne connaissent que le fini et le discret. Les solutions approchées seront calculées en définitive comme des collections de valeurs discrètes sous la forme de composantes d'un vecteur solution.

1.2 Taille des problèmes et stockage mémoire

En simulation numérique, mettant en jeu des modèles mathématiques discrets, la taille du stockage mémoire est gouvernée par deux critères :

- La précision des résultats qui conduit à des exigences sur la finesse de la discrétisation utilisée pour le calcul. En particulier, il est nécessaire, afin d'assurer une bonne précision de la solution approchée, de raffiner la discrétisation dans les zones de fort gradient.
- La nécessité d'une bonne représentation du domaine physique dans le cas de problèmes extérieurs : écoulements autour d'obstacles, propagation d'ondes, etc, problèmes qui, théoriquement, se posent en milieu infini et qui doivent évidemment être réduits dans des domaines de taille finie lorsqu'ils sont modélisés sur un ordinateur.

Disons, pour simplifier, que l'ordre de grandeur du nombre de nœuds d'une discrétisation, qui va conditionner le nombre d'inconnues et donc la taille des systèmes à résoudre évolue comme une puissance de la dimension. Par exemple,

si une équation demande une centaine de points pour une résolution en dimension un, elle demandera environ un million de points en dimension trois.

1.3 Mémoires d'ordinateur

On distingue, sur un ordinateur, plusieurs types de mémoire, selon leur technologie et la rapidité d'accès à un élément stocké. Comme tous les matériels informatiques, ces mémoires évoluent avec la miniaturisation de la technologie.

- Le premier niveau est la mémoire cache, la plus rapide d'accès, mais aussi la plus intégrée. L'ordre de grandeur aujourd'hui (en 2002) pour un PC est de l'ordre du mégaoctet, avec des temps d'accès de l'ordre de la nanoseconde.
- La mémoire vive (ou la RAM) caractérise la taille maximum d'un exécutable statique pour l'ordinateur en question. La taille de ces mémoires pour un PC est aujourd'hui de l'ordre du gigaoctet, sachant que pour une machine à 32 bits, on ne pourra pas, pour des raisons d'adressage, dépasser les 2 gigaoctets. Les temps d'accès sont inférieurs à 100 nanosecondes.
- Les mémoires disques ont une capacité quasiment illimitée (plusieurs dizaines de gigaoctets) mais avec des temps d'accès plus longs : de l'ordre de la milliseconde.

  Lors de l'écriture du programme il faut donc éviter au maximum la lecture et l'écriture sur le disque, ainsi que les"défauts de cache". Ce dernier point est particulièrement important. Il faut exploiter au maximum une information se trouvant dans la mémoire cache. Considérons, par exemple, la boucle ci-dessous :

```
real x(n), u(n)
do i=1,n
u(i)=h(x(i))
u(i)=g(u(i),x(i))
u(i)=f(u(i),x(i))
end do
```

  Ici à chaque ligne, on fait un défaut de cache (on interrompt les calculs pour aller chercher des informations ou en écrire ailleurs). Le programme ci-dessous est bien plus efficace :

```
do i=1,n
! lecture et stockage dans une variable locale et temporaire
xtemp=x(i)
```

```
utemp=h(xtemp)
utemp=g(utemp,xtemp)
utemp=f(utemp,xtemp)

! sauvegarde globale uniquement en final
u(i)=utemp
end do
```

Le gain sera d'autant plus important que le nombre d'opérations par indice de boucle est grand. En résumé, quand une information est lue, il faut l'utiliser au maximum. En optimisant l'utilisation du cache on peut gagner un facteur 20 lors de la programmation des algorithmes numériques.

1.4 Vitesse de calcul

Les performances en temps de calcul se mesurent d'une part par la vitesse d'horloge, actuellement de l'ordre du gigahertz (on prévoit environ un facteur deux tous les deux ans pour l'évolution de la puissance des processeurs), et plus précisément en nombre d'opérations flottantes par secondes. Par exemple une résolution de système linéaire plein de n équations à n inconnues, prend, par la méthode du pivot de Gauss, de l'ordre de $\frac{n^3}{3}$ opérations. La résolution d'un système 1000×1000 prendra donc environ $1/3$ secondes sur une machine à 1 gigaflops, en supposant que chaque opération compte pour 1 flop (floating point operation). Actuellement, compte-tenu de la difficulté à comparer les performances de machines de technologies différentes, on préfère mesurer les vitesses de calcul sur des batteries de tests normalisés.

1.5 Parallélisme et scalabilité

En ce qui concerne le calcul parallèle (voir aussi le chapitre 20), un critère important qui permet d'optimiser la rapidité globale des calculs est la "scalabilité". La scalabilité caractérise la qualité de la parallélisation d'un algorithme. Un défaut de scalabilité mesure la perte de l'efficacité due aux communications et opérations redondantes. L'efficacité théorique suit une courbe linéaire car en doublant le nombre de processeurs, on voudrait aller deux fois plus vite. Alors que les communications et opérations redondantes font plutôt apparaître une progression logarithmique.

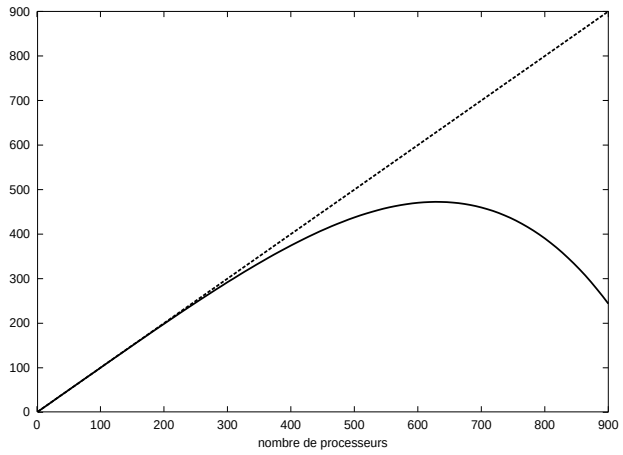


FIGURE 1.1 – Accélération théorique et observée : à taille de problèmes constante, l'augmentation du nombre de processeurs permet une accélération des calculs, puis les communications et opérations redondantes devenant prohibitives, les performances se dégradent.

1.6 Stratégies de calcul

Les caractéristiques techniques des matériels conditionnent les stratégies de calcul. Par exemple, il est souvent impossible de stocker en mémoire vive les systèmes d'équations provenant de problèmes industriels réels tridimensionnels. Et ceci, même en appliquant des techniques de numérotation des inconnues et de stockage optimales. Pour ces gros cas de calcul, on devra choisir des méthodes de résolution itératives qui évitent le stockage des matrices et n'exigent que celui de vecteurs de même taille que le vecteur des inconnues. Par contre, ces méthodes ne convergent pas forcément rapidement et peuvent donc être plus gourmandes en temps de calcul. Il y a souvent opposition entre gain en stockage mémoire et gain en temps de calcul. Il faudra donc choisir un compromis en fonction de son matériel et du problème à résoudre.

1.7 Représentation des nombres dans un ordinateur

Les calculateurs électroniques utilisent la numérotation binaire. L'unité d'information est le bit (binary digit en anglais). En numérotation binaire un nombre entier X s'écrit uniquement avec les chiffres 0 et 1. On obtient sa représentation

en le développant en puissances de 2. Ainsi :

$$X = \sum_{i=0}^{p} a_i \, 2^i \quad \text{s'écrit} \quad X = a_p a_{p-1} ... a_0$$

Donc avec n digits binaires, on peut représenter tous les nombres entiers de 0000...0 à 1111...1, soit 2^n nombres compris entre 0 et $2^n - 1$. Par exemple avec 1 octet, c'est-à-dire 8 bits, on dispose de 256 possibilités de représentation (on peut représenter tous les nombres compris entre 0 et 255).

Nous renvoyons aux ouvrages d'informatique générale pour tout ce qui concerne les codages classiques de l'information. Nous voudrions par contre entrer dans le détail pour ce qui concerne la représentation des nombres dans les unités de calcul qui est fondamentale pour le numéricien. C'est elle qui explique les erreurs d'arrondis.

On distingue sur un ordinateur les "entiers", les "réels" ou nombres flottants et dans certains langages les "réels" ou nombres flottants en double précision. Le problème pour l'utilisateur provient de ce que ces dénominations sont trompeuses car les catégories entiers, réels, en informatique, ne correspondent pas aux définitions mathématiques.

1.7.1 Les entiers

Ils sont représentés par une suite de bits organisés en octets (8 bits). Par exemple un entier à 2 octets occupera 16 bits ($2^{16} = 65536$). Un entier 2 - octets machine correspond donc à un entier mathématique compris entre - 32 768 et 32 767.

Le type entier 4 - octets ($2^{32} = 4294967296$) permet la représentation des valeurs entières comprises entre - 2 147 483 648 et 2 147 483 647.

Les opérations sur les entiers, dans la mesure où le résultat est un entier représentable en machine, s'effectuent exactement.

1.7.2 Les réels ou nombres flottants

Chiffres significatifs

C'est la notion importante qui justifie l'écriture en virgule flottante. Le but de cette représentation est d'obtenir, pour un encombrement mémoire donné, un éventail de valeurs suffisant avec la plus grande précision possible. Il s'agit de ne pas perdre de place mémoire en caractères inutiles.

Pour plus de clarté, nous nous placerons dans le système de numérotation décimale. Supposons que nous disposions de 10 caractères dont la virgule et le signe. Dans une représentation à virgule fixe, nous ne pourrons écrire que les décimaux du type suivant :

$$\pm 123,45678$$

On voit donc que l'éventail des nombres possibles est très limité (comme pour les entiers). Le plus petit nombre, en valeur absolue, représentable étant :

$$\pm,00000001$$

Sur ce dernier exemple, 7 caractères ont été occupés par la représentation de zéros. Le seul chiffre significatif est le 1.

Nombres flottants

Un nombre flottant s'écrit sous la forme suivante :

$$X = \pm a.b^n \tag{1.1}$$

a est la mantisse, b la base, n l'exposant.

Plaçons-nous à nouveau en base 10 et supposons que nous disposions de 10 caractères. Répartissons-les comme suit :

- 1 pour le signe du nombre,
- 3 pour l'exposant dont un pour son signe,
- les 6 caractères restants seront affectés à la mantisse.

On obtient ainsi des représentations du type suivant :

$$\pm,123456\,10^{\pm 78}$$

L'adoption d'une mantisse normalisée telle que le premier caractère à droite de la virgule soit différent de zéro assure un nombre maximal de chiffres significatifs et donc une meilleure précision. D'autre part il devient inutile de stocker la virgule, si l'on convient que, par définition, la mantisse représentée par un nombre entier naturel à 6 chiffres sera un nombre décimal compris entre $0,100000$ et $0,999999$. Il est également inutile de stocker la base qui sera une constante pour tous les nombres flottants.

Dans cette représentation, on dispose d'un éventail beaucoup plus large entre $\pm 0,100000\,10^{-99}$ et $\pm 0,999999\,10^{+99}$.

Exemple : Supposons que l'on cherche à représenter π avec la plus grande précision possible. Dans ce système à mantisse normalisée on obtient :

$$+0,314159\,10^{+1}$$

Considérons à présent la représentation de $\frac{\pi}{10000}$ avec une mantisse normalisée, nous obtenons : $+0,314159\,10^{-3}$ et donc le même nombre (6) de chiffres significatifs que pour p et une précision qui passe à 10^{-9}. (Si l'on n'avait pas imposé un chiffre non nul comme premier chiffre après la virgule, on aurait obtenu 0,000314101 avec une précision de 10^{-5} seulement).

1.7.3 La représentation standard

La brève introduction précédente permet de comprendre le choix technologique fait pour représenter les réels en machine. Nous reprenons la forme

$$X = \pm a.\beta^n$$

avec a : mantisse, β : base, n : exposant, β est égal à 2, a et n sont représentés par des binaires

Voici le format standard d'un nombre réel en simple précision :

Ce nombre occupe 4 octets.

- son exposant est stocké sur un octet
- son signe sur 1 bit
- sa mantisse occupe les 23 bits restants.

Donc l'exposant prend toutes les valeurs entières entre - 128 et + 127 . La mantisse est représentée par $t = 23$ caractères binaires selon l'écriture : $d_1 d_2 ... d_t$ avec $d_1 = 1$.

Elle correspond au nombre : $X = \frac{d_1}{2} + \frac{d_2}{2^2} + \frac{d_3}{2^3} + ... + \frac{d_t}{2^{23}}$.

Conséquences de ce choix de représentation

1) Le plus petit nombre en valeur absolue ou zéro machine est le nombre d'exposant minimal L et de mantisse 100....0.
Le plus petit nombre obtenu est ainsi : $X = \frac{1}{\beta}\,\beta^L \;=\; \beta^{L-1}$
Dans notre exemple $L = -128$, $b = 2$ Le zéro machine vaut : $2^{-129} \simeq 1,47\,10^{-39}$.

2) Le plus grand nombre en valeur absolue ou infini machine est le nombre d'exposant maximal U et de mantisse 111.....1. La mantisse vaut donc

$$a = \frac{1}{\beta} + \frac{1}{\beta^2} + \frac{1}{\beta^3} + ...\frac{1}{\beta^t} = 1 - \beta^{-t}$$

avec $\beta = 2$ et $U = 127$, on obtient un infini machine de : $(1-2^{-23})2^{127} = 1,7\ 10^{38}$

3) L'écart entre deux nombres successifs d'exposant n (ainsi que l'écart entre le plus grand nombre d'exposant n et le plus petit nombre d'exposant $n+1$) vaut :

$$2^{-t}.2^n$$

Cet écart croît exponentiellement avec n. Ceci explique pourquoi la précision en valeur absolue des calculs flottants est d'autant moins bonne que les nombres sont grands.
Ainsi, la meilleure précision possible pour des calculs sur des nombres de l'ordre de 1, correspondant à des nombres d'exposant $n = 0$ sera :

$$2^{-t} = 2^{-23} \simeq 1,19\ 10^{-7}$$

Pour des nombres de l'ordre de 1000, correspondant à un exposant $n = 10$ ($210 = 1024$), la meilleure précision possible tombe à

$$2^{-23}2^{10} = 2^{13} \simeq 1,22\ 10^{-4}$$

Un grand nombre d'erreurs d'arrondis catastrophiques s'explique ainsi : on cherche à obtenir, lors d'un calcul sur ordinateur, un nombre petit par différence de nombres grands. Ceci est une des motivations de la technique d'adimensionnement que nous présentons chapitre 3.

1.8 Erreurs d'arrondis

Le fait d'utiliser un nombre limité de bits pour représenter un nombre réel a donc comme conséquence la propagation des erreurs d'arrondis. Un certain nombre de calculs sont particulièrement sensibles aux erreurs d'arrondis, c'est le cas de l'addition (et la soustraction), contrairement à la multiplication où les erreurs d'arrondis sont moindres. Un numéricien choisira les méthodes numériques ayant une moindre sensibilité aux erreurs numériques.

1.8.1 Erreurs d'arrondis par multiplication

Considérons la multiplication de deux réels f et g écrits selon la représentation ci-dessus :

$$fg = (x.2^n)(y.2^m) \qquad \frac{1}{2} \leq x \leq 1 \qquad \frac{1}{2} \leq y \leq 1$$

Elle se décompose selon les étapes suivantes :

1. Addition des exposants n+m,
2. Multiplication des mantisses,

3. Normalisation de la mantisse, (Le produit fg étant dans l'intervalle $[\frac{1}{4}, 1]$ on normalisera en se ramenant à l'intervalle $[\frac{1}{2}, 1]$ en multipliant par 2 et en réduisant l'exposant de 1)
4. Troncature de la mantisse.

L'erreur relative de cette opération est évaluée en considérant de petites variations pour les variables et en prenant en compte l'erreur commise lors de la troncature de la mantisse ($\delta f \sim \delta g \sim \epsilon \sim 10^{-8}$) :

$$(f + \delta f)(g + \delta g) + \epsilon|fg| = fg + f\delta g + g\delta f + \epsilon|fg|$$

Ainsi, l'erreur relative (en divisant par fg) est :

$$E_* = \frac{\delta g}{g} + \frac{\delta f}{f} \pm \epsilon \sim \epsilon$$

On constate que les trois sources d'erreurs s'ajoutent et ont le même ordre de grandeur.

1.8.2 Erreurs d'arrondis par addition

Considérons maintenant l'addition de deux réels :

$$f + g = (x.2^n) + (y.2^m) = (x + y.2^{m-n})2^n \qquad \frac{1}{2} \leq x \leq 1 \qquad \frac{1}{2} \leq y \leq 1$$

qui se décompose selon :

1. Décalage de la mantisse pour avoir le même exposant,
2. Addition des mantisses,
3. Normalisation de la mantisse comme pour la multiplication,
4. Troncature de la mantisse.

L'erreur relative est calculée de la même manière que précédemment :

$$(f + \delta f) + (g + \delta g) + \epsilon(|f + g|) = (f + g) + (\delta g + \delta f) + \epsilon(|f + g|)$$

$$E_+ = \frac{\delta f + \delta g + \epsilon(|f + g|)}{f + g} >> \epsilon \text{ si } |f + g| << 1$$

Cette fois, si $f + g << 1$ l'écart peut être non négligeable. Ceci est une cause majeure des erreurs numériques. En voici quelques exemples.

Un premier exemple d'erreurs d'arrondis

Essayons de calculer sur un ordinateur donnant 7 chiffres significatifs la quantité suivante :

$$S = 1000 - \sqrt{999999}$$

La machine opère de la manière suivante : elle calcule $\sqrt{999999} = 999,9995$ avec 7 chiffres significatifs. Puis elle fait la soustraction.
Pour faire une soustraction, l'ordinateur donne aux termes de la soustraction le même exposant. L'exposant choisi est le plus grand :

$$1000 \text{ s'écrit } 0,1000000\,10^4 \text{ avec 7 chiffres significatifs}$$

$$999,9995 \text{ s'écrit } 0,9999995\,10^3 \text{ avec 7 chiffres significatifs}$$

L'ordinateur commence par choisir comme exposant 4. Du coup, nous perdons un chiffre significatif pour $999,9995$ qui s'écrit : $0,0999999\,10^4$.

La soustraction donne alors le résultat suivant : $0,0000001\,10^4$ soit : $0,001$

En conclusion, sur une machine calculant avec 7 chiffres significatifs de précision, on obtient : $S \simeq 0,001$

Si l'on avait disposé d'une machine plus précise, à 9 chiffres significatifs exacts, on aurait obtenu les calculs suivants :

$$1000 = 0,1000000000\,10^4$$
$$\sqrt{999999} = 0,999999500\,10^3 = 0,099999950\,10^4$$
$$1000 - \sqrt{999999} = 0,000000050\,10^4$$

soit 0,0005. Avec une machine à 9 chiffres significatifs, on obtient donc un bon résultat :

$$S = 0,0005$$

On constate qu'avec 7 chiffres significatifs le résultat était totalement erroné avec une erreur relative de 100 %.

Cependant en conduisant les calculs différemment, on aurait pu obtenir le bon résultat même dans le cas de calculs avec 7 chiffres exacts. Mathématiquement, on a l'égalité suivante :

$$S = 1000 - \sqrt{999999} = \frac{1}{1000 + \sqrt{999999}}$$

Calculons S selon cette deuxième formule sur notre machine à 7 chiffres : on obtient :

$$1000 \text{ soit } 0,1000000\,10^4$$

$$+\sqrt{999999} \text{ soit } 0,0999999\,10^4$$

$$= 0,1999999\,10^4$$

La division $\dfrac{1,000000}{0,1999999\,10^4}$ donne cette fois le bon résultat $0,00050$.

Un autre exemple classique : le développement de l'exponentielle

Calculons une valeur approchée de $e^{-10,2}$ sur un ordinateur. Si l'on utilise le développement de Taylor :

$$e^{-10,2} \simeq 1 - \frac{10,2}{1!} + \frac{10,2^2}{2!} - \frac{10,2^3}{3!} .... + (-1)^n \frac{10,2^n}{n!}$$

on est dans la situation catastrophique où l'on cherche à calculer un nombre de l'ordre de $3.7\ 10^{-5}$ par différences entre nombres dont le plus grand est de l'ordre de $3.3\ 10^3$. Comme sur des nombres de cet ordre la précision n'est que de 10^{-4} environ , on ne pourra évidemment pas obtenir, ne serait-ce qu'un bon chiffre significatif pour le résultat. Par contre on a également :

$$e^{-10,2} = \frac{1}{e^{10,2}}$$

et on obtiendra un bon résultat par le calcul préalable de

$$e^{10,2} \simeq 1 + \frac{10,2}{1!} + \frac{10,2^2}{2!} + \frac{10,2^3}{3!} .... + \frac{10,2^n}{n!}$$

En réalité, le calcul de l'exponentielle sur un ordinateur se fait de la façon suivante. On isole les puissances entières de e que l'on calcule par multiplications successives. Il ne reste alors qu'à prendre le développement de Taylor de l'exponentielle pour la partie restante de l'argument qui est donc < 1. Ce développement est donc très précis avec peu de termes.

1.8.3 Problèmes stables et instables

Considérons le problème différentiel suivant :

$$\begin{cases} y' = -xy \\ y(0) = 1 \end{cases} \tag{1.2}$$

Ce problème a pour solution

$$y(x) = e^{-\frac{x^2}{2}}$$

Si l'on change la condition initiale $y(0) = 1$ en la condition voisine $y(0) = 1 + \epsilon$ avec ϵ très petit, on obtient

$$y_\epsilon = (1 + \epsilon)e^{-\frac{x^2}{2}}$$

Cette nouvelle solution est voisine de la première, et de plus, tend vers celle-ci quand x augmente. On dit que ce problème est stable. Il est peu sensible aux erreurs d'arrondis commises dans la représentation des nombres ou lors des calculs.

Considérons, à l'inverse, le problème différentiel :

$$\begin{cases} y' - 10y = -11e^{-x} \\ y(0) = 1 \end{cases} \tag{1.3}$$

Ce problème a pour solution

$$y(x) = e^{-x}$$

Si l'on change la condition initiale $y(0) = 1$ en la condition voisine $y(0) = 1 + \epsilon$ avec ϵ très petit, on obtient cette fois

$$y_\epsilon = e^{-x} + \epsilon\, e^{10x}$$

Cette nouvelle solution s'écarte infiniment de la première à mesure que x tend vers l'infini. On dit que ce problème est instable. Une petite perturbation des conditions initiales entraîne un écart qui s'amplifie en tendant vers l'infini avec x. Il est impossible de résoudre sur un ordinateur un tel problème pour de grandes valeurs de x. On peut dire que ce problème est non numérisable.

En conclusion, ces exemples montrent qu'entre le calcul mathématique exact et le calcul sur ordinateur, les arrondis introduisent des erreurs qui peuvent avoir des effets extrêmement importants. Certaines formulations mathématiques sont moins sensibles aux erreurs d'arrondis, on dit qu'elles sont plus stables. Ce sont elles qu'il faut le plus possible utiliser.

1.9 Langages et outils algorithmiques

Il existe plusieurs niveaux d'implémentation des méthodes numériques. En particulier, on distingue 3 niveaux d'abstraction pour la mise en œuvre des algorithmes :

— langages de bas-niveau,
— outils d'algorithmique génériques,
— codes industriels boîte-noire.

1.9.1 Langages de bas-niveau

En utilisant des langages de bas-niveau, Fortran, C, C++, on doit décrire les détails des opérations.

Certains langages permettent cependant d'étendre, à travers la représentation objet, les opérations et concepts existants initialement. Un exemple, classiquement cité, est la définition des classes matrice et vecteur et des opérations associées, en utilisant les classes prédéfinies réel et entier et les opérations disponibles dans la librairie native du langage. La surcharge d'opérateurs permet de garder une syntaxe proche de l'écriture mathématique pour les opérations sur les nouveaux types.

```
matrice A(n,m), B(m,k), C(n,m); vecteur b(m), x(n);

x=A*b;  ! produit valide matrice-vecteur
C=A*B;  ! produit valide matrice-matrice
D=A+C;  ! somme valide matrice-matrice
```

On voit le travail à accomplir en amont, pour la prévision de toutes les configurations possibles. Par exemple, l'opérateur produit matrice∗vecteur s'écrit en bas-niveau (langage C) :

```
float A[n-1,m-1],X[n-1],b[m-1];
integer i,j;
for( i=0, i < n-1 ,i++)
 {
 X[i]=0;
 for (j =0, j < m-1, j++) {X[i] = X[i]+A[i,j]*b[j]; }
 }
```

Ainsi, en faisant $X = A * b$, on fera appel à ces lignes "encapsulées" dans l'opérateur "$*$" qui surcharge donc le produit classique de réels prévu par le langage. De nombreuses librairies existent pour diverses classes, il faut éviter, autant que possible, d'en écrire de nouvelles. Il est préférable de s'adapter à l'une des librairies existantes. D'autant plus que l'écriture de codes efficaces dans un langage objet n'est pas chose aisée et demande une bonne connaissance de la problématique ainsi que du langage.

1.9.2 Outils d'algorithmique génériques

Le premier contact avec les méthodes numériques se fait désormais au travers d'outils permettant la manipulation de concepts mathématiques évolués. Ces outils utilisent une combinaison du calcul symbolique et de la programmation bas-niveau. Comme exemples, on peut citer `Matlab, Scilab, Maple, Mathematica`. Leur but est de rester le plus proche possible de la formulation mathématique et de garder sous-jacents les aspects techniques de l'implémentation. Ces outils sont génériques. Ils permettent de considérer tout type de problèmes avec une efficacité moyenne. Ils constituent une bonne introduction à la simulation. Ces outils restent intéressants, même dans le cas où l'on est amené à écrire son propre code, car on peut les utiliser pour générer le code bas-niveau qui sera ensuite inclus, après de légères modifications, dans l'outil en développement. Bien sûr ceci ne peut concerner que des ensembles à faible nombre d'opérations, sinon, l'efficacité se dégradera.

1.9.3 Codes industriels boîte-noire

Enfin, si le domaine d'application est spécifique et si le but principal est l'exploitation d'outils numériques sur des configurations complexes, le meilleur choix reste le choix de codes "industriels" ou "spécialisés", développés pour l'application en question. Ces codes sont manipulés en boîtes-noires, ce qui signifie que l'utilisateur n'en connaît pas les détails et le contenu algorithmique exact. Il ne peut pas accéder au code source et ne dispose que d'un exécutable. Cependant il peut souvent proposer des routines "utilisateurs" pour certaines extensions, notamment pour la prise en compte de conditions aux limites ou de termes sources plus sophistiqués dans les équations. Le code boîte-noire utilisera toujours la même méthode numérique pour la résolution du problème, même en présence de ces nouveaux ingrédients, ce qui peut être parfois inadapté. En règle générale, l'utilisateur fournit la configuration de calcul dans la classe des applications pour lesquelles l'outil a été développé. Un code de crash de voiture ne sera pas appliqué en thermique ou pour une application en mécanique des fluides. En conclusion, la meilleure façon d'utiliser ces outils est de rester dans leur domaine de validité, précisé dans le guide d'utilisateur. Règle fondamentale toutefois : il ne faut jamais baser une analyse sur une seule simulation, mais il faut procéder à plusieurs simulations sur des discrétisations géométriques (maillages) différentes, avec des données physiques légèrement différentes pour pouvoir identifier d'éventuels biais et instabilités numériques.

Dans l'ensemble, la simulation numérique s'oriente de plus en plus vers l'utilisation des outils algorithmiques génériques ou bien de codes industriels et essaie d'éviter au maximum une descente au langage de bas-niveau.

1.9.4 Shell et appel d'exécutable

Une approche puissante pour la mise en œuvre des algorithmes consiste en l'utilisation d'outils disjoints, couplés à travers leurs interfaces respectives. Ceci peut être fait par la réunion de commandes dans un fichier de commande que l'on appelle `shell`. Dans cette approche, les outils communiquent entre eux par fichiers, ou bien en utilisant des commandes de passage de messages disponibles dans les librairies de communication (voir chapitre 20 pour le calcul parallèle). Le second choix est préférable si le nombre de communication est important. En effet, dans ce cas, l'écriture et la lecture des fichiers seront pénalisantes en temps de calcul, car elles exigent un accès au disque.

Les configurations où cette approche est souhaitable concernent le couplage entre modèles, l'optimisation et en régle générale, dès que plusieurs concepts interviennent dans le calcul. L'autre avantage de cette approche est la possibilité de coupler des codes industriels, des outils d'algorithmique génériques, ainsi que des parties directement développées par l'utilisateur. Cette dernière partie est destinée à être minimale et concerne une particularité éventuelle de l'application traîtée par l'utilisateur pour laquelle ni un outil générique, ni un code industriel n'existe.

Prenons par exemple, le calcul multi-disciplinaire d'un ouvrage, mettant en jeu le comportement de la structure sous le vent (on étudiera plus en détails ce type de problèmes au chapitre 13).

Les outils nécessaires à la simulation sont alors :

— 1. Un outil de manipulation géométrique ou de CAO pour la définition des géométries.

— 2. Un solveur industriel en mécanique des fluides, qui calcule la distribution de pression et la vitesse de l'écoulement autour de l'ouvrage.

— 3. Un outil d'interpolation des efforts, qui à partir des valeurs discrétisées sur le maillage fluide, produira les efforts correspondants sur la peau du maillage de la structure. Cet outil est rarement disponible, car il est trop spécifique, et dépend des discrétisations utilisées dans les deux solveurs. L'utilisateur aura donc probablement à le développer.

— 4. Un solveur industriel en structure, qui connaissant les efforts exercés sur l'ouvrage, solution du calcul fluide, prédit les déformations et contraintes dans la structure discrétisée. Ces déformations changent le domaine de simulation du solveur fluide. On doit itérer pour obtenir la solution. Il s'agit donc d'un calcul couplé.

— 5. Utilisation inverse de l'outil (3) pour reporter les déformations de la structure sur la peau du maillage fluide (les deux discrétisations sont différentes).

— 6. Un outil de déformation de maillage ou de remaillage du domaine fluide,

pour la prise en compte des modifications des géométries, et pour une nouvelle évaluation des caractéristiques de l'écoulement.

Mêler, au niveau du code, les outils ci-dessus est tout simplement impossible car on ne dispose, en général, pas de tous les codes sources. Dans l'exemple précédent, seul le point (3) a été développé par l'utilisateur. Le choix le plus pratique consiste donc à communiquer, par fichier, les informations nécessaires à chaque outil.

1.10 Opérations de base

Nous présentons certaines opérations de base et expressions qui sont utiles lors de l'écriture et l'évaluation d'un algorithme, aussi bien en langage bas-niveau que lors de l'utilisation d'un outil d'algorithmique générique.

— L'**affectation :** elle a pour objet d'affecter le contenu d'une variable à une seconde variable (le contenu de cette variable est alors perdu).

$$a = b \text{ signifie} \quad a \longleftarrow b$$

dans ce cas si : $a = 3$ et $b = 2$ alors $a = b = 2$.
Il faut que les deux types a et b soient identiques (pensez à l'affectation des valeurs d'une matrice à une autre). Ainsi, l'affectation doit être, elle aussi, surchargée dans une approche objet.
Si l'on veut permuter deux éléments (nécessairement de la même classe), il faut, évidemment, faire appel à trois affectations :

```
temporaire = x;
x = y;
y = temporaire;
```

— Les **instructions conditionnelles** sont très utiles et permettent de définir dans quelles conditions telles ou telles tâches seront effectuées :

```
Si ( test1 ) Alors
        faire tache1
      Sinon ( test2 ) Alors
        faire tache2
      ...
      Sinon
        faire tache3
Fin
```

— Les **boucles** permettent d'itérer des séquences d'instructions. On peut combiner les boucles et les instructions conditionnelles, soit en début de chaque itération, soit à la fin, pour décider s'il faut poursuivre ou s'arréter. Par exemple, avec un test en début de boucle on obtient :

```
Tant que {test} Faire
  ...
 instructions
  ...
Fin
```

et en plaçant le test en fin de l'itération on a :

```
Faire
  ...
 instructions
  ...
Tant que {test} Fin
```

Il est courant d'avoir des boucles imbriquées (par exemple des itérations en temps et en espace).

— En programmation de bas-niveau en particulier, il est bon d'identifier les actions et de les représenter au travers de **fonctions** ou de **sous-programmes**. Ainsi, la manipulation d'objets complexes a été introduite pour la première fois sous forme de subroutines. Dans notre exemple de produit matrice-vecteur, cette action peut être introduite soit par un sous-programme, soit par une fonction :

```
sousprogramme matvec(n,m,M,b,X)
(avec comme appel) :
appel matvec(n,m,M,b,X)

fonction matvec(n,m,M,b)
(avec comme appel) :
X=matvec(n,m,M,b)
```

— La **récursivité**, proche de la récurrence mathématique, est utile pour représenter une action faisant appel à elle-même (exemple type : la factorielle).

```
fonction factoriel(n)
```

```
{Si (n =  1 ) Alors
 Renvoyer 1
 Sinon
 Renvoyer  n * factoriel(n-1)
 }
```

L'idée est attrayante, mais le domaine d'application n'est pas très large.

— La **complexité** représente l'effort en terme de stockage et temps de calcul à fournir afin d'effectuer une tâche. Il existe en général une dualité entre la mémoire utilisée et les calculs à faire. En d'autres termes, l'on peut être amené à faire des calculs redondants pour diminuer la taille mémoire utilisée.

Nous allons étudier quelques algorithmes de tri, très souvent utilisés, et examiner leur complexité.

1.11 Algorithmes de tri

Il est souvent nécessaire de **trier** les éléments d'un ensemble, suivant leur appartenance à une classe d'équivalence à travers une relation d'ordre connue. Par exemple, pour trier un ensemble d'entiers $a(i), i = 1, .., n$ dans l'ordre croissant, on utilise l'algorithme de **tri par permutation**. Cet algorithme est valable quels que soient la relation d'ordre et l'ensemble d'application choisi.

```
Tri(1,n) { i=1, Tant que (i < n+1) Faire
Si (a(i+1) < a(i) ) permuter (a(i),a(i+1)); i=i+1; Fin }
```

À chaque application de cet algorithme, on fait monter, parmi les éléments restant à trier, un élément en haut de la pile. Il faut donc appliquer n fois cet algorithme, ce qui implique une complexité en n^2 comme suit :

```
j=0, Tant que (j < n+1) Faire Tri(1,n); j=j+1; Fin
```

Il est facile d'améliorer légèrement cet algorithme en évitant de boucler jusqu'à n à chaque itération, mais jusqu'à j :

```
j=1, Tant que (j < n+1) Faire Tri(1,n-j+1); j=j+1; Fin
```

qui a une complexité de $n(n-1)...1 = n(n+1)/2$.

Une optimisation plus conséquente consiste à :
— partager l'ensemble par paquet de taille paire et croissante (2, 4, ...),
— trier chaque paquet,
— permuter les paquets d'éléments.

La complexité est alors $nlog_2(n)$ qui représente le produit de l'effort à fournir sur un niveau par le nombre de fois où on peut couper par deux l'ensemble ($log_2(n)$). On n'écrit en général pas d'outils de tri et on utilise plutôt les outils existants. Cependant il est bon de connaître la complexité des algorithmes.

Chapitre 2

Méthodes numériques de base

Le but de ce chapitre est de fournir, de manière succincte, les principes de base des méthodes numériques les plus utilisées. Même si désormais le premier contact avec les méthodes numériques ne se fait plus au travers d'une programmation de ces méthodes dans les langages de bas-niveau, mais plutôt par l'utilisation d'outils tels que `Matlab, Scilab, Maple, Mathematica...` permettant la manipulation de concepts mathématiques évolués, il reste indispensable de connaître les fondements des principales méthodes pour les utiliser de façon pertinente.

2.1 Résolution des équations de type $f(x) = 0$

Nous nous restreignons, par souci de simplicité, à la recherche de racines réelles, zéros de fonctions réelles continues.

L'existence et une première localisation des solutions utilisent le théorème des valeurs intermédiaires.

Théorème 2.1.1 (Théorème des valeurs intermédiaires) *Si f est une fonction continue sur $[a, b]$, et si $f(a)f(b) \leq 0$, alors il existe au moins un point $c \in [a, b]$ tel que $f(c) = 0$.*

Si de plus f est strictement monotone sur $[a, b]$, la racine est unique dans $[a, b]$.

2.1.1 Méthode de dichotomie

La méthode de dichotomie est un procédé systématique de raffinement de la localisation d'une racine. Le mot dichotomie (dicho = deux, tomie = coupe) exprime clairement le principe de la méthode.

Soit $[a, b]$ un intervalle initial de localisation de la racine cherchée s. Supposons que l'on ait $f(a)f(b) < 0$, l'algorithme de la dichotomie s'écrit :

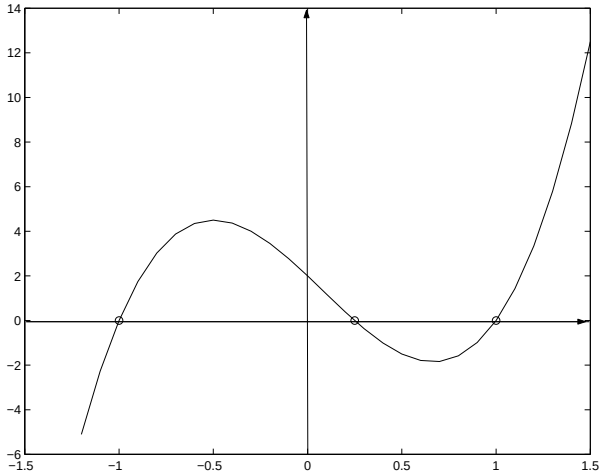


FIGURE 2.1 – Fonction présentant 3 racines.

```
Tant que( abs( b-a ) > epsilon ) faire   ! test d'arret
  calcul de m=(a+b)/2
  calcul de f(m)
    test  sur le signe de (f(a) * f(m))
    si (f(a) * f(m)) < 0 faire b=m       ! s est dans  [a,m]
    sinon faire a = m                    ! s est dans  [m,b]
Fin de boucle
```

Cet algorithme réduit à chaque pas l'amplitude de la localisation d'un facteur 2. L'erreur est donc réduite d'un facteur 2 à chaque itération. En 20 itérations, par exemple l'erreur sera 10^{-6} fois l'erreur initiale. Cette méthode est relativement lente. Par contre elle converge dans tous les cas où la fonction change de signe au voisinage de la racine (ce qui exclut les racines de multiplicités paires). C'est une méthode que nous qualifierons de méthode tout-terrain, lente mais quasiment infaillible.

2.1.2 Méthodes de point-fixe

Les méthodes de point-fixe permettent de construire des algorithmes plus rapides que la dichotomie (parfois) mais surtout des algorithmes qui se généralisent simplement au cas de problèmes en dimension supérieure à un. On ramène l'équation $f(x) = 0$ à une équation équivalente de forme point-fixe

$$x = g(x)$$

Ceci nous permettra d'obtenir simplement une méthode itérative de la forme

$$\begin{cases} x_0 \text{ donné :} & \text{initialisation} \\ x_{n+1} = g(x_n) \end{cases} \tag{2.1}$$

Si cette itération converge, elle converge vers le point-fixe de g, donc de manière

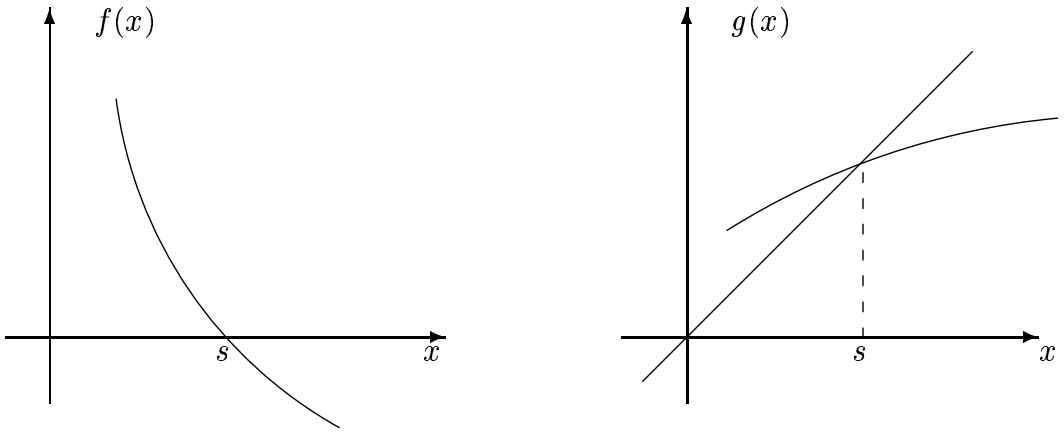


FIGURE 2.2 – Forme $f(x) = 0$ et forme point-fixe équivalente d'une équation.

équivalente vers le zéro recherché de f. La condition de convergence essentielle est une condition de contraction sur la fonction g.

Définition 2.1.1 (Application contractante) *On dit qu'une application définie de $[a, b]$ dans $[a, b]$ est contractante, ou que c'est une contraction, s'il existe un nombre $0 \leq k < 1$ tel que, pour tout couple de points distincts (x_1, x_2) de $[a, b]$, on ait :*

$$|g(x_1) - g(x_2)| \leq k|x_1 - x_2|$$

k est le facteur de contraction. Il donne la vitesse de convergence de l'itération.

Dans le cas où g est dérivable, la condition de contraction se ramène à la condition suivante sur la dérivée : $|g'(x)| \leq k < 1 \quad \forall x \in [a, b]$

Remarque 2.1.1 *Les notions de contraction et le théorème de convergence des itérations associé peuvent s'écrire et se démontrer dans le cadre général des espaces vectoriels normés (espaces de Banach). Cette possibilité de généralisation très large est un des intérêts principaux des méthodes de point-fixe (voir la démonstration générale du théorème 2.10.1).*

2.1.3 Vitesse de convergence et ordre d'une méthode itérative

Nous nous plaçons dans le cas d'une itération $x_{n+1} = g(x_n)$ convergente et nous supposons la fonction g suffisamment dérivable. L'application de la formule de Taylor au voisinage de la racine s donne :

$$x_{n+1} - s = g(x_n) - g(s) = (x_n - s)g'(s) + \frac{(x_n - s)^2}{2}g''(s) + O((x_n - s)^3)$$

Méthodes d'ordre un

Si $g'(s) \neq 0$, la limite du rapport des écarts est :

$$\lim_{n\to\infty} \frac{|x_{n+1} - s|}{|x_n - s|} = |g'(s)|$$

L'écart au pas $n + 1$ est donc du même ordre que l'écart au pas n. Le facteur de réduction d'erreur est asymptotiquement donné par $|g'(s)|$. Plus petite sera la valeur de $|g'(s)|$, plus vite se fera la convergence.

Méthodes d'ordre deux

Si $g'(s) = 0$, l'erreur au pas $n+1$ est un infiniment petit d'ordre ≥ 2 par rapport à l'erreur au pas n. On obtient en effet :

$$\lim_{n\to\infty} \frac{|x_{n+1} - s|}{|x_n - s|^2} = \frac{1}{2}|g''(s)|$$

La convergence est dite quadratique. La réduction du nombre des itérations est spectaculaire dès l'ordre 2. À partir d'une erreur de 0.1, on obtient 10^{-8} en trois itérations.

On peut essayer, pour chaque problème particulier, de construire une itération de point-fixe convergente. Il est évidemment plus intéressant d'utiliser des méthodes générales applicables pour toute équation $f(x) = 0$. Voici une famille de méthodes classiques très utiles dans la pratique.

2.1.4 Méthode de Newton et Quasi-Newton

On obtient évidemment une forme point-fixe équivalente à $f(x) = 0$ en considérant la famille $x = x - \lambda(x)f(x)$, avec λ définie et non nulle sur un intervalle $[a, b]$ contenant la racine s. Parmi tous les choix possibles pour λ, le choix qui conduit à la convergence la plus rapide est $\lambda = \dfrac{1}{f'}$ (f' dérivée de f).

La méthode obtenue ainsi est la méthode de Newton. Il est facile de vérifier que l'itération

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{2.2}$$

est d'ordre deux si elle converge. Évidemment la méthode de Newton n'est plus efficace si f' s'annulle, donc dans le cas de racines multiples. Il existe des résultats de convergence globale de Newton. Ils supposent des hypothèses de monotonie et de concavité de signe constant sur f (pas de point d'inflexion). La méthode de Newton est très classique. Elle s'interprète géométriquement comme méthode de la tangente.

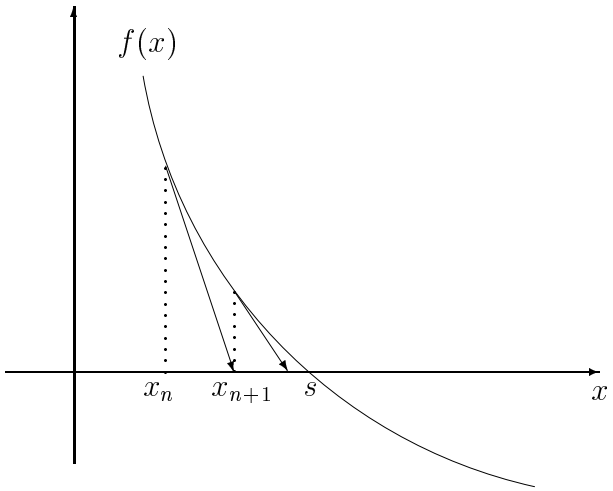


FIGURE 2.3 – Interprétation géométrique de la méthode de Newton.

La méthode de Newton nécessite le calcul des dérivées $f'(x_n)$. C'est un inconvénient dans la pratique où l'on ne dispose pas toujours d'expression analytique pour la fonction f.

Une solution simple est fournie par la *méthode de la sécante ou de fausse-position* dans laquelle on remplace le calcul de $f'(x_n)$ par l'approximation

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

Ce qui donne

$$x_{n+1} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})} \tag{2.3}$$

Les procédures de résolution d'équations de type $f(x) = 0$ que l'on trouve dans les outils génériques d'algorithmique (Matlab par exemple), combinent en général une première approche de la racine par quelques pas de dichotomie, suivis, pour la convergence fine, par une méthode rapide afin d'obtenir une grande précision en peu d'itérations. L'algorithme proposé par Matlab est dû à Dekker (1969). La méthode rapide utilisée est une interpolation quadratique inverse, assurant l'ordre deux de convergence (comme Newton), sans calcul de la dérivée de f.

2.1.5 Méthode de Newton et Quasi-Newton pour les systèmes

La méthode de Newton se généralise en dimension supérieure à un. On peut en effet montrer que le choix du $n+1^e$ itéré x_{n+1} est tel que

$$f(x_{n+1}) = O(|x_{n+1} - x_n|^2)$$

En effet un développement simple donne :

$$f(x_{n+1}) = f(x_n) + (x_{n+1} - x_n)f'(x_n) + O(|x_{n+1} - x_n|^2)$$

On voit donc que le choix

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

assure bien

$$f(x_{n+1}) = O(|x_{n+1} - x_n|^2)$$

On retrouve ainsi l'ordre 2 de la méthode de Newton.

Dans le cas d'un système de N équations non-linéaires, on peut écrire

$$F(X_{n+1}) = F(X_n) + F'(X_n)(X_{n+1} - X_n) + O(\|X_{n+1} - X_n\|^2) \tag{2.4}$$

la même idée conduit au choix suivant pour l'itération vectorielle :

$$X_{n+1} = X_n - \{F'(X_n)\}^{-1} F(X_n) \tag{2.5}$$

où $F'(X_n)$ désigne la matrice jacobienne de coefficients $\dfrac{\partial F_i}{\partial x_j}(X_N)$. Pratiquement le $n+1^e$ itéré est calculé à chaque itération par résolution de systèmes linéaires

$$F'(X_n)[X_{n+1} - X_n] = -F(X_n) \tag{2.6}$$

La matrice $F'(X_n)$ doit être assemblée et recalculée et le système doit être résolu à chaque itération. Ceci rend la méthode de Newton très coûteuse en temps de calcul.

Pour éviter ces inconvénients, on utilise des méthodes dites de *Quasi-Newton* dans lesquelles sont proposées des approximations de l'inverse de la matrice Jacobienne. Une méthode classique et efficace de ce type est la méthode BFGS (Broyden-Fletcher-Goldfarb-Shanno).

Application à l'optimisation

Dans un contexte d'optimisation, la recherche du minimum d'une fonctionnelle J peut être ramenée à la recherche du point qui annule son gradient $\nabla J(x)$, x désignant la paramétrisation du problème (voir chapitre 17). On pourra alors utiliser les méthodes de type Newton ou Quasi-Newton (nous renvoyons à la littérature pour les méthodes d'optimisation en dimension un du type de la méthode de la section dorée qui ne nécessite pas le calcul des dérivées). Ces méthodes demandent le calcul exact ou approché de la matrice des dérivées partielles secondes ou matrice Hessienne. La méthode BFGS permet de construire directement une approximation de l'inverse de la Hessienne $\mathbf{H}$, en démarrant de la matrice identité ($\mathbf{H}_0 = Id$) et en appliquant l'itération suivante :

$$\mathbf{H}_{p+1} = \mathbf{H}_p + \left(1 + \frac{{\gamma_p}^{\mathrm{T}} \mathbf{H}_p \gamma_p}{{\delta x_p}^{\mathrm{T}} \gamma_p}\right) \frac{\delta x_p {\delta x_p}^{\mathrm{T}}}{{\delta x_p}^{\mathrm{T}} \gamma_p} - \frac{1}{2} \frac{{\delta x_p}^{T} (\mathbf{H}_p + (\mathbf{H}_p)^T) \gamma_p}{\delta x_p \gamma_p} \tag{2.7}$$

où p indique l'itération d'optimisation, $\delta x_p = x_p - x_{p-1}$ la variation du paramètre et $\gamma_p = \nabla J(x_{p+1}) - \nabla J(x_p)$. Voir chapitre 17 pour des développements sur l'optimisation.

2.2 Interpolation

Une collection de valeurs notées y_i étant données pour un ensemble d'abscisses x_i, pour $i = 0$ à n, l'interpolation est le procédé qui consiste à déterminer une fonction, d'une famille choisie a priori, prenant les valeurs données aux abscisses correspondantes. Le choix de fonctions polynomiales est le plus classique. Dans ce cas, le polynôme d'interpolation est le polynôme de degré minimal passant par les $n + 1$ points donnés. Ce polynôme est unique et il est de degré inférieur ou égal à n. Si l'on exprime le polynôme recherché dans la base canonique sous la forme

$$P(x) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$$

on doit résoudre un système linéaire de $n + 1$ équations à $n + 1$ inconnues. Sous cette forme le calcul est donc coûteux. La solution de ce système est très sensible aux erreurs d'arrondis.

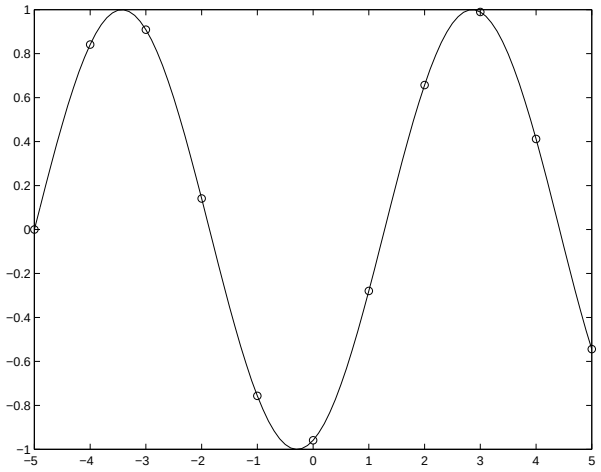


FIGURE 2.4 – Interpolation polynomiale de degré 10 pour une fonction sinusoïde.

2.2.1 Polynômes de Lagrange

Une solution simple, élégante et économique de ce problème est fournie par l'utilisation de la base des polynômes de Lagrange.

On considère les $n + 1$ polynômes L_i de degré $\leq n$ qui vérifient, pour tout i et j compris entre 0 et n, les égalités :

$$\begin{cases} L_i(x_i) = 1 \\ L_i(x_j) = 0 \end{cases} \tag{2.8}$$

Les polynômes L_i sont déterminés de façon unique par les $n+1$ équations ci-dessus. Il est facile de montrer qu'ils forment une base de l'espace des polynômes de degré inférieur ou égal à n et qu'ils s'écrivent :

$$L_i(x) = \prod_{j=0 j \neq i}^{n} \frac{x - x_j}{x_i - x_j} \tag{2.9}$$

Exprimé dans cette nouvelle base, le polynôme d'interpolation s'écrit

$$P(x) = \sum_{i=0}^{n} y_i \, L_i(x) \tag{2.10}$$

La relation ci-dessus, facile à vérifier, explique l'intérêt de la base de Lagrange. Les coefficients du polynôme d'interpolation cherché sont, dans cette base, tout simplement les valeurs y_i données. Exprimé autrement, le changement de base, de la base canonique à la base de Lagrange, a transformé le système à résoudre en un système à matrice identité.

2.2.2 Limites de l'interpolation polynomiale

L'interpolation polynomiale est la base de nombreuses techniques numériques, en particulier les techniques d'intégration approchée. Elle se généralise de façon naturelle aux cas de dimension supérieure à un.

Cependant elle a des limites :

- théoriques : on n'est pas assuré de la convergence du polynôme d'interpolation vers la fonction interpolée lorsque l'on fait tendre le nombre de points d'interpolation (et donc le degré du polynôme) vers l'infini (voir le phénomène de Runge pour la fonction $f(x) = \frac{1}{1+x^2}$) ;
- numériques : même dans le cas où la convergence théorique est assurée, les instabilités de calcul provenant de l'accumulation des erreurs d'arrondis, auxquelles le procédé d'interpolation polynomiale est particulièrement sensible, limite l'usage de cette technique dès que le nombre de points d'interpolation dépasse la dizaine ;
- pratiques : remarquons que dans de nombreux cas, les valeurs données résultent d'expériences ou de calculs préalables. Ces valeurs sont donc approximatives. Le problème réel n'est alors plus un problème d'interpolation, mais plutôt un problème de meilleure approximation pour lequel les méthodes de moindres carrés, présentées plus bas, sont mieux adaptées.

2.2.3 Interpolation par des splines

Pour éviter l'inconvénient, signalé plus haut, de l'augmentation du degré du polynôme et de l'instabilité qui en résulte, lorsque le nombre de points est grand,

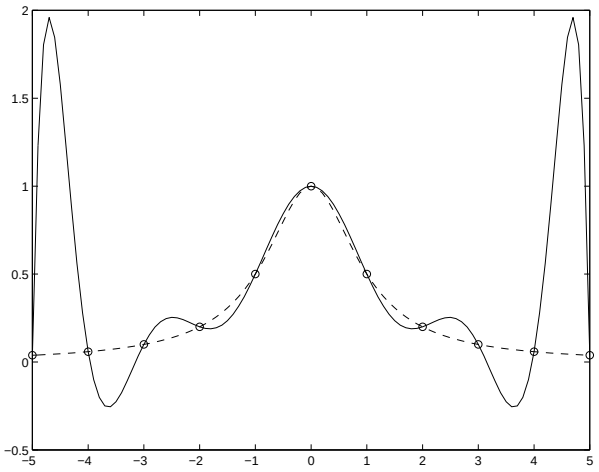


FIGURE 2.5 – Divergence de l'interpolation polynomiale pour la fonction $y = \frac{1}{1+x^2}$. Phénomène de Runge. En pointillés : la fonction, en traits pleins : le polynôme d'interpolation de degré 10 construit sur 11 points régulièrement espacés.

tout en restant dans un procédé d'interpolation, on subdivise l'ensemble des points donnés en plusieurs sous-ensembles. On réalise les interpolations sur ces petits sous-ensembles, ce qui permet de se limiter à des polynômes de bas degré. Les fonctions polynomiales par morceaux obtenues sont à la base des éléments finis de Lagrange.

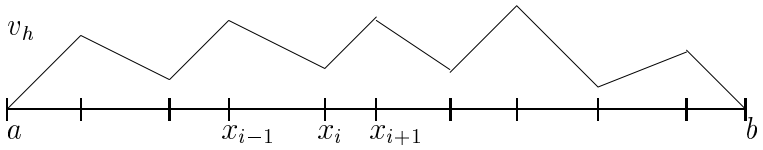


FIGURE 2.6 – Une fonction affine par morceaux.

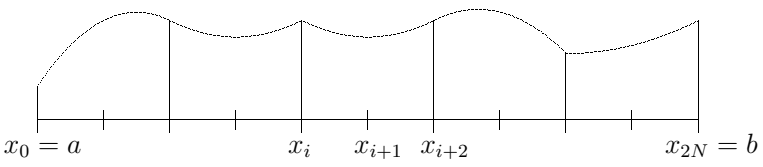


FIGURE 2.7 – Une fonction polynomiale de degré deux par morceaux.

Les interpolations ci-dessus produisent des fonctions globalement continues mais non continûment dérivables.

Les *splines cubiques* d'interpolation sont des fonctions cubiques par morceaux, globalement C^2. On obtient leur expression analytique, segment par segment, en imposant les conditions suivantes aux points x_i d'interpolation

$$s(x_i) = y_i \text{ donné pour } i = 0, ..n, \quad s' \text{ et } s'' \text{ continues}$$

Les inconnues du problème sont alors les dérivées secondes C_i de la spline aux

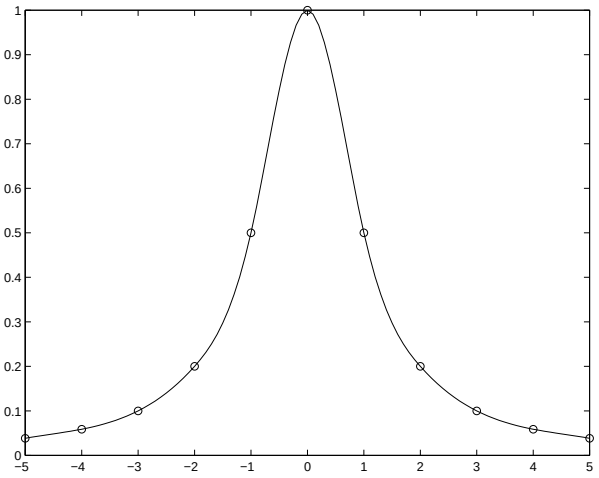


FIGURE 2.8 – Spline cubique d'interpolation pour $y = \frac{1}{1+x^2}$.

points x_i. On suppose la dérivée seconde de la spline affine par intervalles. On intègre deux fois en prenant en compte les conditions de continuité de la dérivée et les valeurs données y_i aux points x_i. On en déduit les expressions suivantes de la spline sur chaque intervalle $[x_i, x_{i+1}]$:

$$\begin{aligned} S_i(x) = \tfrac{C_i}{6}[\tfrac{(x_{i+1}-x)^3}{h_i} - h_i(x_{i+1}-x)] + \tfrac{C_{i+1}}{6}[\tfrac{(x-x_i)^3}{h_i} - h_i(x-x_i)] \\ + y_i\tfrac{(x_{i+1}-x)}{h_i} + y_{i+1}\tfrac{(x-x_i)}{h_i} \end{aligned} \tag{2.11}$$

avec $h_i = x_{i+1} - x_i$, et où les C_i sont solutions du système tridiagonal :

$$\frac{h_{i-1}}{6}C_{i-1} + \frac{h_{i-1}+h_i}{3}C_i + \frac{h_i}{6}C_{i+1} = \frac{y_{i+1}-y_i}{h_i} - \frac{y_i - y_{i-1}}{h_{i-1}}$$
$$\text{pour } i = 1, ..n-1, \text{ complété, en général, par } C_0 = C_n = 0.$$

Voici par exemple (Figure 2.8) la spline cubique d'interpolation de la fonction $f(x) = \dfrac{1}{1+x^2}$ sur 10 intervalles. On observe la stabilité de cette interpolation par contraste avec le résultat obtenu (Figure 2.5) par interpolation polynomiale.

2.3 Approximation au sens des moindres carrés

L'instabilité du procédé d'interpolation polynomiale lorsque le nombre de points augmente, d'une part, l'incertitude des résultats de mesure, d'autre part, conduisent à préférer à l'interpolation des méthodes d'approximation. Ainsi il est clair que l'expérimentateur qui relèvera 100 points quasiment alignés sera plus intéressé par la droite passant " au mieux " par ces 100 points plutôt que par le polynôme de degré 99 réalisant l'interpolation exacte.

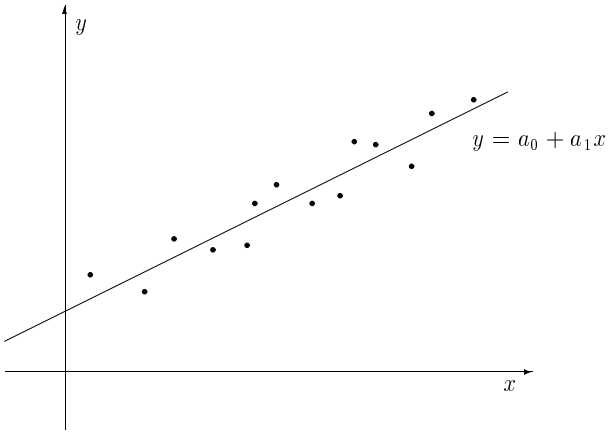


FIGURE 2.9 – Droite des moindres carrés

La plus célèbre et la plus utile des méthodes d'approximation est la méthode des moindres carrés. La formalisation de l'idée intuitive d'une droite représentant "au mieux" un nuage de points au sens des moindres carrés se fait de la manière suivante.

2.3.1 Droite des moindres carrés

Soient N valeurs $y_1, y_2, ...y_i, ...y_N$ données aux N abscisses $x_1, x_2, ...x_i, ...x_N$. Le polynôme P de degré un : $P(x) = a_0 + a_1 x$ (représenté par une droite) qui réalise la meilleure approximation au sens des moindres carrés des valeurs y_i données aux points x_i est celui qui minimise la somme des carrés des écarts entre les y_i et les $P(x_i)$, soit

$$S(a_0, a_1) = \sum_{i=1}^{N} [y_i - (a_0 + a_1 x_i)]^2 \tag{2.12}$$

S apparaît comme le carré de la norme euclidienne du vecteur de composantes $y_i - (a_0 + a_1 x_i)$. La minimisation de S s'interprète donc comme la recherche du

vecteur le plus proche du vecteur $Y \in I\!R^N$ de composantes y_i, dans le sous-espace de dimension deux engendré par les vecteurs U, de composantes toutes égales à 1, et X, de composantes x_i. Comme la norme utilisée est la norme euclidienne, le vecteur le plus proche est le projeté orthogonal. On obtient ses composantes a_0 et a_1 en écrivant les relations d'orthogonalité :

$$\begin{cases} (Y - a_0U - a_1X|U) = \sum_{i=1}^{N} [y_i - (a_0 + a_1x_i)]\ 1 = 0 \\ (Y - a_0U - a_1X|X) = \sum_{i=1}^{N} [y_i - (a_0 + a_1x_i)]\ x_i = 0 \end{cases} \tag{2.13}$$

Ceci conduit au système dit des équations normales pour a_0 et a_1, coefficients de la droite des moindres carrés (ou de régression) cherchée.

$$\begin{pmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{pmatrix} \begin{pmatrix} a_0 \\ \\ a_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i y_i \end{pmatrix} \tag{2.14}$$

2.3.2 Généralisation : polynôme des moindres carrés

Il est facile de généraliser le calcul précédent au cas de la recherche du polynôme de degré $\leq m$, avec $m << N$, qui réalise la meilleure approximation au sens des moindres carrés des y_i. Ce polynôme minimise

$$S(a_0, a_1, ...a_m) = \sum_{i=1}^{N} [y_i - (a_0 + a_1x_i + a_2x_i^2 + ...a_mx_i^m)]^2 \tag{2.15}$$

On obtient les relations d'orthogonalité :

$$\begin{cases} (Y - a_0U - a_1X - ...a_mX^m|U) = 0 \\ (Y - a_0U - a_1X - ...a_mX^m|X) = 0 \\ .... \\ (Y - a_0U - a_1X - ...a_mX^m|X^m) = 0 \end{cases} \tag{2.16}$$

où l'on a noté X^m le vecteur de composantes x_i^m. Les valeurs des coefficients a_k du polynôme des moindres carrés s'en déduisent après résolution du système linéaire déduit de (2.16).

Remarque 2.3.1 *Le système des moindres carrés ci-dessus est mal conditionné (il est de plus en plus sensible aux erreurs d'arrondis à mesure que m augmente).*

On se limite habituellement à des polynômes de degré peu élevé. La résolution pratique des problèmes de moindres carrés se fait par des algorithmes spécifiques d'élimination pour des systèmes rectangulaires surdéterminés. Ces algorithmes utilisent la factorisation QR (technique de Householder).

On peut envisager, comme dans le cas de l'interpolation, la recherche d'une meilleure approximation par des splines en découpant l'ensemble des points en sous-ensembles plus petits. Cette idée de meilleure approximation au sens des moindres carrés par des splines est à la base de nombreuses techniques d'approximation et de représentation de données.

2.4 Intégration numérique

La plupart des formules d'intégration numérique proviennent de méthodes d'interpolation polynomiales. Le calcul de l'intégrale d'une fonction est approché par le calcul exact de l'intégrale d'un polynôme interpolant cette fonction en certains points x_k pour $k = 1, ..p$. On obtient ainsi une forme générale pour les quadratures numériques

$$\int_a^b f(x)dx = \sum_{k=1}^{p} A_k f(x_k) \tag{2.17}$$

Les x_k sont alors les points d'intégration et les coefficients A_k les poids de la formule de quadrature.

Le coût d'une formule est mesuré par le nombre de calculs de f nécessaires, donc par le nombre de points d'intégration. Le critère d'efficacité choisi est le degré maximal de l'espace des polynômes pour lequel la formule est exacte. La précision d'une formule d'intégration numérique est mesurée par son ordre.

Définition 2.4.1 *On dit qu'une formule est d'ordre k si k est le plus grand entier pour lequel elle donne l'intégrale exacte de tout polynôme de degré $\leq k$. On montre que l'erreur d'intégration est alors, pour toute fonction suffisamment régulière, un infiniment petit d'ordre k du pas d'intégration h.*

2.4.1 Formules des rectangles

Ce sont les formules à un point d'intégration qui proviennent d'interpolation par des polynômes constants.

$$\int_a^b f(x)dx \approx (b-a)f(\alpha) \tag{2.18}$$

Le coût d'une telle formule à un point est celui d'une évaluation de f. Les choix classiques sont $\alpha = a$, $\alpha = b$ et le choix donnant la meilleure formule dite *formule du point-milieu* (exacte pour les fonctions affines) est $\alpha = \frac{a+b}{2}$.

2.4.2 Formule des trapèzes

On considère cette fois l'interpolation par un polynôme de degré un construit sur les points a et b. On obtient la formule classique des trapèzes :

$$\int_a^b f(x)dx \approx \frac{b-a}{2}\,[f(a)+f(b)] \tag{2.19}$$

Cette formule, à deux points, est clairement exacte pour les polynômes de degré ≤ 1.

2.4.3 Formule de Simpson

En utilisant l'interpolation sur les trois points $a, \frac{a+b}{2}, b$, on obtient la formule de Simpson, que l'on vérifie être exacte pour les polynômes de degré ≤ 3.

$$\int_a^b f(x)dx \approx \frac{b-a}{6}[f(a) + 4\, f(\frac{a+b}{2}) + f(b)] \tag{2.20}$$

Cette formule à trois points nécessite donc trois évaluations de f par segment (voir cependant, paragraphe 2.4.7 dans le cas d'un segment global subdivisé en sous-intervalles, les formules composites de calcul d'intégrales par les méthodes des trapèzes et de Simpson).

2.4.4 Formules de Gauss

Dans les formules précédentes le choix des points d'intégration était fixé (extrémités et/ou milieu des intervalles d'intégration). Dans les formules de type Gauss, les points d'intégration sont choisis de manière à obtenir la précision la plus élevée.
La **Formule de Gauss Legendre** à 2 points, est exacte pour les polynômes de degré ≤ 3 :

$$\int_a^b f(x) \approx \frac{b-a}{2}[f(\xi_1) + f(\xi_2)]. \tag{2.21}$$

$$\text{avec} \quad \xi_1 = \frac{a+b}{2} - \frac{b-a}{2}\frac{\sqrt{3}}{3} \quad \text{et} \quad \xi_2 = \frac{a+b}{2} + \frac{b-a}{2}\frac{\sqrt{3}}{3}$$

2.4.5 Intégration en dimension deux

Dans le cas de domaines **quadrangulaires**, on ramène les intégrales sur le carré unité C de sommets $A_1(0,0), A_2(1,0), A_3(1,1),\ A_4(0,1)$.

Les formules de quadrature sur le carré se déduisent facilement des formules en dimension un. On obtient ainsi à partir de la formule des trapèzes, la formule :

$$\iint_C F(x,y)dxdy \approx \frac{1}{4}[F(A_1) + F(A_2) + F(A_3) + F(A_4)] \tag{2.22}$$

exacte pour les polynômes de type Q_1 (de degré un par rapport à chaque variable, voir leur définition chapitre 8).

La formule de Simpson donne :

$$\begin{aligned}\iint_C F(x,y)dxdy &\approx \frac{1}{36}[F(A_1) + F(A_2) + F(A_3) + F(A_4)] \\ &+\frac{1}{9}[F(A_{12}) + F(A_{23}) + F(A_{34}) + F(A_{41})] + \frac{4}{9}F(G)\end{aligned} \tag{2.23}$$

La formule de Gauss-Legendre devient :

$$\iint_C F(x,y)dxdy \approx \frac{1}{4}[F(B_1)+F(B_2)+F(B_3)+F(B_4)] \tag{2.24}$$

où

$$B_1=(\xi_1,\xi_1),\ B_2=(\xi_2,\xi_1),\ B_3=(\xi_2,\xi_2),\ B_4=(\xi_1,\xi_2),$$

$$\text{avec} \quad \xi_1=\frac{1}{2}-\frac{1}{2}\frac{\sqrt{3}}{3} \quad \text{et} \quad \xi_2=\frac{1}{2}+\frac{1}{2}\frac{\sqrt{3}}{3}$$

Dans le cas de **triangles**, on dispose d'une formule d'intégration exacte pour tout produit de fonctions barycentriques (voir leur définition chapitre 8) sur un triangle quelconque T :

$$\iint_T \lambda_1^n\,\lambda_2^p\,\lambda_3^q\,dxdy = 2\,Aire(T)\,\frac{n!p!q!}{(n+p+q+2)!} \tag{2.25}$$

Cas particuliers importants (pour une programmation efficace des éléments finis) :

$$\iint_T \lambda_i\,dxdy = \frac{Aire(T)}{3} \quad \forall i=1,2,3$$

$$\iint_T \lambda_i^2\,dxdy = \frac{Aire(T)}{6} \quad \forall i=1,2,3$$

$$\iint_T \lambda_i\,\lambda_j\,dxdy = \frac{Aire(T)}{12} \quad \forall i,j=1,2,3$$

Il existe également des formules de quadrature approchée. Par exemple la formule suivante exacte pour les polynômes de degré inférieur ou égal à un, construite sur les sommets du triangle :

$$\iint_T F(x,y)\,dxdy = \frac{Aire(T)}{3}[F(A_1)+F(A_2)+F(A_3)] \tag{2.26}$$

et la formule exacte pour les polynômes de degré inférieur ou égal à deux, construite sur les milieux des côtés :

$$\iint_T F(x,y)\,dxdy = \frac{Aire(T)}{3}[F(A_{12})+F(A_{23})+F(A_{31})] \tag{2.27}$$

2.4.6 Intégration en dimension trois

Dans le cas d'éléments hexaédriques (briques) on ramène les intégrations sur le cube de référence. Les formules pour le cube se déduisent aisément des formules classiques en dimension un (exactement comme on l'a fait pour le carré en dimension deux).

Dans le cas d'éléments tétraédriques, on dispose, comme pour les triangles, d'une formule d'intégration exacte pour tout produit de fonctions barycentriques sur un tétraèdre quelconque T :

$$\iiint_T \lambda_1^n \, \lambda_2^p \, \lambda_3^q \, \lambda_4^r \, dxdydz = 6 \, Volume(T) \, \frac{n!p!q!r!}{(n+p+q+r+3)!} \tag{2.28}$$

Cas particuliers importants :

$$\iiint_T \lambda_i \, dxdydz = \frac{Volume(T)}{4} \quad \forall i = 1,2,3,4$$

$$\iiint_T \lambda_i^2 \, dxdydz = \frac{Volume(T)}{10} \quad \forall i = 1,2,3,4$$

$$\iiint_T \lambda_i \, \lambda_j \, dxdydz = \frac{Volume(T)}{20} \quad \forall i,j = 1,2,3,4$$

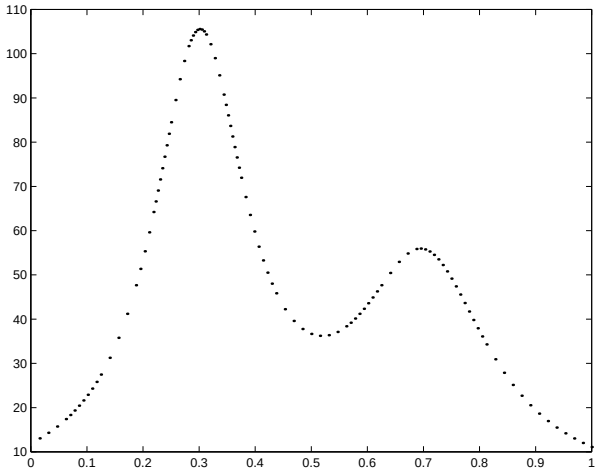


FIGURE 2.10 – Choix optimal de points d'intégration par Matlab pour la fonction $f(x) = \frac{1}{(x-0.3)^2+0.01} + \frac{1}{(x-0.7)^2+0.02}$.

2.4.7 Formules composites, maillages et méthodes adaptatives

Toutes les formules présentées ci-dessus sont des formules de base, utilisables sur de petits éléments en dimension un, deux ou trois. Pour faire un calcul réel, il faut préalablement découper le domaine d'intégration global en un ensemble de petits sous-domaines élémentaires. Voici, pour fixer les idées, le calcul global, de l'intégrale d'une fonction f, par la méthodes des trapèzes, sur un intervalle $[a, b]$ découpé uniformément en N sous-intervalles de longueur h (le pas d'intégration) :

$$\int_a^b f(x)dx \approx \frac{h}{2}(f(a) + f(b)) + h\sum_{i=1}^{N-1} f(a + ih)$$

et le même calcul par Simpson

$$\int_a^b f(x)dx \approx \frac{h}{3}(f(a) + f(b)) + \frac{2h}{3}\sum_{i=1}^{P-1} f(a + 2ih) + \frac{4h}{3}\sum_{i=1}^{P} f(a + (2i - 1)h)$$

où $P = N/2$.

Cependant, un découpage géométrique uniforme en sous-intervalles égaux n'est pas optimal. L'opération de discrétisation géométrique ou "maillage" que l'on retrouvera dans le contexte des méthodes de différences, d'éléments ou de volumes finis est déterminante pour la précision du résultat. Il faut mettre plus de points d'intégration là où la fonction présente des variations relatives rapides. Dans les outils génériques comme Matlab, on trouve des procédures de quadrature adaptatives qui choisissent automatiquement les points d'intégration (ou la taille des mailles) selon les variations de la fonction à intégrer. Typiquement, ces méthodes utilisent des indicateurs d'erreurs locaux basés sur l'écart entre deux calculs effectués avec deux méthodes de base différentes, trapèzes et Simpson par exemple (voir Fig 2.10).

2.5 Résolution des équations différentielles

Nous limiterons l'exposé au cas simple d'équations différentielles ordinaires de la forme

$$\begin{cases} \text{Trouver la fonction } y : \quad x \to y(x) \quad \text{ telle que :} \\ y'(x) = f(x, y(x)) \qquad \forall x \in [a, b] \\ y(a) = y_0 \end{cases} \tag{2.29}$$

Les problèmes différentiels de ce type sont appelés problèmes de Cauchy ou problèmes à valeurs initiales. Si f est continue et si elle vérifie une condition de Lipschitz par rapport à la deuxième variable (Il existe $L > 0$ tel que $\forall x \in [a, b]$ et $\forall y_1,\ y_2$ on ait : $|f(x, y_1) - f(x, y_2)| \leq L|y_1 - y_2|$), alors le problème admet une solution unique y pour toute valeur initiale. On dit qu'il est "bien posé". On a vu, lors du chapitre précédent, que certains problèmes différentiels bien posés du point de vue théorique pouvaient s'avérer impossibles à résoudre numériquement car instables. Numériquement, il faudra en effet considérer l'ensemble des solutions voisines de la solution exacte cherchée, solutions voisines correspondant à de petites perturbations des conditions initiales. Si ces solutions ne s'écartent pas trop de la solution de référence exacte, on aura un problème stable et on pourra construire des approximations numériques convenables.

En généralisant l'écriture de l'équation (2.29) au cas d'une fonction inconnue vectorielle, on pourra traiter des systèmes différentiels et des équations ou systèmes d'ordre supérieur à un par des extensions naturelles des techniques que nous présentons ci-dessous dans le cas de la dimension un par souci de simplicité.

2.5.1 Principe général des méthodes numériques

La solution exacte d'une équation différentielle du type (2.29) est une fonction continue. Les ordinateurs ne peuvent fournir qu'un nombre fini de résultats numériques. Tout commence donc par un choix préalable d'un nombre fini de points x_i sur $[a, b]$. Ceci s'appelle une discrétisation ou un maillage du domaine géométrique (ici le segment $[a, b]$). On limitera le calcul au calcul approché de la solution en ces points.

Le choix des points x_i est évidemment crucial pour la qualité de la solution numérique obtenue. Le maillage doit permettre de représenter de façon précise la solution. Comme cette solution est au départ inconnue, on procède par des techniques d'adaptation de maillage a posteriori. On calcule une première solution sur un premier maillage. On déduit de ce premier calcul les zones de forte variation de la solution. On raffine le maillage dans ces zones.

Encore une fois dans un souci de simplicité, nous présenterons ici les méthodes numériques dans le cas de maillage à pas uniformes. Le problème de l'adaptation

de maillage sera traité dans un cadre général chapitre 18.

On peut construire des schémas d'intégration d'équations différentielles de diverses manières. Par exemple :

- les schémas d'intégration à un pas peuvent être obtenus en utilisant des formules d'intégration numérique approchée. En introduisant des sous-pas, on obtient les schémas de Runge et Kutta qui sont les plus pratiques et les plus utilisés ;
- les schémas multi-pas peuvent être construits par développement de Taylor.

Deux critères principaux gouvernent le choix d'une méthode :

- l'ordre, qui mesure la précision du schéma ou erreur de troncature faite en remplaçant l'équation différentielle exacte par son approximation. L'ordre de la méthode donnera, à convergence, l'ordre de l'erreur commise en fonction du pas de discrétisation.
- la stabilité, qui concerne le comportement de la solution approchée discrète et la propagation des erreurs d'arrondis dans le cas d'un calcul réel pour un pas de discrétisation fixé. Le schéma est stable si la solution discrète reste bornée quel que soit le nombre de pas de discrétisation.

Remarque 2.5.1 *Un critère supplémentaire et important de choix des schémas concerne la facilité de mise en œuvre, notamment lors d'un redémarrage des calculs. Imaginons un calcul instationnaire ne pouvant faire l'objet d'un calcul complet, soit en raison de la modélisation (comme en météorologie par exemple, où de nouvelles conditions initiales et aux limites doivent être assimilées chaque jour par le modèle), soit en raison de l'implémentation et de la durée du calcul. Par exemple, sur un calculateur parallèle, le temps d'attente est lié au temps de calcul et à la quantité de mémoire requise. Dans le premier cas, comme dans le second, on aura recours aux approches à un pas de type Runge et Kutta pour assurer une précision élevée, plutôt qu'aux schémas multi-pas. En effet, quelle que soit la précision demandée, les méthodes de Runge et Kutta ne nécessitent que le stockage d'un seul état pour un redémarrage des calculs.*

2.5.2 Méthodes à un pas

Dans ces méthodes la valeur approchée y_{n+1} de la fonction inconnue y pour l'abscisse x_{n+1} est calculée en fonction des seules valeurs de l'abscisse précédente x_n, de l'approximation y_n et du pas de discrétisation h.
Si y_{n+1} s'obtient par une formule explicite de la forme

$$y_{n+1} = y_n + \Phi(x_n, y_n, h)$$

on dit que la méthode est *explicite*.

Si par contre y_{n+1} est donnée par une relation de la forme générale

$$y_{n+1} = y_n + \Phi(x_n, y_n, y_{n+1}, h)$$

on ne pourra l'obtenir que par la résolution d'une équation. On dit que la méthode est *implicite*.

La fonction Φ définit la méthode utilisée.

Ces schémas sont obtenus, par exemple, en intégrant l'équation différentielle et en utilisant des formules d'intégration numérique pour le second membre. L'ordre du schéma sera égal au degré du polynôme pour lequel l'intégration est exacte + 1.

$$\int_{x_n}^{x_{n+1}} y'(x)\,dx = y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} f(x, y(x))\,dx$$

À titre d'exemple, on obtient les schémas suivants :

— A l'ordre 1, avec une formule exacte pour les polynômes constants par morceaux, on obtient le **schéma d'Euler explicite :**

$$y_{n+1} - y_n = hf(x_n, y_n)$$

— Toujours à l'ordre 1, mais en utilisant le point d'arrivée, on obtient le **schéma d'Euler implicite :**

$$y_{n+1} - y_n = hf(x_{n+1}, y_{n+1})$$

— A l'ordre 2, en utilisant la méthode des trapèzes, on obtient le **schéma des trapèzes ou de Crank-Nicolson :**

$$y_{n+1} - y_n = \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1})].$$

En général un schéma explicite a une complexité plus faible mais impose des conditions de stabilité plus restrictives (voir plus loin).

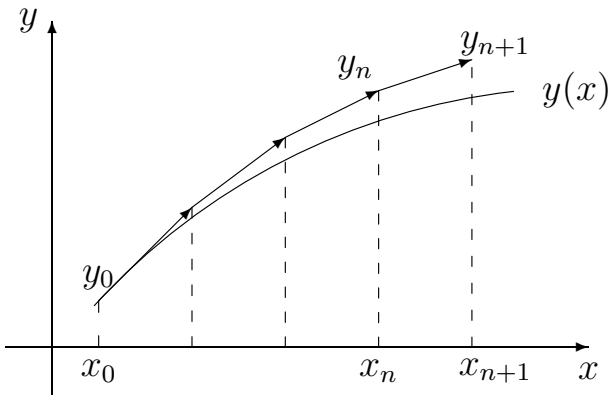


FIGURE 2.11 – Interprétation de la méthode d'Euler comme méthode de la tangente. En chaque point x_n, utiliser un développement de Taylor à l'ordre un revient à remplacer la courbe solution par sa tangente.

2.5.3 Interprétations de la méthode d'Euler explicite

La méthode d'Euler est le prototype le plus simple des méthodes numériques de résolution des équations différentielles. Pour l'équation

$$\begin{cases} y'(x) = f(x, y(x)) & \forall x \in [a, b] \\ y(a) = y_0 \end{cases}$$

elle s'écrit :

$$\begin{cases} y_0 \text{ donné} \\ y_{n+1} = y_n + hf(x_n, y_n) \end{cases} \tag{2.30}$$

C'est la plus simple des méthodes explicites à un pas.

1. La méthode d'Euler provient du développement de Taylor d'ordre un de y au voisinage de x_n. On peut montrer que, lorsque cette méthode converge, l'erreur est un infiniment petit d'ordre un en h.
2. Géométriquement, elle revient à remplacer localement en chaque point x_n, la courbe solution par sa tangente. On voit donc qu'au cours du processus numérique, on va passer, à chaque pas d'une courbe solution à une courbe voisine correspondant à une condition initiale légèrement différente. Si le problème est stable, on pourra obtenir la convergence. Par contre, les résultats peuvent vite devenir catastrophiques dans un cas instable (exemple 1.3 du chapitre 1).
3. Enfin, en utilisant l'équivalence entre un problème différentiel et une équation intégrale, on peut interpréter, on l'a vu plus haut, la méthode

d'Euler comme le résultat de l'application de la formule des rectangles basée sur le point x_n

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(t, y(t))dt \Rightarrow y_{n+1} = y_n + hf(x_n, y_n)$$

2.5.4 Méthodes de Runge et Kutta

Dans les méthodes de résolution des problèmes à valeurs initiales, le processus de calcul est un processus fini. On avance de N pas, à partir du temps initial jusqu'au temps final et on s'arrête. Chaque valeur est donc calculée une fois pour toutes. Sauf technique plus complexe d'adaptation de maillages, il n'y a pas de réitération pour améliorer le résultat. Il faudra donc utiliser des méthodes suffisamment précises. Ceci explique le recours à des méthodes d'ordre élevé. Les méthodes de Runge et Kutta sont les généralisations de la méthode d'Euler à des ordres supérieurs à un. Elles s'obtiennent à partir de formules d'intégration numériques plus précises que la formule des rectangles.

Considérons tout d'abord l'utilisation de la formule des trapèzes. Elle conduit à la méthode

$$\begin{cases} y_0 \text{ donné} \\ y_{n+1} = y_n + \dfrac{h}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1})] \end{cases} \tag{2.31}$$

Cette méthode est une méthode implicite. Le calcul de la nouvelle valeur y_{n+1} nécessite la résolution d'une équation. Si l'on veut obtenir une méthode explicite du même ordre, on peut procéder de la manière suivante :

$$\begin{cases} y_0 \text{ donné} \\ {y_{n+1}}^* = y_n + hf(x_n, y_n) \\ y_{n+1} = y_n + \dfrac{h}{2}[f(x_n, y_n) + f(x_{n+1}, {y_{n+1}}^*)] \end{cases} \tag{2.32}$$

Ceci peut s'interpréter comme une itération de point-fixe (limitée ici à un pas) pour résoudre l'équation du schéma implicite des trapèzes (voir plus bas 2.5.8). On obtient ainsi la méthode de **Runge et Kutta d'ordre 2 : RK2**.

De même l'utilisation de la formule d'intégration de Simpson est à la base de la formule de **Runge et Kutta d'ordre 4 : RK4**. C'est l'une des formules les plus utilisées, elle s'écrit :

$$\begin{cases} y_0 \text{ donné, puis pour } n \geq 0 \\ k_1 = hf(x_n, y_n) \\ k_2 = hf(x_n + h/2, y_n + k_1/2) \\ k_3 = hf(x_n + h/2, y_n + k_2/2) \\ k_4 = hf(x_n + h, y_n + k_3) \\ y_{n+1} = y_n + \dfrac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4] \end{cases} \tag{2.33}$$

Pour réduire la complexité en stockage de ce schéma (pas de stockage des coefficients k_i), on peut utiliser le schéma de Runge-Kutta sans stockage suivant :

$$y_{n+1} = y_n + \theta_p h f(y_{n+1}), p = 1...q, \quad \theta_p \in]0, 1]$$

où l'on passe de n à $n+1$ après q sous-itérations, sans stocker les valeurs intermédiaires. Les coefficients θ_p doivent être calés pour réaliser une intégration exacte à un ordre donné pour une fonction f donnée (ce qui est une limitation de cette technique).

Considérons l'équation $y'(x) = \lambda y(x)$, et prenons $q = 2$, le schéma s'écrit alors (on introduit pour la compréhension les v_i intermédiaires) :

$$\begin{cases} v_0 = y_n \\ v_1 = v_0 + h\theta_1 \lambda v_0 \\ v_2 = v_0 + h\theta_2 \lambda v_1 \\ y_{n+1} = v_2 \end{cases}$$

on a donc :

$$y_{n+1} = y_n + h\theta_2 \lambda y^n + h^2 \theta_1 \theta_2 \lambda^2 y_n$$

Or $y''(x) = \lambda^2 y(x)$ d'où par identification : $\theta_2 = 1$ et $\theta_1 \theta_2 = \frac{1}{2}$.

2.5.5 Application aux systèmes différentiels

On peut généraliser simplement l'application des méthodes de Runge et Kutta aux systèmes. Pour un système de deux équations couplées de la forme

$$\begin{cases} y_0, z_0 \text{ donnés} \\ y'(x) = f(x, y(x), z(x)) \\ z'(x) = g(x, y(x), z(x)) \end{cases} \tag{2.34}$$

on obtient l'algorithme suivant pour Runge et Kutta d'ordre 4 :

$$\begin{cases} y_0, z_0 \text{ donnés, puis pour } n \geq 0 & \\ k_1 = hf(x_n, y_n, z_n) & l_1 = hg(x_n, y_n, z_n) \\ k_2 = hf(x_n + h/2, y_n + k_1/2, z_n + l_1/2) & l_2 = hg(x_n + h/2, y_n + k_1/2, z_n + l_1/2) \\ k_3 = hf(x_n + h/2, y_n + k_2/2, z_n + l_2/2) & l_3 = hg(x_n + h/2, y_n + k_2/2, z_n + l_2/2) \\ k_4 = hf(x_n + h, y_n + k_3, y_n + l_3) & l_4 = hg(x_n + h, y_n + k_3, y_n + l_3) \\ y_{n+1} = y_n + \dfrac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4] & z_{n+1} = z_n + \dfrac{1}{6}[l_1 + 2l_2 + 2l_3 + l_4] \end{cases} \tag{2.35}$$

Voici trois applications classiques

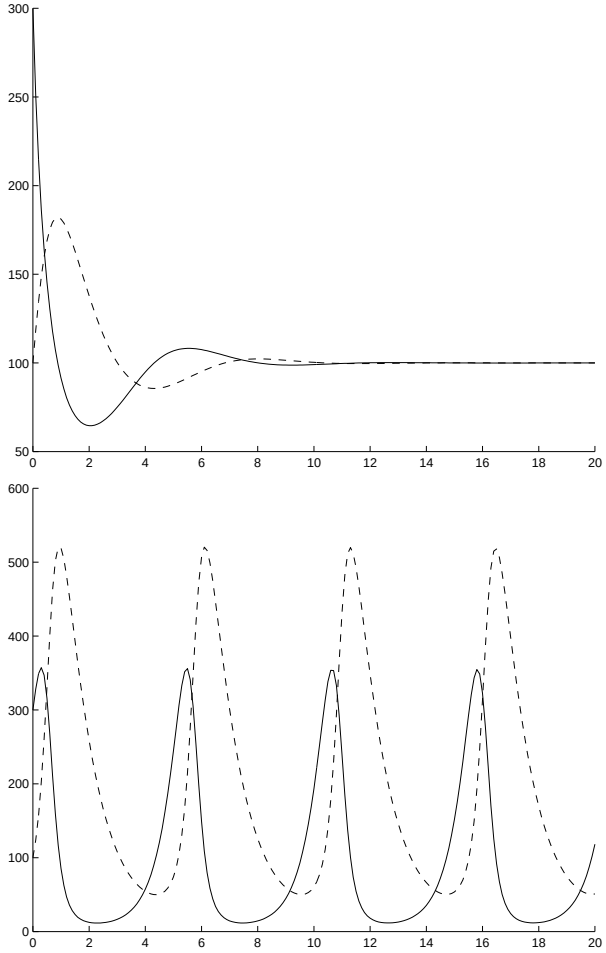


FIGURE 2.12 – Système proie-prédateur : les proies sont représentées en trait continu, les prédateurs en pointillés. En haut le cas $b = 0.01$: proies en milieu fini, en bas le cas $b = 0$: pas de limite au développement de la population des proies, hormis la présence de prédateurs.

1. Le système proies-prédateurs :

$$\begin{cases} y_1' = (\, a - by_1 - cy_2 \,)\, y_1 \\ \\ y_2' = (\, -\alpha + \gamma y_1 \,) y_2 \\ \\ y_1(0) = 300 \quad y_2(0) = 150 \\ \\ a = 2, b = 0.01 \text{ (ou } b = 0), \, c = 0.01, \alpha = 1, \gamma = 0.01 \end{cases} \tag{2.36}$$

dont voici le programme en langage Matlab utilisant le schéma RK2.

```
x=0:0.1:20;
h=0.1;
a= 2;
b=0.01;
c=0.01;
alpha=1;
gamma=0.01;

y(1)=300;
z(1)= 100;

for i=1:200
   k1=h*(a -b*y(i)-c*z(i))*y(i);
   l1=h*(-alpha +gamma*y(i))*z(i);
   k2=h*(a-b*(y(i)+k1) -c*(z(i)+l1))*(y(i)+k1);
   l2=h*(-alpha +gamma*(y(i)+k1))*(z(i)+l1);
   y(i+1)=y(i)+(k1+k2)/2;
   z(i+1)=z(i)+(l1+l2)/2;
end
hold on
plot(x,y)
plot(x,z,'--')
hold off
```

2. L'équation du pendule amorti :

$$\begin{cases} \theta''(t) + \alpha\theta'(t) + k^2\, sin\,(\theta(t)) = 0 \qquad \text{sur} \quad [0, 4\pi] \\ \theta(0) = \dfrac{\pi}{2} \text{ et } \theta'(0) = 0 \end{cases} \tag{2.37}$$

   $k = 5$ et $\alpha = 0.1$

3. Le système dynamique de Lorentz :

$$\dot{x} = \sigma(y - x), \quad \dot{y} = \rho x - y - xz, \quad \dot{z} = xy - \beta z \tag{2.38}$$

   dont les conditions initiales sont précisées dans le programme ci-dessous :

```
! Etude du systeme dynamique de Lorentz
```

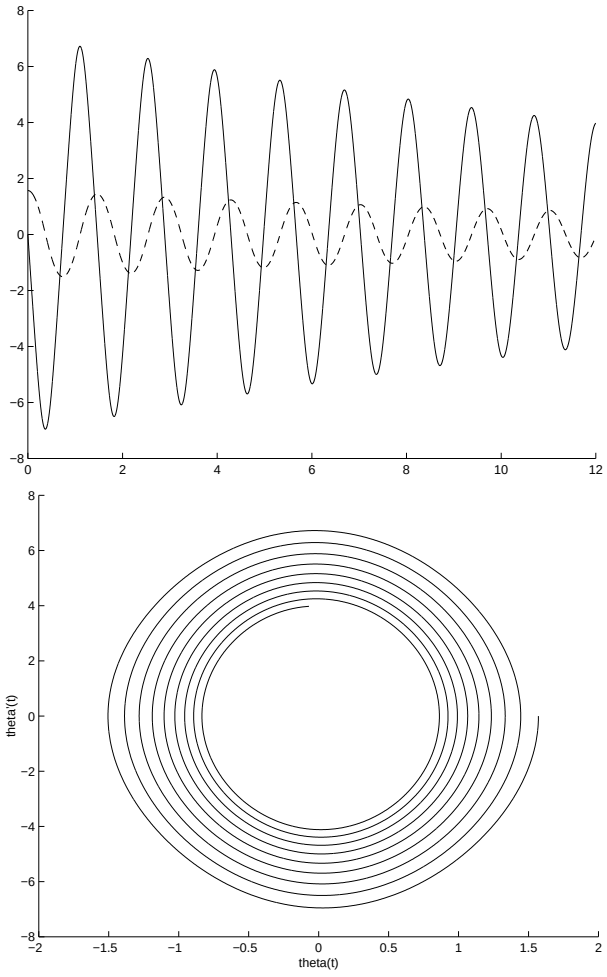


FIGURE 2.13 – Oscillation du pendule amorti : les angles sont représentés en pointillés, leurs dérivées en trait continu. En bas le diagramme de phase, angles / vitesses angulaires

```
x1=x0=-10.;
y1=y0=20.;
z1=z0=-5.;
epsilon=1.e-2  ! perturbation 1 pourcent
sigma=10.*(1.+epsilon)
ro=28.*(1.+epsilon)
beta=2.6667*(1.+epsilon)
irkmax=4  ! Runge et Kutta a 4 pas
dt=1.e-3  ! pas de temps
```

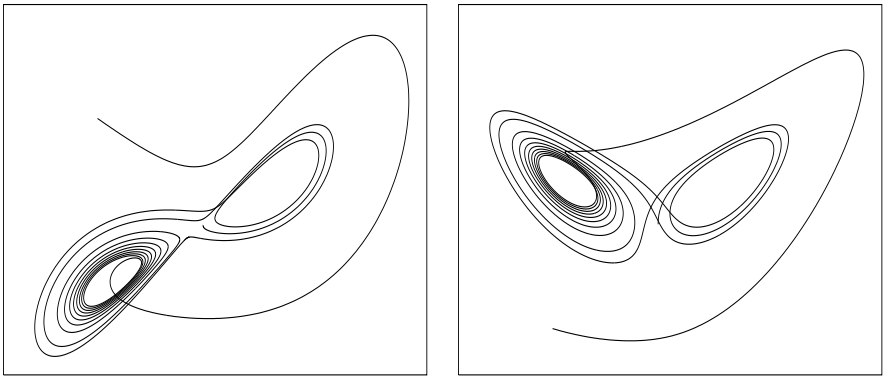


FIGURE 2.14 – À gauche : trajectoire du système dynamique de Lorentz dans le plan (x, y) et à droite en (x, z). Intégration avec un schéma Runge et Kutta sans stockage à 4 pas.

```
do kt=1,10000  ! boucle en temps
  do irk=1,irkmax ! boucle Runge et Kutta sans stockage
    alpha=1./(irkmax+1-irk);
    x1=x0+dt*alpha*(sigma*(y1-x1));
    y1=y0+dt*alpha*(ro*x1-y1-x1*z1);
    z1=z0+dt*alpha*(x1*y1-beta*z1);
  enddo
    x0=x1;  y0=y1;   z0=z1;
enddo
```

Ci-dessus le pas de temps a été fixé a priori, à la suite d'essais numériques. Nous pouvons, par une analyse de stabilité, proposer un critère pour son choix. Cependant, l'analyse de stabilité s'avère souvent difficile pour les systèmes couplés. Dans ce cas, on effectue l'analyse pour chaque équation, en laissant invariantes les autres variables. Le pas de temps sera alors le minimum des pas de temps produits par ces analyses. Par exemple, dans le cas du système de Lorentz on obtient :

$$dt = \min(\frac{2}{\sigma}|\frac{x}{x-y}|,\ \ 2|\frac{y}{-\rho x + y + xz}|,\ \ 2|\frac{z}{\beta z - xy}|)$$

2.5.6 Méthodes à pas multiples

Les méthodes de Runge et Kutta sont des méthodes à un pas. Pour obtenir des méthodes d'ordre de précision élevé, on peut aussi augmenter le nombre de

pas. Dans ce cas, la solution au point $n+1$ sera fonction des solutions aux p pas précédents avec $p > 1$. La construction de ces schémas peut se faire en utilisant la dérivation numérique et la précision du schéma sera donnée par la précision de cette dérivation. Ainsi, pour un schéma à l'ordre 2, on peut combiner les expressions ci-dessous :

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2}y''(x_n) + ...$$

$$y(x_{n-1}) = y(x_n) - hy'(x_n) + \frac{h^2}{2}y''(x_n) + ...$$

pour aboutir au schéma "saute-mouton" (leap-frog en anglais) :

$$\frac{y_{n+1} - y_{n-1}}{2h} = f(x_n, y_n) \tag{2.39}$$

Un autre schéma à deux pas largement utilisé est le schéma implicite d'ordre deux de Gear :

$$\frac{3}{2}y_{n+1} - 2y_n + \frac{1}{2}y_{n-1} = hf(x_{n+1}, y_{n+1}) \tag{2.40}$$

On trouvera dans les ouvrages spécialisés un grand nombre de formules multipas d'ordre élevé (voir cependant la remarque 2.5.1 sur les avantages pratiques des méthodes à un pas de type Runge-Kutta). Remarquons toutefois, qu'un schéma d'ordre élevé n'a d'intérêt que si l'on est assuré de la régularité suffisante de la solution. Pour une solution discontinue ou peu régulière des méthodes d'ordre bas sont mieux adaptées. D'autre part, il y a, en général, sauf évidemment dans le cas de schémas implicites, antagonisme entre ordre élevé et stabilité.

2.5.7 Stabilité

Par opposition à l'ordre de précision qui utilise la solution du problème continu, le concept de **stabilité** est basé sur la solution discrète. Il rend compte du comportement réel de la solution approchée pour une valeur pratique, donc non nulle, du pas h.

Lors d'un calcul réel, les erreurs d'arrondis, inévitables, s'accumulent. Ceci est particulièrement évident dans le processus de résolution d'une équation différentielle où l'on progresse pas à pas à partir d'une valeur initiale. Il existe diverses conditions de stabilité. Tout d'abord la solution numérique doit rester bornée. Cette exigence minimale de stabilité peut se révéler insuffisante dans la pratique, la borne obtenue étant souvent une exponentielle de la durée qui donc croît infiniment lorsque celle-ci augmente.

On introduit alors des critères de stabilité plus exigeants afin que la solution numérique reproduise le comportement physique de la solution exacte. Par

exemple, pour des problèmes dissipatifs, on imposera des conditions de stabilité permettant d'obtenir une solution de norme décroissante. Le concept de A-stabilité, sous sa forme la plus simple, est basé sur l'analyse du comportement, selon les valeurs du pas h, des solutions numériques de l'équation modèle

$$y'(t) = -ay(t) \qquad \text{avec } y(0) \text{ donné et } a \text{ réel} > 0 \tag{2.41}$$

dont la solution exacte est $y(t) = y(0)e^{-at}$.

Étudions le cas du schéma d'Euler explicite. On obtient $y_{n+1} = y_n - ahy_n$, soit

$$y_n = (1 - ah)^n y_0$$

Si l'expression exacte e^{-at} est toujours décroissante en temps et positive, ce n'est pas le cas de l'expression approchée $(1 - ah)^n$, qui selon les valeurs du pas $h > 0$, peut tendre vers l'infini, vers zéro ou prendre alternativement les valeurs 1 et -1. Pour que la solution approchée reproduise le comportement de la solution exacte, donc reste positive et décroissante, il faut imposer une condition de stabilité sur le pas h. On doit avoir $h < \dfrac{1}{a}$. Pour des systèmes différentiels de type $X'(t) + AX(t) = F$ (A est supposée symétrique définie positive, donc à valeurs propres réelles positives), la condition de stabilité imposera, pour toute valeur propre λ_i de la matrice A, les conditions : $h < \dfrac{1}{\lambda_i}$. Donc $h < \dfrac{1}{max(\lambda_i)}$. Si certaines valeurs propres sont grandes, ceci imposera des pas h très petits, et donc des difficultés pour le calcul des solutions sur de longues durées. On dit, dans ce cas, que le système différentiel est *raide* (stiff en anglais).

Étudions maintenant le cas d'un schéma implicite, le schéma d'Euler implicite :

$$y_{n+1} = y_n + f(x_{n+1}, y_{n+1})$$

Son application à (2.41) donne :

$$(1 + ah)^n y_n = y_0 \Longrightarrow y_n = \frac{y_0}{(1 + ah)^n}$$

Dans ce cas, quel que soit $h > 0$, la solution numérique est bornée, positive et décroissante au cours du temps. On dit que le schéma est inconditionnellement stable.

Remarque 2.5.2 *On retrouve ici les approximations stables et instables de la fonction exponentielle présentées au chapitre précédent.*

2.5.8 Point-fixe explicite pour schéma implicite

Il est souvent intéressant pour l'utilisateur de pouvoir spécifier le pas d'intégration sans se soucier des contraintes de stabilité numérique. Ceci implique alors l'utilisation de schémas implicites qui sont inconditionnellement stables. Mais ces schémas ont l'inconvénient de nécessiter la résolution d'une équation à chaque pas. On présente ci-dessous une méthode de point-fixe permettant d'implémenter de façon explicite des méthodes purement implicites.

Considérons le problème de Cauchy suivant :

$$y'(x) = f(x, y(x)), \quad y(0) = y_0$$

Si on le résout par une méthode d'Euler explicite :

$$\frac{y_{n+1} - y_n}{h} = f(x_n, y_n), \quad y(x_0) = y_0, \quad n = 1...N$$

où n désigne l'indice de la solution approchée aux points d'intégration successifs, le pas h devra satisfaire une condition de stabilité de type $h \leq H(f(y))$, H étant obtenu, par exemple, par une analyse de A-stabilité.

L'itération du point-fixe explicite suivante permet de choisir librement h :

$$\left\{ \begin{array}{l} y(x_0) = y_0, \quad \tilde{y}_0 = y_0, \\ n = 0, .., N, \\ \quad m = 0, .., M, \\ \quad\quad \dfrac{\tilde{y}_{m+1} - \tilde{y}_m}{\tilde{h}} + \dfrac{\tilde{y}_{m+1} - y_n}{h} = f(x_{n+1}, \tilde{y}_m) \\ \quad \text{Si} \quad (|\tilde{y}_{m+1} - \tilde{y}_m| \leq \varepsilon) \quad y_{n+1} = \tilde{y}_{m+1} \end{array} \right. \tag{2.42}$$

m désigne les sous-itérations de point-fixe. $\tilde{h}$ est soumis à une condition de stabilité de la forme $\tilde{h} \leq min(h, H(f(y)))$, mais pas h. Ainsi, à chaque pas en n, après convergence de l'itération de point-fixe en m (donc lorsque $|\tilde{y}_{m+1} - \tilde{y}_m| \to 0$), nous réalisons une itération implicite. Bien entendu, pour augmenter la précision de l'intégration en n, on peut utiliser un schéma de type Runge et Kutta au lieu de la méthode d'Euler explicite.

Cette approche est très utile en couplage de modèles, comme nous le verrons dans le chapitre 13. En effet, une implicitation totale est parfois impossible lorsque les équations sont résolues par des solveurs boîte-noire.

Un autre intérêt de cette approche est la possibilité de l'utilisation de pas d'intégration locaux non-physiques pour $\tilde{h}$. Ce pas de temps est donné par la condition de stabilité en chaque point. Ceci permet une accélération de la convergence des itérations de point-fixe.

2.6 Problèmes à valeurs aux limites

Nous venons de voir comment résoudre les problèmes de Cauchy ou à valeurs initiales dans lesquels les conditions sont imposées en un point ou un temps donné. Il existe d'autres types de problèmes différentiels où les conditions sont imposées en des points distincts, et en particulier aux limites du domaine de résolution. Ce type de problème est très courant. Voici, par exemple, un problème de Poisson en dimension un avec conditions de Dirichlet :

$$\begin{cases} -u''(x) = f(x) \\ u(0) = 1, \quad u(L) = 2 \end{cases} \tag{2.43}$$

C'est un problème aux limites à travers ses conditions aux points 0 et L.

De même, l'équation des ondes ci-dessous :

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - c^2 \dfrac{\partial^2 u}{\partial x^2} = 0.001x(1-x)\sin(10x) \\ u(t, x=0) = 1, \quad u(t, x=1) = 2 \\ u(t=0, x) = x+1, \quad u(t=1, x) = 1 + x + 5x(1-x) \end{cases} \tag{2.44}$$

fait intervenir des conditions aux limites en espace et en temps. On constate que par rapport aux problèmes de Cauchy, où les conditions en temps sont exprimées au même instant, elles sont ici imposées à deux instants différents.

La résolution de ces problèmes ne fait pas appel aux mêmes techniques. Dans le premier cas, nous utiliserons, par exemple, les différences finies (voir chapitre 4) sur une discrétisation en $N+2$ points du domaine $[0, L]$:

$$\begin{cases} -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} = f(x_j) \quad \text{pour } j = 1...N \\ \text{avec} \quad u_0 = 1, \quad u_{N+1} = 2 \end{cases} \tag{2.45}$$

Cette approche aboutit à la résolution d'un système linéaire tridiagonal $(N \times N)$ pour trouver la solution $u_1, u_2, ..., u_N$ aux points de discrétisation $x_1, x_2, ..., x_N$.

Pour le deuxième problème, la même approche s'applique naturellement pour le problème aux limites spatial, mais la complexité en stockage la rend inutilisable pour prendre en compte les conditions aux limites temporelles. En effet, considérons une discrétisation explicite de l'équation des ondes (voir chapitre 11) :

$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} - c^2 \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = f(x_j) \tag{2.46}$$

Ici, au contraire d'un problème à valeurs initiales, où l'on dispose de deux états u^0 et u^1 pour démarrer l'itération, on ne dispose que de u^0 et de u^M avec $M\Delta t = 1$.

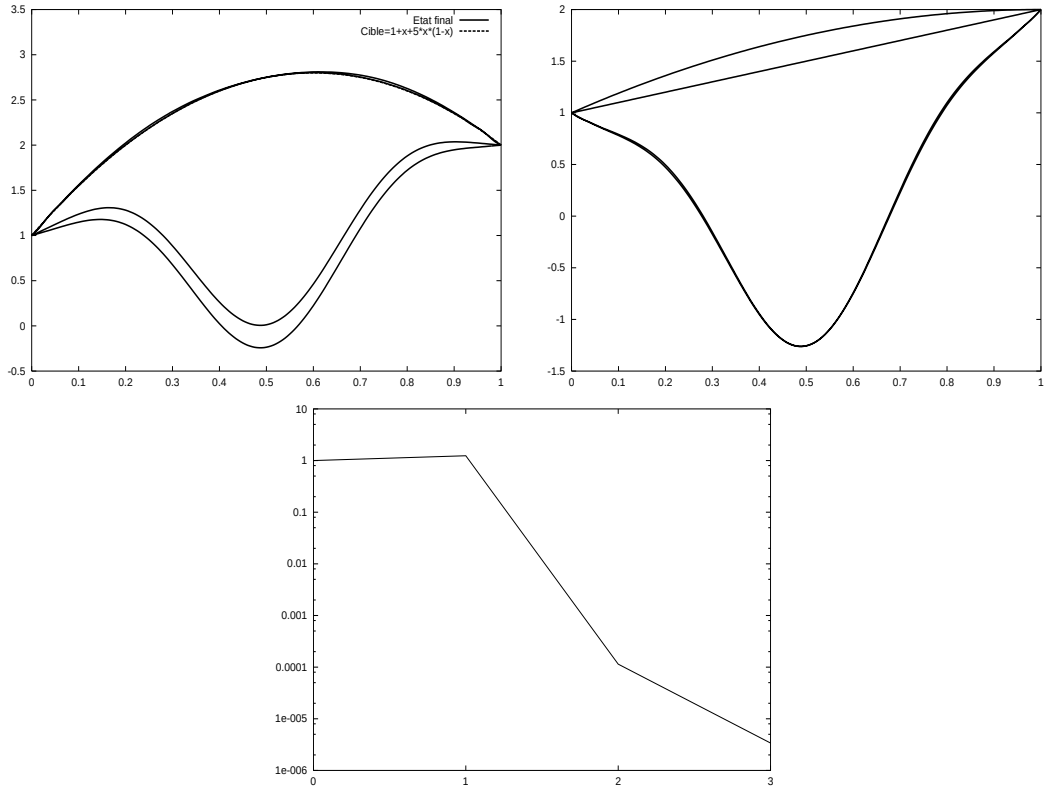


FIGURE 2.15 – Résolution du problème aux limites (2.44) avec le schéma itératif ci-dessus. En haut à gauche : évolution de u^M pour différents u^1, on atteint après quatre essais l'état désiré. En haut à droite : évolution de u^1. En bas : historique de (2.48).

On est alors amené à résoudre un système linéaire de taille, non plus $(N \times N)$, mais $(NM \times NM)$. Ce qui est beaucoup plus cher. Et la situation s'aggrave si la dimension d'espace est plus élevée.

On utilise alors l'une des méthodes itératives de résolution d'équations présentées plus haut pour trouver le zéro de la fonction $h(u^1) = (1 + x + 5x(1 - x) - u^M)$. Le but est la détermination de la bonne condition u^1 pour atteindre la valeur limite souhaitée en $t = 1$. L'itération suivante, appelée *méthode de tir*, est basée

sur l'utilisation de la méthode de fausse-position (2.3) :

$$\begin{cases} g_1(x) \quad \text{et} \quad g_2(x) \quad \text{données,} \\ \text{Faire :} \\ \quad \text{Calculer } u^M, \text{ solution de (2.46) avec } u^0 = x+1 \text{ et } u^1 = g_p, \\ \quad \text{Si} \quad p \geq 2, \quad g_{p+1} = g_p + \dfrac{h(g_p)(g_p - g_{p-1})}{u_p^M - u_{p-1}^M}, \\ \text{Tant que} \quad ||h(g_p)|| > TOL, \quad p \leftarrow p+1. \end{cases} \tag{2.47}$$

La norme utilisée ci-dessus peut être, par exemple, une norme L^2 discrète qui mesure la distance entre la condition finale réalisée et celle souhaitée sur les points de la discrétisation spatiale, au temps $t = 1$:

$$||h(u^1)|| = (\sum_j ((h(u^1)_j)^2)^{1/2} \tag{2.48}$$

Remarque 2.6.1 *La convergence de l'algorithme de tir ci-dessus est possible uniquement si la fonction h n'est pas trop sensible aux valeurs de* u^1.

Remarque 2.6.2 *Nous montrerons, chapitre 17, le lien entre les problèmes à valeurs aux limites et les problèmes d'optimisation.*

2.7 Méthodes de résolution des systèmes linéaires

Les systèmes linéaires sur-déterminés (plus grand nombre d'équations que d'inconnues) sont résolus, en général, en utilisant les techniques de moindres carrés. Nous nous limitons donc ici au cas des sytèmes carrés comportant autant d'équations que d'inconnues.

On considère le système suivant :

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + ...a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + ...a_{2n}x_n = b_2 \\ \qquad\qquad ..... \\ \qquad\qquad ..... \\ \qquad\qquad ..... \\ a_{n1}x_1 + a_{n2}x_2 + ...a_{nn}x_n = b_n \end{cases}$$

qui s'écrit matriciellement

$$AX = B$$

Soit ϕ l'application linéaire représentée par la matrice A, si X représente x et B, b, $AX = B$ correspond à $\phi(x) = b$.

2.7.1 Existence et unicité de la solution

Théorème 2.7.1 (Systèmes de Cramer) *Une condition nécessaire et suffisante pour qu'un système linéaire de n équations à n inconnues admette une solution unique pour tout second membre $B \in \mathbb{R}^n$ est de manière équivalente que*

- $det(A) \neq 0$
- $Ker(\phi) = \{0\}$
- $rg(\phi) = dim(Im(\phi)) = n$
- *les vecteurs lignes ou les vecteurs colonnes de A sont indépendants*
- *la seule solution de $AX = 0$ est $X = 0$*

2.7.2 Méthodes directes

Les méthodes directes de résolution des systèmes linéaires sont des méthodes dans lesquelles la solution est obtenue de façon exacte en un nombre fini d'opérations. De façon exacte s'entend, sur un ordinateur, aux erreurs d'arrondis machine près. Le prototype de méthode directe est la méthode du pivot de Gauss. Cette méthode permet de ramener la résolution d'un système général à la résolution d'un système triangulaire supérieur. La triangularisation de Gauss consiste à annuler, par étape, colonne par colonne, les coefficients sous-diagonaux de la matrice par combinaison des lignes avec une ligne de référence ou ligne

"pivot". À la première étape, la ligne pivot est la première ligne, puis la k^e à l'étape k et ainsi de suite jusqu'à l'étape $n-1$. Le coefficient diagonal $A_{k,k}$ de la ligne de référence s'appelle le pivot. L'élimination se fait selon

$$A_{i,j}^{(k+1)} = A_{i,j}^{(k)} - \frac{A_{i,k}^{(k)}}{A_{k,k}^{(k)}} A_{k,j}^{(k)}$$

pour $k = 1...N-1,\ i = k+1...N$ et $j = k+1...N$, sans oublier de combiner de la même façon les composantes du second-membre. La rencontre de pivot nul peut nécessiter la permutation de lignes du système. Cependant pour certaines classes de matrices, en particulier les matrices symétriques définies positives, on est assuré de pouvoir trianguläriser le système par Gauss sans permutation. La méthode du pivot équivaut alors à une factorisation de type

$$A = LU$$

de la matrice A. L est une matrice triangulaire inférieure à diagonale unité et U une matrice triangulaire supérieure. Une fois obtenu le système triangulaire supérieur équivalent au système initial, sa résolution se fait ensuite explicitement par un processus de remontée. On commence par calculer la dernière composante du vecteur inconnu en utilisant la dernière équation et on remonte équation par équation en déterminant les composantes correspondantes.

Dans le cas d'une matrice A symétrique, on peut utiliser la symétrie pour obtenir une factorisation de Crout :

$$A = LDL^T$$

avec D diagonale.

Dans le cas d'une matrice A symétrique définie positive, la méthode de Choleski conduit à une factorisation :

$$A = LL^T$$

On trouve la matrice L, qui cette fois n'est plus à diagonale unité, par un algorithme d'identification de coefficients.

De

$$A_{ij} = \sum_{k=1,i} L_{ik} L_{jk} \quad \text{pour} \quad j \leq i$$

on déduit pour tout i :

$$L_{ii} = \sqrt{A_{ii} - \sum_{k=1,i-1} L_{ik}^2}$$

et pour tout $j < i$

$$L_{ij} = \frac{(A_{ij} - \sum_{k=1,j-1} L_{ik} L_{jk})}{L_{jj}}$$

Les méthodes directes présentent l'avantage de fournir la solution exacte (aux erreurs d'arrondis machine près) en un nombre fini d'opérations (de l'ordre de $\frac{n^3}{3}$ pour Gauss). La méthode du pivot de Gauss s'applique à tout système inversible. Par contre les méthodes directes ont un coût important en stockage mémoire (bien que des techniques de stockage minimal associées à des algorithmes de numérotation optimale des inconnues permettent de le réduire sensiblement). Ceci rend leur application pratiquement impossible, en l'état actuel de la technologie, pour de gros systèmes à plus de 10^5 inconnues, et donc en particulier pour la résolution de problèmes industriels en dimension 3 d'espace.

Nous renvoyons à la littérature pour plus de précisions sur les méthodes directes (voir Lascaux-Théodor, Ciarlet, Lucquin-Pironneau).

2.7.3 Méthodes itératives

Le principe général des méthodes itératives est le suivant. Le vecteur solution du système est obtenu comme limite (quand elle existe) d'une suite itérative de vecteurs définie par une récurrence linéaire de la forme :

$$\begin{cases} X^{(0)} \quad \text{donné} \\ \\ X^{(k+1)} = M\ X^{(k)} + C \end{cases} \tag{2.49}$$

où M est une matrice $n \times n$ dite matrice d'itération et C un vecteur de $\mathbb{R}^n$.

À convergence on a

$$X = MX + C$$

donc cette équation de type point-fixe doit évidemment être équivalente à l'équation initiale

$$AX = B$$

L'utilisation pratique des méthodes itératives nécessite, non seulement la convergence de la méthode, mais aussi que cette convergence ne soit pas trop lente. Il est souvent nécessaire pour accélerer la convergence d'utiliser des techniques de préconditionnement (voir la définition 2.7.3 du conditionnement). Malheureusement les bons préconditionneurs ne sont pas toujours faciles à trouver. En conclusion, il faut, à notre avis, utiliser les méthodes directes dès que cela est possible.

2.7.4 Conditions de convergence

On aura convergence de la suite vectorielle $\{X^{(k)}\}$ vers la solution X à condition que pour une norme vectorielle on ait

$$\lim_{k\to\infty} \|X^{(k)} - X\| = 0$$

On peut également introduire le vecteur **résidu**

$$R^{(k)} = B - AX^{(k)} = A(X - X^{(k)})$$

On aura de manière équivalente (A est inversible) convergence des itérations vers la solution si

$$\lim_{k\to\infty} \|R^{(k)}\| = 0$$

Définition 2.7.1 (Norme matricielle) *On appelle norme matricielle induite par une norme vectorielle $\|.\|$, l'application de l'espace des matrices dans $\mathbb{R}^+$ définie pour toute matrice A par*

$$\|A\| = \max_{X\neq 0} \frac{\|AX\|}{\|X\|} \tag{2.50}$$

Définition 2.7.2 (Rayon spectral) *On appelle rayon spectral d'une matrice A le nombre positif*

$$\rho(A) = \max_i |\lambda_i| \tag{2.51}$$

où les λ_i sont les valeurs propres de la matrice A.

Théorème 2.7.2 *Si A est symétrique réelle, son rayon spectral $\rho(A)$ est égal à sa norme induite par la norme vectorielle euclidienne.*

Démonstration : Une matrice symétrique réelle admet une base orthonormée de vecteurs propres. Soit X un vecteur de $\mathbb{R}^n$ de composantes x_i dans la base des vecteurs propres $\{V^i\}$ de A. On aura alors :

$$AX = \sum_{i=1,n} \lambda_i x_i V^i \quad \text{et donc} \quad \|AX\| = (\sum_{i=1,n} \lambda_i^2 x_i^2)^{\frac{1}{2}}$$

On en déduit

$$\|A\| = \max_{X\neq 0} \frac{\|AX\|}{\|X\|} = \max_i |\lambda_i| = \rho(A)$$

Théorème 2.7.3 *L'itération*

$$\begin{cases} X^{(0)} \quad donné \\ \\ X^{(k+1)} = M\ X^{(k)} + C \end{cases}$$

converge vers la solution X si pour une norme matricielle donnée

$$\|M\| < 1$$

Démonstration
De $X^{(k+1)} = M\ X^{(k)} + C \quad$ et $\quad X = M\ X + C$,
on déduit : $X - X^{(k+1)} = M\ (X - X^{(k)}) \quad$ soit $\quad X - X^{(k)} = M^k\ (X - X^{(0)})$
Donc $\|X - X^{(k)}\| = \|M^k\ (X - X^{(0)})\| \leq \|M\|^k \|X - X^{(0)}\|$

D'où la convergence de la suite $\{X^k\}$ vers la solution X dès que $\|M\| < 1$.

Remarque 2.7.1 *Dans le cas d'une matrice M symétrique on aura donc convergence si*

$$\rho(M) < 1 \tag{2.52}$$

On admettra que ce résultat est vrai quelle que soit la matrice d'itération M.

On voit que le problème de l'estimation des valeurs propres de la matrice d'itération est crucial pour l'étude de la convergence des méthodes. Le théorème suivant est parfois un outil commode.

Théorème 2.7.4 (de Gershgörin-Hadamard) *Soit une matrice quelconque A. Si λ est valeur propre, réelle ou complexe, de A, il existe un indice de ligne i tel que*

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \tag{2.53}$$

Autrement dit, toutes les valeurs propres de la matrice A appartiennent à l'union des disques du plan complexe de centre a_{ii} et de rayon $\sum_{j \neq i} |a_{ij}|$

Démonstration. Soit v un vecteur propre de A associé à la valeur propre λ. Supposons que v_i soit la composante maximale de v. Le vecteur $w = \dfrac{v}{v_i}$ est également vecteur propre de A associé à λ, et vérifie, par construction, $w_i = 1$ et $|w_j| \leq 1 \quad \forall j \neq i$.
De $Aw = \lambda w$, on déduit : $\sum_j a_{ij} w_j = \lambda_i w_i$
et donc $\sum_{j \neq i} a_{ij} w_j = (\lambda_i - a_{ii}) w_i$
D'où : $|\lambda_i - a_{ii}||w_i| \leq \sum_{j \neq i} |a_{ij}||w_j|$
Et le résultat en utilisant $w_i = 1$ et $|w_j| \leq 1 \quad \forall j \neq i$

2.7.5 Méthode de Jacobi

Soit

$$AX = B$$

le système linéaire à résoudre. La méthode de Jacobi correspond à la décomposition de la matrice A sous la forme

$$A = D - E - F \quad \text{ou} \quad A = D + A - D$$

avec D matrice diagonale constituée des éléments diagonaux a_{ii} de A,

$-E$ matrice triangulaire inférieure stricte, constituée des éléments strictements sous-diagonaux de A : a_{ij} pour $i > j$, et

$-F$ matrice triangulaire supérieure stricte, constituée des éléments strictements sur-diagonaux de A : a_{ij} pour $i < j$. On définit alors la méthode de Jacobi comme la méthode itérative :

$$\begin{cases} X^{(0)} \quad \text{donné} \\ \\ X^{(k+1)} = D^{-1}(E+F)\ X^{(k)} + D^{-1}B \end{cases} \tag{2.54}$$

ou ce qui revient au même :

$$\begin{cases} X^{(0)} \quad \text{donné} \\ \\ X^{(k+1)} = [I - D^{-1}A]\ X^{(k)} + D^{-1}B \end{cases}$$

La matrice d'itération de Jacobi est donc

$$J = D^{-1}(E+F) = I - D^{-1}A$$

et la condition de convergence s'exprime par

$$\rho(J) < 1$$

On observe que la méthode de Jacobi correspond à l'écriture ligne par ligne suivante :

$$x_i^{k+1} = \frac{b_i - \sum_{j \neq i} a_{ij} x_j^k}{a_{ii}} \tag{2.55}$$

La méthode de Jacobi n'utilise donc pas les informations les plus récentes sur les composantes des vecteurs itérés. En conséquence, c'est une méthode qui converge lentement. Par contre, c'est une méthode adaptée à la vectorisation, car à chaque itération on peut calculer les composantes indépendamment les unes des autres.

2.7.6 Méthode de Gauss-Seidel ou de relaxation

Soit encore

$$AX = B$$

le système linéaire à résoudre. La méthode de Gauss-Seidel correspond aussi à la décomposition de la matrice A sous la forme

$$A = D - E - F$$

avec D matrice diagonale constituée des éléments diagonaux a_{ii} de A,

$-E$, matrice triangulaire inférieure stricte, constituée des éléments strictements sous-diagonaux de A : a_{ij} pour $i > j$, et

$-F$, matrice triangulaire supérieure stricte, constituée des éléments strictements sur-diagonaux de A : a_{ij} pour $i < j$.

Mais on définit cette fois la méthode de Gauss-Seidel comme la méthode itérative :

$$\begin{cases} X^{(0)} \quad \text{donné} \\ \\ X^{(k+1)} = (D-E)^{-1}F \; X^{(k)} + (D-E)^{-1}B \end{cases} \tag{2.56}$$

La matrice d'itération de Gauss-Seidel est donc

$$\mathcal{L} = (D-E)^{-1}F$$

et la condition de convergence s'exprime par

$$\rho(\mathcal{L}) < 1$$

On observe que la méthode de Gauss-Seidel correspond à l'écriture ligne par ligne suivante :

$$x_i^{k+1} = \frac{b_i - \sum_{j<i} a_{ij} x_j^{k+1} - \sum_{j>i} a_{ij} x_j^k}{a_{ii}} \tag{2.57}$$

Elle utilise les informations les plus récentes sur les composantes des itérés. La méthode de Gauss-Seidel converge en général plus vite que la méthode de Jacobi. Par contre, Gauss-Seidel n'est pas adaptée à la vectorisation.

Les méthodes de **relaxation** sont des extrapolations de la méthode de Gauss-Seidel dépendant d'un paramètre ω. Ce paramètre est déterminé pour obtenir une convergence plus rapide. Les méthodes de relaxation s'écrivent :

$$\begin{cases} X^{(0)} \quad \text{donné} \\ \\ X^{(k+1)} = (D-\omega E)^{-1}(\omega F + (1-\omega)D) \; X^{(k)} + \omega (D-\omega E)^{-1}B \end{cases} \tag{2.58}$$

La méthode de relaxation s'écrit ligne par ligne comme une extrapolation de la composante obtenue par Gauss-Seidel. On a :

$$X_i^{k+1}(relax) = \omega\, X_i^{k+1}(GS) + (1-\omega)\, X_i^k(relax)$$

La méthode de Gauss-Seidel correspond au cas $\omega = 1$. Dans le cas de matrices A symétriques définies positives, les méthodes de relaxation convergent pour $0 < \omega < 2$.

2.7.7 Méthodes de descente - Méthode du gradient

Les méthodes de descente sont des méthodes itératives qui utilisent l'équivalence entre les problèmes suivants (voir annexe A) :

$$\begin{cases} \text{Trouver } X \in I\!R^N \quad \text{tel que} \\ \\ A\,X = B \end{cases} \tag{2.59}$$

et

$$\begin{cases} \text{Trouver } X \in I\!R^N \quad \text{tel que} \\ \\ J(X) = \dfrac{1}{2}(A\,X, X) - (B, X) \;\; \text{soit minimal} \end{cases} \tag{2.60}$$

et qui sont donc limitées au cas des systèmes linéaires dont la matrice A est symétrique définie positive. Elles se généralisent, par contre, au cas non-linéaire de la minimisation de fonctionnelles strictement convexes quelconques (voir chapitre 17).

Les méthodes de descente sont basées sur le calcul de la solution comme limite d'une suite minimisante de la forme quadratique J. Cette suite est construite comme une suite récurrente :

$$\begin{cases} X^{(0)} \quad \text{donné} \\ \\ X^{(k+1)} = X^{(k)} - \alpha_k\, d^{(k)} \end{cases} \tag{2.61}$$

avec $d^{(k)} \in I\!R^N$ vecteur donnant la direction de descente à l'étape k et $\alpha_k \in I\!R$ coefficient déterminé de manière à minimiser J dans la direction $d^{(k)}$.

$$J(X^{(k)} - \alpha_k\, d^{(k)}) \leq J(X^{(k)} - \alpha\, d^{(k)}) \quad \forall \alpha \in I\!R$$

Après développement, on obtient α_k comme valeur annulant la dérivée de J par rapport à α, soit :

$$\alpha_k = \frac{(G^{(k)}, d^{(k)})}{(Ad^{(k)}, d^{(k)})}$$

avec

$$G^{(k)} = grad(J(X^{(k)})) = AX^{(k)} - B$$

On peut montrer que la méthode de Gauss-Seidel est la méthode de descente correspondant aux choix des axes de coordonnées comme directions successives de descente. Dans le cas de la **méthode du gradient**, on choisit comme direction de descente la direction du vecteur gradient de J au point $X^{(k)}$. Cette direction est la direction de variation maximale de J. On dit encore direction de plus profonde descente, d'où le nom :"steepest descent method" , employé dans certains ouvrages en anglais.

L'itération de la méthode du gradient s'écrit :

$$\begin{cases} X^{(0)} \quad \text{donné} \\ \\ X^{(k+1)} = X^{(k)} - \alpha_k \, G^{(k)} \end{cases} \tag{2.62}$$

avec

$$\alpha_k = \frac{\|G^{(k)}\|^2}{(AG^{(k)}, G^{(k)})}$$

Vitesse de convergence. Conditionnement.

À partir de $G^{(k+1)} = AX^{(k+1)} - B$ et de $X^{(k+1)} = X^{(k)} - \alpha_k \, G^{(k)}$, on obtient la récurrence suivante sur les gradients :

$$G^{(k+1)} = G^{(k)} - \alpha_k \, AG^{(k)}$$

On en déduit $\|G^{(k+1)}\|^2 = \|G^{(k)}\|^2 - 2\alpha_k(AG^{(k)}, G^{(k)}) + \alpha_k^2 \|AG^{(k)}\|^2$

$$\text{avec} \quad \alpha_k = \frac{\|G^{(k)}\|^2}{(AG^{(k)}, G^{(k)})}$$

on obtient :

$$\|G^{(k+1)}\|^2 = \|G^{(k)}\|^2 \big[\frac{\|G^{(k)}\|^2 \|AG^{(k)}\|^2}{(AG^{(k)}, G^{(k)})^2} - 1\big]$$

Et en utilisant les valeurs propres et les vecteurs propres de la matrice A supposée définie positive, on déduit :

$$\|G^{(k+1)}\|^2 \leq \|G^{(k)}\|^2 \big[\frac{\lambda_{max}}{\lambda_{min}} - 1\big]$$

Définition 2.7.3 *Pour une matrice symétrique réelle le nombre*

$$K(A) = \frac{\lambda_{max}}{\lambda_{min}}$$

rapport des plus grandes et plus petites valeurs propres de A est appelé nombre de conditionnement de la matrice A.

Plus il est proche de 1, plus vite la méthode du gradient converge. Les techniques de préconditionnement ont, entre autres, pour but de rapprocher les valeurs propres extrêmes afin d'accélerer la convergence des méthodes itératives.

Remarque 2.7.2 *Le nombre de conditionnement est, en particulier, égal à 1 pour la matrice identité. D'ailleurs, dans ce cas, l'expression à minimiser décrirait, en dimension deux, un ensemble de cercles concentriques. Les gradients ayant la direction d'un rayon, on obtiendrait la solution en une itération. Cette remarque est à la base de l'idée de la méthode du gradient conjugué.*

2.7.8 Méthode du gradient conjugué

Dans le cas d'une matrice symétrique définie positive quelconque A, les iso-J sont des hyper-ellipses. Elles redeviennent des cercles pour le produit scalaire définie par A :

$$X, Y \longrightarrow (X, Y)_A = (AX, Y)$$

D'où l'idée de remplacer les directions de descente dans la méthode du gradient, par des directions conjuguées, c'est à dire orthogonales au sens du produit scalaire $(.,.)_A$.

On obtient alors l'algorithme :

$$\begin{cases} X^0 \in I\!R^N \quad \text{donné} \\ d^0 = G^0 = AX^0 - B \\ \text{puis pour } k = 0, 1... \\ \alpha_k = \dfrac{\|G^k\|^2}{(Ad^k, d^k)} \\ X^{k+1} = X^k - \alpha_k\, d^k \\ G^{k+1} = G^k - \alpha_k\, Ad^k \\ \beta_{k+1} = \dfrac{\|G^{k+1}\|^2}{\|G^k\|^2} \\ d^{k+1} = G^{k+1} + \beta_{k+1}\, d^k \end{cases} \tag{2.63}$$

Nous renvoyons à la littérature pour une étude exhaustive de la méthode du gradient conjugué. Retenons que la propriété de A-orthogonalité entraîne une convergence théorique en N itérations. Ce qui en ferait une méthode directe. Cependant, elle serait alors plus chère que Choleski. De plus, les erreurs d'arrondis font que les propriétés d'orthogonalité ne sont pas vérifiées exactement. Il faut donc considérer le gradient conjugué comme une méthode itérative. On montre que sa vitesse de convergence dépend également du conditionnement de la matrice A. On est conduit à préconditionner A pour obtenir des performances intéressantes. Parmi les préconditionnements classiques, on citera le préconditionnement SSOR (O. Axelsson) et par Choleski incomplet (Meijerink et Van der Vorst).

Dans le cas où la matrice du système n'est pas symétrique définie positive, il est plus difficile d'obtenir des méthodes itératives performantes. Mentionnons la méthode GMRES (Saad et Schultz), dont il existe également une version pour la résolution de systèmes non-linéaires.

2.7.9 Application des méthodes de gradient au cas non-linéaire

Dans le cas plus général de minimisation d'une fonctionnelle J strictement convexe non nécessairement quadratique, le gradient ∇J est non-linéaire. Cependant les méthodes de gradient peuvent s'appliquer (on peut même les appliquer sans garantie de convergence, car dans tous les cas on réduira J, voir chapitre 17). La détermination du pas optimal α_k à l'itération k se fait alors par une méthode de recherche de l'argument α_k qui minimise la fonction $J(X^k - \alpha \nabla J(X^k))$. On est ramené à un problème en dimension un pour lequel diverses techniques existent, en particulier les algorithmes de recherche de type section dorée.

L'extension de la méthode du gradient conjugué au cas non-linéaire nécessite, de plus, la définition des directions de descente "conjuguées". L'algorithme le plus efficace est celui de Polak-Ribière. Les directions de descente successives sont données par

$$d^{k+1} = \nabla J(X^{k+1}) + \beta_{k+1}\, d^k \tag{2.64}$$

avec cette fois

$$\beta_{k+1} = \frac{(\nabla J(X^{k+1}), \nabla J(X^{k+1}) - \nabla J(X^k))}{\|\nabla J(X^k)\|^2}$$

2.8 Calcul des valeurs et vecteurs propres

Les méthodes efficaces de calcul des éléments propres d'une matrice, dont on verra plus loin des applications, en vibrations de structures, par exemple, utilisent

toujours des techniques itératives. Commençons par exposer la méthode la plus simple, la méthode de la **puissance**.

2.8.1 La méthode de la puissance

La méthode de la puissance est basée sur le fait suivant. Si la matrice A admet une seule valeur propre de plus grand module, l'itération vectorielle

$$V^{k+1} = \frac{AV^k}{\|AV^k\|}$$

lorsqu'elle converge, admet comme limite un vecteur propre de la matrice A associé à cette valeur propre. La division par la norme a pour but d'éviter d'atteindre des valeurs trop grandes (ou trop petites) des composantes lors des itérations vectorielles. Un cas pratique important où l'on peut démontrer la convergence est celui des matrices symétriques réelles dont la valeur propre de plus grand module est unique. En effet, une matrice symétrique réelle admet une base de vecteurs propres orthonormés $\{e_i\}$. Considérons l'initialisation :

$$V^0 = \sum_{i=1}^{N} \alpha_i e_i$$

et supposons que e_1 soit vecteur propre associé à la valeur propre λ_1 de plus grand module. Il est alors facile d'obtenir :

$$V^{k+1} = \frac{\sum_{i=1}^{N} \lambda_i^{k+1} \alpha_i e_i}{(\sum_{i=1}^{N} (\lambda_i^{k+1} \alpha_i)^2)^{\frac{1}{2}}}$$

On en déduit :

$$\lim_{k\to\infty} V^k = \frac{\lambda_1^{k+1}\alpha_1}{|\lambda_1^{k+1}\alpha_1|}\, e_1 = \pm e_1$$

La valeur propre de module maximal s'obtenant simplement par

$$\lambda_1 = \lim_{k\to\infty} \frac{(AV^k, V^k)}{(V^k, V^k)}$$

Les limites de cette méthode sont les suivantes :

— elle ne converge efficacement que dans le cas de valeurs propres simples de plus grand module. La vitesse de convergence dépend de la vitesse à laquelle les rapports $(\frac{\lambda_i}{\lambda_1})^k$ tendent vers zéro quand k augmente. La convergence sera donc d'autant plus rapide que l'écart entre la plus grande valeur propre et les suivantes est grand. Dans le cas de valeurs propres

multiples, la convergence est ralentie. Dans le cas de valeurs propres distinctes, mais de module égal (valeurs propres opposées ou complexes conjuguées) l'algorithme se complique, on ne peut obtenir directement les valeurs propres et les vecteurs propres associés ;
— elle ne fournit qu'une valeur propre. Les techniques de "déflation", pour en déduire les suivantes, sont peu efficaces et trop sensibles aux erreurs d'arrondis.

Voici ses avantages :

— elle est facile à implémenter ;
— elle permet de calculer efficacement les vecteurs propres et d'améliorer sensiblement la précision des valeurs propres, si l'on a obtenu, par une autre méthode, des approximation des valeurs propres ;
— enfin, il est facile de l'adapter à la recherche d'une autre valeur propre que la plus grande, comme on va le voir plus loin.

Remarque 2.8.1 *On pourrait penser que le cas $\alpha_1 = 0$, toujours possible théoriquement si l'on choisit, par hasard, un vecteur initial V^0 orthogonal au vecteur propre e_1 recherché, pose des problèmes. En réalité, voici une situation où paradoxalement, les erreurs d'arrondis sont bénéfiques. Dans la pratique la composante α_1 sera toujours non-nulle. Et même si elle est faible, l'itération finira par converger vers le vecteur propre e_1 associé à la valeur propre de plus grand module.*

Méthode de la puissance inverse

Il est clair que si l'on peut obtenir la plus grande valeur propre, on peut obtenir la plus petite (et le vecteur propre associé) à partir d'une itération utilisant la matrice inverse. On sait, en effet, que les valeurs propres de A^{-1} sont les inverses des valeurs propres de A. Ceci conduit à l'itération

$$\left\{\begin{array}{l} V^0 \text{ donné,} \\ \text{ puis pour } k = 0, 1, ... \\ A\,\tilde{V}^{(k+1)} = V^{(k)} \\ \mu_{k+1} = |\tilde{V}^{(k+1)}| \\ V^{(k+1)} = \frac{\tilde{V}^{(k+1)}}{\mu_{k+1}} \end{array}\right. \tag{2.65}$$

Les $V^{(k)}$ convergent vers le vecteur propre recherché, la plus petite valeur propre en module de A s'obtient comme $\lim_{k\to\infty} \dfrac{1}{\mu_k}$. La matrice A est en général factorisée une fois pour toutes, la résolution du système se réduit à celle de deux systèmes triangulaires.

Remarque 2.8.2 *Dans les applications aux vibrations, la plus petite valeur propre correspond au mode fondamental. C'est donc souvent celle que l'on recherche. Il arrive également que l'on soit intéressé par les modes de fréquences proches d'une fréquence donnée. C'est en particulier le cas lorsque l'on cherche à éviter des phénomènes de résonance. Dans ce cas la valeur propre recherchée est la plus proche d'une valeur μ donnée. Il est facile d'utiliser alors la méthode de la puissance inverse sur la matrice décalée $A - \mu I$, dont la plus petite valeur propre, en module, donnera bien la valeur propre de A la plus proche de μ.*

2.8.2 Méthodes des sous-espaces

Afin de calculer simultanément plusieurs valeurs propres et leurs vecteurs propres associés, on itère sur plusieurs vecteurs au lieu d'un seul. Bien sûr, il faut empêcher tous les vecteurs de converger vers le même vecteur propre (celui associé à la plus grande valeur propre). Pour cela, la méthode des sous-espaces consiste à initialiser l'itération par un ensemble de m vecteurs orthonormés de $\mathbb{R}^N$, puis à réorthonormaliser régulièrement les vecteurs itérés. La réorthonormalisation est une opération coûteuse en temps de calcul. On évite de la faire à chaque itération. Elle utilise une factorisation QR (Q matrice orthogonale, c'est à dire telle que $Q^T Q = I$, et R matrice triangulaire supérieure de taille $m \times m$) de type Gram-Schmidt que l'on peut implémenter par la technique de Householder (voir Lascaux-Théodor). En pratique, si l'on veut obtenir p valeurs propres et vecteurs propres avec une précision acceptable, il faut itérer sur un ensemble de $m > 2p$ vecteurs.

2.8.3 Méthode QR

Pour obtenir l'ensemble des valeurs propres et des vecteurs propres d'une matrice, la principale méthode utilisée est la méthode QR. C'est une itération de type sous-espace, avec comme sous-espace, l'espace $\mathbb{R}^N$ tout entier. L'algorithme QR s'écrit :

$$\left\{ \begin{array}{l} A_0 = A \quad \text{! initialisation} \\ \text{puis pour } k = 0, 1... \\ \quad Q_k\, R_k = A_k \quad \text{! factorisation } QR \\ \quad A_{k+1} = R_k\, Q_k \quad \text{! itération par calcul du produit} \end{array} \right. \tag{2.66}$$

Il est facile de vérifier qu'à chaque itération $A_{k+1} = Q_k^T\, A_k\, Q_k$, donc que les matrices A_k obtenues sont semblables à A pour tout k.

La méthode QR est un des algorithmes les plus coûteux de l'analyse numérique matricielle. En pratique, on n'itère pas sur la matrice initiale, mais sur une matrice semblable H de forme Hessenberg (ses coefficients $H_{i,j}$ sont nuls pour

$i > j + 1$) ou tridiagonale dans le cas symétrique. Ce qui réduit sensiblement le nombre d'opérations. De nombreux travaux ont été menés pour démontrer et aussi pour accélerer la convergence de la méthode QR. On peut citer en particulier les méthodes avec "shift" dues à Wilkinson. Le cas de matrices A symétriques réelles se rencontre dans de nombreux problèmes intéressants en pratique comme en vibrations de structures par exemple. On a montré la convergence de la suite de matrices tridiagonales symétriques vers une matrice diagonale dont les coefficients sont donc les valeurs propres recherchées (Wilkinson, Parlett, Saad).

2.8.4 Méthode de Lanczos

Si l'on ne recherche qu'une partie des valeurs propres et des vecteurs propres on utilisera la méthode de Lanczos. Cette méthode permet de réduire la dimension du problème avant de poursuivre par une méthode générale, QR par exemple. La méthode de Lanczos est une méthode de calcul de valeurs propres de type sous-espace appliquée au cas de matrices symétriques réelles. Elle permet de construire itérativement, colonne par colonne, une matrice tridiagonale de taille réduite (matrice de Rayleigh) dont les éléments propres sont des approximations d'un sous-ensemble d'éléments propres de la matrice A. On obtient les valeurs propres de plus grand module par itération directe, et comme pour la méthode de la puissance, celles de plus petit module par itération inverse. Elle s'interprète en fait comme une méthode de projection. On projette le problème $AV = \lambda V$ dans un sous-espace de dimension $m << N$, selon $(AV - \lambda V, q_i) = 0$ pour $i = 1...m$. Voici l'algorithme de Lanczos pour la recherche des m plus grandes valeurs propres et des vecteurs propres associés d'une matrice A symétrique réelle $N \times N$:

$$\left\{\begin{array}{l} \text{Soit } q_0 = 0 \text{ et } q_1 \in I\!R^N \text{ tel que } \|q_1\|_2 = 1 \\ \text{On calcule successivement pour } k = 1, 2...m-1 \\ w_k = Aq_k - \beta_{k-1}q_{k-1} \quad \text{! itération directe} \\ \alpha_k = (w_k, q_k) \quad \text{! produit scalaire} \\ v_k = w_k - \alpha_k q_k \quad \text{! orthogonalisation} \\ \beta_k = \|v_k\|_2 \quad \text{! Calcul de la norme euclidienne} \\ q_{k+1} = v_k/\beta_k \quad \text{! normalisation} \end{array}\right. \tag{2.67}$$

On obtient ainsi m vecteurs orthonormés q_i et une matrice T tridiagonale symétrique $m \times m$ (matrice de Rayleigh) constituée d'éléments diagonaux $T_{i,i} = \alpha_i$ et extradiagonaux $T_{i,i+1} = T_{i+1,i} = \beta_i$. Les valeurs propres et vecteurs propres de T se calculent par la méthode QR. On obtient ainsi les m plus grandes valeurs propres de A et les composantes des vecteurs propres de A dans la base des q_i. Cet algorithme est assez délicat. Les erreurs d'arrondis produisent, comme pour le gradient conjugué, des défauts d'orthogonalité. Il est nécessaire

de réorthogonaliser de temps en temps les q_i (voir Lascaux-Théodor et Parlett pour plus de détails).

2.9 Analyse en fréquence

On décrit ici de façon succincte les méthodes d'analyse fréquentielle des signaux. Ces méthodes peuvent être utilisées pour remplacer la discrétisation spatiale des équations aux dérivées partielles linéaires par des méthodes de type différences ou éléments finis, et donc, pour ramener la résolution du problème à celle soit d'équations algébriques, soit d'équations différentielles ordinaires.

Cependant le domaine d'application le plus courant, pour l'analyse en fréquence, est le traitement du signal et la compression des données. Il est facile de voir qu'une sinusoïde infinie à amplitude constante nécessitera une infinité de points pour sa représentation spatiale exacte, mais qu'elle peut être représentée en fréquence par la donnée unique de deux scalaires (amplitude et fréquence). On recherche donc les bases les plus adaptées à ce problème de représentation. La transformée de Fourier utilisera une base trigonométrique globalisante, tandis que les ondelettes s'attacheront à une localisation de la liaison entre les caractéristiques temporelles et spatiales du signal.

2.9.1 Transformée de Fourier

La transformée de Fourier analyse le "contenu fréquentiel" d'un signal en le décomposant sur une base trigonométrique particulière. Pour une fonction f dans $L^1(I\!R) \cap L^2(I\!R)$, la transformée de Fourier est donnée par :

$$\hat{f}(\omega) = \int_{I\!R} f(t) \exp(-2i\pi\omega t) dt.$$

Connaissant $\hat{f}(\omega)$, on peut reconstruire f par une sommation infinie de sinusoïde :

$$f(t) = \int_{I\!R} \hat{f}(\omega) \exp(2i\pi\omega t) d\omega$$

La manipulation des opérateurs de dérivation est facilitée par cette opération car les sinusoïdes en sont les fonctions propres :

$$\widehat{f^{(p)}}(\omega) = (2i\pi\omega)^p \, \hat{f}(\omega)$$

On constate que les formules ci-dessus donnent une représentation globale du signal : l'information sur tout le domaine est utilisée pour passer en fréquence et, dans l'autre sens, toute l'information fréquentielle est nécessaire pour trouver la valeur en un point d'espace. Ceci rend la compression d'informations non-bijective : on est certain de perdre de l'information en revenant dans le domaine spatial, après une opération de cut-off en fréquence (dans lequel on se contente d'un nombre limité de fréquences pour représenter le signal).

Remarque 2.9.1 *La transformation de Fourier permet d'obtenir des informations sur la régularité locale d'une fonction. En particulier : une fonction f est bornée, p fois différentiable et à dérivées bornées, si* $I = \int_{\mathbb{R}} |\hat{f}(\omega)|(1 + |\omega|^p)d\omega$ *est borné.*

2.9.2 Transformée de Fourier discrète

En pratique, on ne dispose pas d'une représentation continue du signal, mais plutôt d'un échantillonnage spatial (penser à une image numérique 1024×614). D'un signal f à n échantillons c'est à dire d'un n-uplet $(f_0, ...f_{n-1})$ de nombres complexes, on passe à un autre n-uplet $(\hat{f}_0, ...\hat{f}_{n-1})$ par la transformée de Fourier discrète suivante :

$$\hat{f}_q = \sum_{p=0}^{n-1} f_p \, \omega_n^{pq} \quad \text{pour} \quad 0 \leq q \leq n-1 \tag{2.68}$$

où $\omega_n = \exp(-2i\pi/n)$.

La transformée inverse se calcule par :

$$f_p = \frac{1}{n} \sum_{q=0}^{n-1} \hat{f}_q \, \omega_n^{-pq} \quad \text{pour} \quad 0 \leq p \leq n-1$$

D'après l'expression (2.68), on constate que $\hat{f}_q = \hat{f}_{q+n}$ car $\exp(2i\pi n) = 1$. De plus, pour $n = 1$, ceci se réduit à $\hat{f} = f$.

Le calcul de l'intégrale discrète (2.68) demande n^2 additions et multiplications complexes. En utilisant la transformée de Fourier rapide, on peut réduire ce coût.

2.9.3 Transformée de Fourier rapide (FFT)

La transformée de Fourier rapide sépare le traitement des indices pairs et impairs lors du calcul d'une transformée de Fourier discrète. Ceci permet de diminuer le nombre d'opérations. Supposons $n = 2m$ pair et séparons les éléments de f d'indices pairs de ceux d'indices impairs :

$$f[pair] = (f_0, f_2, ..., f_{n-2}), \quad f[impair] = (f_1, f_3, ..., f_{n-1}).$$

Alors, en séparant la somme en deux termes regroupant d'une part les $p = 2n$ et d'autre part les $p = 2n + 1$ et en utilisant $\omega_{2m}^{2jq} = \omega_m^{jq}$ on a :

$$\hat{f}_p = \sum_{p=0}^{n-1} f_p \omega_n^{pq} \quad \text{pour} \quad 0 \leq p \leq n-1$$

$$= \sum_{j=0}^{m-1} f_{2j}\omega_{2m}^{2jp} + \sum_{j=0}^{m-1} f_{2j+1}\omega_{2m}^{(2j+1)p}$$

$$= \sum_{j=0}^{m-1} f_{2j}\omega_{m}^{jp} + \omega_{2m}^{p} \sum_{j=0}^{m-1} f_{2j+1}\omega_{m}^{jp}$$

$$\hat{f}_p = \hat{f}[pair]_p + \omega_n^p \hat{f}[impair]_p$$

Ainsi, pour calculer la transformée de Fourier sur n points, il suffit de calculer deux transformées de Fourier sur $n/2$ points, de faire $n/2$ multiplications et $n/2$ additions. Si n est une puissance de 2, cette décomposition peut s'appliquer de façon récursive, introduisant la relation suivante entre le coût de la FFT pour deux niveaux successifs : $T(n) = 2T(n/2) + O(n)$ au lieu de $4T(n/2)$ pour (2.68). Par récurrence, on montre que le nombre d'opérations nécessaires au calcul de la transformée de Fourier par cette méthode est en $O(n \log_2(n))$. La récursivité de la démarche permet une programmation récursive, donc compacte, de la FFT.

2.9.4 Transformée en ondelettes

La transformée de Fourier est une relation globale entre les distributions dans les espaces physique et fréquentiel.

La transformée en ondelettes utilise, au lieu de la sinusoïde de la transformée de Fourier, une famille de translations et dilatations d'une même fonction, appelée mère d'ondelette. L'ondelette doit être une fonction de moyenne nulle, centrée au voisinage de 0 et d'énergie finie, à support compact. Cette compacité permettra la localisation des informations obtenues sur un signal. Les paramètres de translation et de dilatation sont les deux arguments de la transformation.

La transformée en ondelettes d'une fonction $f(t)$ est définie, comme dans le cas de Fourier, par :

$$W(f)(u,s) = < f, \psi_{u,s} > = \frac{1}{\sqrt{s}} \int_{\mathbb{R}} f(t)\psi(\frac{t-u}{s})dt$$

où $\psi_{u,s}$ est une fonction centrée en u, à valeurs réelles. Par ailleurs, si le centre de fréquence de l'ondelette est en ω, le centre de fréquence de la fonction dilatée est en ω/s.

2.10 Méthodes intégrales

Les méthodes intégrales de résolution de problèmes aux limites sont basées sur la recherche des fonctions et valeurs propres des opérateurs différentiels. Elles utilisent les représentations intégrales des fonctions harmoniques (formule

de Green). Dans le cas des problèmes aux limites, elles aboutissent, après discrétisation, à des systèmes de taille plus faible que lors de l'utilisation de méthodes de discrétisation locale de type différences, volumes ou éléments finis. Leur utilisation est intéressante si l'on connaît les solutions élémentaires du problème aux limites.

En ce qui concerne les problèmes de Cauchy, l'approche intégrale est très naturelle. Elle a permis de construire les méthodes de type Runge et Kutta.

Nous commençons par une étude des équations intégrales dans un cadre monodimensionnel. L'existence et l'unicité des solutions, ainsi que la convergence des méthodes numériques utilisent le théorème du point-fixe que nous démontrons dans le cas général des espaces de Banach. Nous avons déjà utilisé ce théorème dans le cas simple des itérations de point-fixe pour la résolution d'équations $f(x) = 0$.

2.10.1 Théorème du point-fixe

Définition 2.10.1 *Soit T une application d'un espace normé E dans lui-même, T est une contraction (ou application contractante) s'il existe une constante k, $0 \leq k < 1$ telle que pour tout x et y de E on ait :*

$$\|T(x) - T(y)\| \ \leq \ k \, \|x - y\| \tag{2.69}$$

Théorème 2.10.1 *Si T est une application* **contractante** *d'un espace de* **Banach** *E dans lui-même, T possède un point fixe unique $s \in E$, qui est par définition la solution unique de l'équation*

$$T(x) = x \tag{2.70}$$

De plus la suite $\{x_n\}$, dite des approximations successives, définie par

$$\begin{cases} x_0 \text{ donné dans } E \\ x_{n+1} = T(x_n) \end{cases} \tag{2.71}$$

converge vers s.

Démonstration :

1) Existence d'une solution

La suite $\{x_n\}$ est une suite de Cauchy. En effet posons p entier positif :

$$\|x_{n+p} - x_n\| = \|T(x_{n+p-1}) - T(x_{n-1})\| \leq k \, \|x_{n+p-1} - x_{n-1}\|$$

donc

$$\|x_{n+p} - x_n\| \leq k^n \, \|x_p - x_0\|$$

D'autre part $\|x_p - x_0\| \leq \|x_p - x_{p-1}\| + \|x_{p-1} - x_{p-2}\| + ... + \|x_1 - x_0\|$,
soit

$$\|x_p - x_0\| \leq \ (k^p + k^{p-1} + ...k + 1)\, \|x_1 - x_0\| = \frac{1 - k^{p+1}}{1 - k}\, \|x_1 - x_0\|$$

On utilise l'hypothèse de contraction $k < 1$ et on obtient :

$$\|x_p - x_0\| \leq \ \frac{1}{1 - k}\, \|x_1 - x_0\|$$

d'où le résultat

$$\|x_{n+p} - x_n\| \leq k^n\, \frac{1}{1 - k}\, \|x_1 - x_0\|$$

Donc la suite x_n est une suite de Cauchy qui converge vers une limite s dans l'espace complet E. Cette limite vérifie

$$s = T(s)$$

C'est donc bien un point fixe de T.

2) Unicité de la solution

Supposons l'existence de 2 solutions s_1 et s_2, l'hypothèse de contraction entraîne :

$$\|s_1 - s_2\| = \|T(s_1) - T(s_2)\| < \|s_1 - s_2\|$$

Donc contradiction et $s_1 = s_2$.

Remarque 2.10.1 *La condition imposant l'existence d'une constante $k < 1$ est indispensable, il ne suffit pas d'avoir :*

$$\|T(x) - T(y)\| \ < \ \|x - y\| \quad \forall\ x \neq y$$

Par contre, il suffit que, pour une certaine puissance p, T^p soit contractante pour avoir l'existence et l'unicité d'un point fixe de T. La démonstration est laissée au lecteur.

2.10.2 Application aux équations intégrales de Fredholm

On considère un réel λ, une fonction numérique réelle g continue sur un intervalle fermé $[a, b]$, et une fonction réelle K de 2 variables réelles continue sur le pavé fermé $[a, b] \times [a, b]$. Le problème intégral de Fredholm s'écrit :

$$\begin{cases} \text{Trouver la fonction } \ f \ \text{ définie sur } \ [a, b] \ \text{ telle que :} \\ f(x) = \lambda \displaystyle\int_a^b \ K(x, y)\, f(y)\, dy \ + \ g(x) \quad \forall x \in [a, b] \end{cases} \tag{2.72}$$

On se place dans l'espace $C[a, b]$ des fonctions continues muni de la norme du max notée norme ∞

$$\|f\|_\infty = \sup_{x \in [a,b]} |f(x)|$$

On pose (K est continue sur $[a,b]\times[a,b]$) :

$$M = \sup_{(x,y)\in[a,b]\times[a,b]} |K(x,y)|$$

et on obtient aisément par application du théorème du point fixe le résultat suivant :

Théorème 2.10.2 *L'équation de Fredholm admet une solution unique dans $C[a,b]$ à la condition suffisante que :*

$$|\lambda|\ M\ (b-a)\ < 1 \tag{2.73}$$

Démonstration :

1) $C[a,b]$ muni de la norme ∞ est un Banach.

2) L'application T définie par

$$\begin{cases} f \longrightarrow Tf \text{ où :} \\ Tf(x) = \lambda \displaystyle\int_a^b K(x,y)\, f(y)\, dy \ + \ g(x) \quad \forall x \in [a,b] \end{cases}$$

est une application de $C[a,b]$ dans lui-même. En effet la fonction Tf est continue sur $[a,b]$.

$$Tf(x_0+h) - Tf(x_0) = \lambda \int_a^b (K(x_0+h,y) - K(x_0,y))\, f(y)\, dy \ + \ g(x_0+h) - g(x_0)$$

d'où

$$|Tf(x_0+h) - Tf(x_0)| \leq |\lambda| \int_a^b |K(x_0+h,y) - K(x_0,y)|\, |f(y)|\, dy \ + \ |g(x_0+h) - g(x_0)|$$

et le résultat en utilisant la continuité de K et celle de g.

3) L'application T est une contraction, en effet on a :

$$Tf_1(x) - Tf_2(x) = \lambda \int_a^b (K(x,y))\, (f_1(y) - f_2(y))\, dy$$

donc comme K est bornée par M dans $[a,b]\times[a,b]$

$$|Tf_1(x) - Tf_2(x)| \leq |\lambda|\, M\, (b-a)\ \|f_1 - f_2\| \quad \forall x \in [a,b]$$

et le résultat

$$\|Tf_1 - Tf_2\| \leq |\lambda|\, M\, (b-a)\ \|f_1 - f_2\|$$

avec

$$|\lambda|\ M\ (b-a)\ < 1$$

Remarque 2.10.2 *On pourrait remplacer la condition (2.73) ci-dessus par la suivante :*

$$\sup_{x\in[a,b]} |\lambda| \int_a^b |K(x,y)| dy \leq k < 1$$

2.10.3 Équations de Volterra

Avec les mêmes hypothèses que précédemment, on considère le problème :

$$\left\{ \begin{array}{l} \text{Trouver la fonction } \ f \ \text{ définie sur } \ [a,b] \ \text{ telle que :} \\ f(x) = \lambda \displaystyle\int_a^x \ K(x,y)\, f(y)\, dy \ + \ g(x) \quad \forall x \in [a,b] \end{array} \right. \tag{2.74}$$

Cette fois la borne supérieure de l'intégrale est égale à la variable x. Ceci permet d'obtenir avec T définie par :

$$\left\{ \begin{array}{l} f \longrightarrow Tf \text{ où :} \\ Tf(x) = \lambda \displaystyle\int_a^x \ K(x,y)\, f(y)\, dy \ + \ g(x) \quad \forall x \in [a,b] \end{array} \right.$$

le résultat d'existence et d'unicité sans l'hypothèse restrictive

$$|\lambda| \ M \ (b-a) \ < 1$$

On reprend la démonstration précédente en observant cette fois que T^p sera contractante à partir d'une certaine puissance de p grâce à l'inégalité (dont la démonstration est laissée au lecteur) :

$$|T^p f_1(x) - T^p f_2(x)| \leq |\lambda^p|\, M^p\, \frac{(b-a)^p}{p!}\, \|f_1 - f_2\| \quad \forall x \in [a,b]$$

2.10.4 Application aux équations différentielles

On considère le problème différentiel :

$$\left\{ \begin{array}{l} y'(x) = f(x, y(x)) \quad \forall x \in [a,b] \\ y(a) = y_0 \end{array} \right.$$

Ce problème est équivalent, si y est une solution continûment différentiable (on dit de classe C^1), au problème intégral :

$$\left\{ \begin{array}{l} \text{Trouver la fonction } \ y \ \text{ définie sur } \ [a,b] \ \text{ telle que :} \\ y(x) = y_0 \ + \displaystyle\int_a^x \ f(t, y(t))\, dt \quad \forall x \in [a,b] \end{array} \right.$$

On considère l'application T définie par :

$$Ty(x) = y_0 \ + \displaystyle\int_a^x \ f(t, y(t))\, dt \quad \forall x \in [a,b]$$

On suppose que la fonction f vérifie les hypothèses suivantes :

1) f est continue sur $[a,b] \times I\!R$

2) f est lipschitzienne par rapport à sa deuxième variable, c'est à dire qu'il existe une constante L telle que :

$$\forall x \in [a,b],\ \forall y,z \in I\!R, \quad |f(x,y) - f(x,z)| \leq L\,|y-z|$$

On montre alors comme précédemment que T est une application de $C[a,b]$ dans lui même qui admet un point fixe unique, solution du problème, car T^p est une contraction stricte à partir d'une certaine puissance p.

2.10.5 Méthodes de résolution numérique de l'équation de Fredholm

On reprend le problème de Fredholm :

$$\left\{ \begin{array}{l} \text{Trouver la fonction } \ f \ \text{ définie sur } \ [a,b] \ \text{ telle que :} \\ f(x) = \lambda \displaystyle\int_a^b K(x,y)\, f(y)\, dy \ + \ g(x) \quad \forall x \in [a,b] \end{array} \right. \tag{2.75}$$

On se propose de le résoudre par une méthode numérique.

Écriture du problème approché

La première étape de toute approche numérique consiste à discrétiser la géométrie du problème. Ici, il suffit très simplement de choisir un certain nombre de points x_i discrètement répartis sur $[a,b]$ de telle sorte que l'intervalle $[a,b]$ soit subdivisé en sous-intervalles $[x_{i-1}, x_i]$. Nous adoptons la répartition suivante :

$$a = x_1 < x_2 < .... < x_i < ... < x_N = b$$

La deuxième étape est la discrétisation de l'opérateur fonctionnel, ici l'intégrale. Nous choisissons une méthode de quadrature numérique qui s'écrit sous la forme générale (voir 2.4) :

$$\int_a^b F(x)\, dx \ \approx \ \sum_{i=1}^{N} A_i\, F(x_i) \tag{2.76}$$

Remarque 2.10.3 *Nous avons choisi les mêmes points x_i comme points d'intégration et comme points de discrétisation. Ceci simplifie l'écriture du problème approché.*

Par exemple on peut prendre la méthode des trapèzes, qui dans le cas de N points régulièrement espacés d'un pas $h = \dfrac{b-a}{(N-1)}$, s'écrit :

$$\int_a^b F(x)\,dx \;\approx\; h\,\Big(\;\frac{F(a)+F(b)}{2} \;+\; \sum_{i=2}^{N-1} F(x_i)\;\Big)$$

ou la méthode de Simpson :

$$\int_a^b F(x)dx \;\approx\; \frac{h}{3}\Big(\;F(a)+F(b)+\;2\sum_{i=1}^{p-1} F(a+2ih)+\;4\sum_{i=1}^{p} F(a+(2i-1)h)\;\Big)$$

avec N **impair**, $p = \dfrac{N-1}{2}$ et $h = \dfrac{b-a}{N-1}$.

Ainsi en notant f_i pour $i = 1...N$, les valeurs approchées des $f(x_i)$ et en remplaçant l'intégration exacte dans l'équation de Fredholm par une formule de quadrature approchée, nous obtenons le système linéaire suivant dont les inconnues sont les valeurs f_i cherchées :

$$\left\{\begin{array}{l} \text{Trouver les valeurs } \; f_i \; \text{ telles que :} \\ \\ f_i = \lambda \displaystyle\sum_{j=1}^{N} A_j \;\; K(x_i,x_j)\, f_j \;+\; g(x_i) \quad \forall i = 1...N \end{array}\right. \tag{2.77}$$

Ce système de N équations à N inconnues peut se mettre sous forme matricielle

$$\Big[\; M \;\Big]\, F \;=\; G$$

avec

$$\Big[\; M \;\Big] \;=\; \Big[\; I \;\; -\lambda\, K \;\Big]$$

où K est la matrice de coefficients :

$$K_{i,j} = A_j\, K(x_i,x_j)$$

F le vecteur des inconnues f_i et G le vecteur second-membre des $g(x_i)$.

Résolution du problème approché

Méthode directe

Dans ce cas, on résout le système matriciel par une méthode directe de type Gauss ou Choleski si la matrice est symétrique définie positive. On obtient alors, aux erreurs d'arrondis près, la solution exacte du système approché. Notons que, si

l'on choisit une formule de quadrature à poids positifs, nous sommes assurés que la matrice M sera inversible car à diagonale strictement dominante.

En effet, nous avons par hypothèse, pour le problème de Fredholm :

$$k = |\lambda| \sup_{(x,y)\in[a,b]\times[a,b]} |K(x,y)| \ (b-a) \ < 1.$$

D'autre part, si la formule de quadrature est à poids positifs, on a :

$$\sum_{j=1}^{N} |A_j| = \sum_{j=1}^{N} A_j \ = \ b-a.$$

Donc pour chaque ligne i de la matrice M

$$\sum_{j\neq i} |M_{ij}| = |\lambda| \sum_{j\neq i} |A_j| \, |K(x_i,x_j)| < 1 - |\lambda| \, |A_i| \, |K(x_i,x_i)| \leq |M_{ii}|$$

D'où le résultat.

Méthodes itératives

On peut utiliser les itérations de point-fixe dans le cas du problème approché.
En effet, avec la même condition que le problème continu, le problème approché vérifie les hypothèses d'un théorème du point-fixe.
On se place dans l'espace $\mathbb{R}^N$ muni de la norme du max

$$\|F\|_\infty = \sup_{i=1,N} |f_i|$$

et on obtient aisément par application du théorème du point-fixe le résultat suivant :

Théorème 2.10.3 *Le problème approché :*

$$\left\{ \begin{array}{l} \textit{Trouver les valeurs } \ f_i \ \textit{ telles que :} \\ \\ f_i = \lambda \sum_{j=1}^{N} A_j \ K(x_i,x_j) \, f_j \ + \ g(x_i) \quad \forall i = 1,..N \end{array} \right. \tag{2.78}$$

discrétisation de l'équation de Fredholm, admet une solution unique dans $\mathbb{R}^N$ à condition que :

$$|\lambda| \ M \ (b-a) \ < 1$$

De plus, l'itération :

$$\left\{ \begin{array}{l} f_i^0 \ \textit{ donné pour tout } i = 1,N \\ \\ f_i^{n+1} = \lambda \sum_{j=1}^{N} A_j \ K(x_i,x_j) \, f_j^n \ + \ g(x_i) \quad \forall i = 1,..N \end{array} \right. \tag{2.79}$$

converge, dans ce cas, vers le vecteur F solution du problème approché.

Remarque 2.10.4 *La méthode itérative ci-dessus peut s'interpréter comme une méthode de résolution itérative du système linéaire obtenu par discrétisation. Elle est assez proche de la méthode de Jacobi. On pourrait également proposer des solutions itératives en utilisant les méthodes de relaxation ou de gradient conjugué.*

Etude de l'erreur

Théorème 2.10.4 *Si la formule de quadrature utilisée pour l'écriture du problème approché est d'ordre k (voir définition 2.4.1), l'erreur d'approximation*

$$E = \max_{i=1,N} |f(x_i) - f_i| \leq C\, h^k. \tag{2.80}$$

Démonstration

On a

$$f(x_i) - f_i = \lambda \int_a^b K(x_i, y)\, f(y)\, dy \;\; - \;\; \lambda \sum_{j=1}^{N} A_j \;\; K(x_i, x_j)\, f_j$$

D'où en ajoutant et retranchant le terme auxiliaire $\lambda \sum_{j=1}^{N} A_j \;\; K(x_i, x_j)\, f(x_j)$ l'inégalité :

$$\begin{aligned} | \; f(x_i) - f_i \; | \leq \big| \; \lambda \textstyle\int_a^b \; K(x_i, y)\, f(y)\, dy \;\; - \;\; \lambda \sum_{j=1}^{N} A_j \;\; K(x_i, x_j)\, f(x_j) \; \big| \\ + \big| \; \lambda \textstyle\sum_{j=1}^{N} A_j \;\; K(x_i, x_j)\, (f(x_j) - fj) \; \big| \end{aligned}$$

Or par hypothèse :

$$|\, \lambda \sum_{j=1}^{N} A_j \;\; K(x_i, x_j)\, (f(x_j) - fj)| \leq |\, \lambda \,|\, M\, (b-a)\, E$$

En majorant l'erreur d'intégration :

$$|\lambda \int_a^b \; K(x_i, y)\, f(y)\, dy \;\; - \;\; \lambda \sum_{j=1}^{N} A_j \;\; K(x_i, x_j)\, f(x_j)| \leq c\, h^k$$

on obtient le résultat :

$$E \leq \frac{1}{1 - |\, \lambda \,|\, M\, (b-a)}\, c\, h^k$$

2.10.6 Application aux problèmes aux limites

Nous présentons l'application des méthodes intégrales à la résolution des problèmes aux limites à travers la solution d'un problème très simple. Il s'agit

d'un problème de Dirichlet intérieur (i.e. on cherche la solution dans un ouvert avec conditions aux limites de Dirichlet au bord) en dimension un.

$$-u''(x) = f(x) \quad \text{sur} \quad \Omega, \quad u(\partial\Omega) = u_{\partial\Omega}, \tag{2.81}$$

où $\Omega =]a, b[\subset \mathbb{R}$ et $\partial\Omega = \{a, b\}$.

La démarche, dans ces méthodes, consiste à séparer le traitement des conditions aux limites de celui des termes sources en utilisant la linéarité du problème. Au contraire des méthodes comme les différences, volumes ou éléments finis, où ces deux points sont traités en même temps, comme nous le verrons au chapitre 4.

— Résolution du problème sur $\mathbb{R}$:

$$-w''(x) = f(x) \quad \text{sur} \quad \mathbb{R}. \tag{2.82}$$

— Résolution du problème de Dirichlet intérieur sans terme source :

$$-v''(x) = 0 \quad \text{sur} \quad \Omega, \quad v(\partial\Omega) = v_{\partial\Omega} = u_{\partial\Omega} - w(\partial\Omega). \tag{2.83}$$

— Assemblage des deux contributions :

$$u = w + v.$$

Remarque 2.10.5 *Dans le cas des équations non-linéaires, nous procédons de la même façon pour l'équation linéarisée.*

Étudions la résolution des deux sous-problèmes ci-dessus.

2.10.7 Résolution du problème en milieu infini

Pour résoudre (2.82), deux approches sont possibles :

— 1. Soit résoudre par une méthode de différences finies rapide le problème étendu suivant :

$$-w''(x) = \tilde{f}(x) \quad \text{sur} \quad D, \quad \text{avec} \quad w(\partial D) = 0. \tag{2.84}$$

où $\Omega \subset D$, $\tilde{f} = f$ sur Ω et $\tilde{f} = 0$ sur $D - \Omega$. Cette approche est utilisée en particulier si f n'est pas explicitement connue.

— 2. Soit, si f est connue explicitement, calculer la solution par convolution du second membre avec le noyau de Green ϕ de l'opérateur dérivée seconde :

$$w(x) = \int_{\mathbb{R}} \phi(x-y) f(y) dy = \int_{\mathbb{R}} \phi(x-y)(-w''(y)) dy$$

$$= \int_{I\!R} -\phi''(x-y)w(y)dy =< \delta(x), w >= w(x).$$

Pour ne pas mailler $I\!R$ en entier, on utilise la transformée de Fourier rapide (FFT). Rappelons que le noyau de Green est une solution élémentaire de :

$$-\phi''(x) = \delta_0 \quad \text{sur} \quad I\!R. \tag{2.85}$$

ϕ est donc harmonique si $x \neq 0$. δ_0 représente une masse de Dirac placée en $x = 0$ définie au sens des distributions par :

$$\forall \psi \in C^0_{comp}(I\!R), < \delta_0, \psi >= \psi(0).$$

Il est facile de construire un noyau de Green dans ce cas en dimension un. On peut prendre, par exemple :

$$\phi(x) = 2x \quad \text{si} \quad x < 0, \quad \phi(x) = x \quad \text{sinon.}$$

2.10.8 Résolution du problème de Dirichlet intérieur sans terme source

Ayant remarqué que ϕ est harmonique pour $x \neq 0$, nous allons chercher une solution de (2.83) de la forme :

$$v(x) = \lambda_1\phi(x-a) + \lambda_2\phi(x-b), \quad \text{sur} \quad]a,b[,$$

$$= \lambda_1(x-a) + 2\lambda_2(x-b),$$

λ_1 et λ_2 sont évalués pour que la solution vérifie les conditions aux limites :

$$\lambda_1 = \frac{\beta}{b-a}, \quad \lambda_2 = \frac{2\alpha}{a-b}.$$

Ainsi, on constate que résoudre cette équation se ramène simplement à résoudre un système de deux équations à deux inconnues, obtenues sur la frontière du domaine.

En dimensions supérieures, la même technique permet de ramener le problème à une équation intégrale de type Fredholm sur la frontière. On aboutit donc à des systèmes linéaires de taille plus petite, la taille dépendant du nombre de points sur la frontière qui est sensiblement inférieur au nombre de noeuds que comporterait un maillage global. Mais les matrices des systèmes obtenus sont pleines, au contraire des matrices résultant des méthodes de discrétisation locales (différences finies, éléments finis ou volumes finis) dont la plupart des coefficients sont nuls. Ceci est une des difficultés essentielles des méthodes intégrales : elles aboutissent à de grands systèmes linéaires pleins à inverser.

De plus, les noyaux de Green en dimensions 2 ou 3 ne sont pas définis aux points de discrétisations frontières. Par exemple, en dimension 3, le noyau de Green pour le Laplacien est de la forme $\phi(x) = 1/|x|$. Lors de la résolution du problème (2.83) équivalent dans $\Omega \subset I\!R^3$:

$$-\Delta v = 0 \quad \text{dans} \quad \Omega, \quad v(\partial\Omega) = v_{\partial\Omega}, \tag{2.86}$$

On cherchera une solution de la forme :

$$v = \sum_{i=1}^{N} \lambda_i \phi(x - x_i) = \sum_{i=1}^{N} \frac{\lambda_i}{|x - x_i|}, \tag{2.87}$$

qui doit satisfaire $v(x_i) = v_{\partial\Omega}(x_i),$ pour $i = 1, .., N$ si la frontière a été discrétisée en N points. Ceci n'est pas possible au sens fort car v n'est pas définie en x_i. On peut cependant établir une formulation faible en écrivant :

$$\int_{\partial\Omega} v(x)\phi(x_j)d\gamma = \int_{\partial\Omega} v_{\partial\Omega}(x)\phi(x_j)d\gamma, \quad \forall j = 1, .., N.$$

Mais même dans ce cas, les coefficients sont indéfinis en $x = x_i$ et il faut recourir à une évaluation des intégrales utilisant des points d'intégration différents des points de discrétisation. De plus, la complexité de l'évaluation de ce système est en N^2. Et la matrice, une fois assemblée, sera difficile à inverser à cause de sa structure pleine.

Une autre difficulté dans l'approche intégrale provient de la prise en compte de conditions aux limites complexes. De même l'extension à des opérateurs quelconques n'est pas aisée et pose, à chaque fois, le problème du calcul du noyau de Green. Dans l'ensemble, les codes de simulation industriels sont rarement basés sur les approches intégrales. On utilise plus volontiers des approches différences, volumes ou éléments finis, que l'on présente au chapitre 4.

Les méthodes intégrales sont cependant inévitables pour les applications en propagation des ondes, en particulier en haute fréquence et spécialement si le domaine de calcul est de grande taille. En effet, sachant que pour capter de façon satisfaisante une onde, il faut prévoir environ une dizaine de points par longueur d'onde, considérons un corps d'une dizaine de mètres éclairé par un signal radar de 100 gigaHertz (correspondant à des longueurs d'onde de l'ordre du centimètre). Un calcul de signature radar (concernant le signal réfléchi) demandera en dimension trois, dans une boîte de 100 mètres de côté, un nombre de points de l'ordre de $(10^5)^3 = 10^{15}$. Ceci rend les calculs différences, volumes ou éléments finis 3D ingérables, ne serait-ce que pour des raisons de stockage. Une approche intégrale s'impose.

L'efficacité des méthodes intégrales peut être améliorée par un couplage avec une méthode multi-polaire pour réduire le nombre de variables du système.

2.11 Approximation multi-pôles

Cette approche consiste à regrouper les opérations redondantes, à ne les effectuer qu'une seule fois et à stocker le résultat pour une réutilisation. On peut voir ces approches comme un outil de prise en compte d'un grand nombre de particules, permettant de reproduire l'effet global de ces particules sur une variable, à travers un nombre réduit de particules nommées pôles.

Considérons une expression du type (2.87) ci-dessus. L'évaluation de cette expression en N points $x_k, k = 1, .., N$ a une complexité en N^2. Une décomposition de cette somme en sous-sommes de P paquets de taille M permet de réécrire :

$$v(x_k) = \sum_{i=1}^{N} \lambda_i \phi(x_k - x_i) = \sum_{p=1}^{P} \left(\sum_{m=1}^{M} \lambda_{m,p} \phi(x_k - x_{m,p}) \right), \quad k = 1, .., N,$$

avec $P << N$ et $PM = N$. Le but maintenant est de remplacer chaque paquet de points par un seul point (pôle) $z_p, p = 1, .., P$, tout en contrôlant l'erreur que l'on commet dans cette opération. À titre d'exemple, pour l'application en propagation d'ondes citée plus haut où $N \sim 10^5$, on cherchera à avoir $P \sim 100$.

Ceci est possible, par exemple, si l'on peut écrire une approximation de la forme :

$$\phi(x_k - x_{m,p}) = \sum_{j=0}^{J} \alpha_j \frac{(x_{m,p} - z)^j}{(x_k - z)^{j+1}} + (\frac{x_{m,p} - z}{x_k - z})^J o(1),$$

pour un pôle z. On veut en particulier que $J << M$.

Considérons par exemple la fonction $1/(x - y)$, on a par division euclidienne la décomposition suivante :

$$\frac{1}{x-y} = \frac{1}{x-z+z-y} = \frac{1}{(x-z)(1+\frac{z-y}{x-z})} = \sum_{j=0}^{J} \frac{(y-z)^j}{(x-z)^{j+1}} + (\frac{z-y}{x-z})^J o(1).$$

Ainsi, $v(x_k), (k = 1, .., N)$ peut s'écrire approximativement :

$$v(x_k) = \sum_{p=1}^{P} \left(\sum_{m=1}^{M} \lambda_{m,p} \left(\sum_{j=0}^{J} \alpha_j \frac{(x_{m,p} - z_p)^j}{(x_k - z_p)^{j+1}} \right) \right).$$

En permutant les sommes on a :

$$v(x_k) = \sum_{p=1}^{P} \sum_{j=0}^{J} \left(\sum_{m=1}^{M} \lambda_{m,p} (x_{m,p} - z_p)^j \right) \frac{\alpha_j}{(x_k - z_p)^{j+1}}.$$

On constate que la somme médiane ne dépend pas de x_k. On peut l'évaluer une seule fois pour $k = 1$ et stocker le résultat pour réutilisation pour $k = 2, .., N$.

L'évaluation a une complexité en PM et le stockage nécessite un tableau de taille $PJ << NJ$:

$$A_{pj} = \sum_{m=1}^{M} \lambda_{m,p}(x_{m,p} - z_p)^j, \quad p = 1, .., P, \quad J = 0, .., J.$$

Une fois cette évaluation effectuée, la complexité du reste de l'évaluation des $v(x_k)$ sera en $NPJ << N^2$:

$$v(x_k) = \sum_{p=1}^{P} \sum_{j=0}^{J} A_{pj} \frac{\alpha_j}{(x_k - z_p)^{j+1}}, \quad k = 2, .., N.$$

Chapitre 3

Équations aux dérivées partielles

3.1 Introduction

Les principaux modèles mathématiques, utilisés dans les applications les plus diverses, s'écrivent sous forme d'équations aux dérivées partielles. Une équation aux dérivées partielles (EDP) est une relation entre une fonction de plusieurs variables et ses dérivées. L'ordre d'une EDP est, par définition, l'ordre maximal des dérivées partielles présentes dans son expression. Notons cependant que les modèles physiques peuvent conduire, dans de nombreux cas dont nous verrons plus loin des exemples, à des solutions qui ne sont pas dérivables au sens classique. On devra alors considérer les dérivées partielles au sens des distributions ou au sens faible. C'est en particulier le cas dans les formulations variationnelles, base de la méthode des éléments finis.

À l'EDP sont associées des informations supplémentaires sur la frontière du domaine en temps et en espace où elle est définie, informations appelées conditions initiales ou finales (pour la dimension temps) et conditions aux limites (pour les dimensions d'espace). Remarquons une différence importante entre une EDP et une équation différentielle ordinaire. Alors que, pour une équation différentielle, la solution générale contient des constantes, la solution générale d'une EDP contient, elle, des fonctions quelconques. Considérons l'exemple suivant :

$$y\frac{\partial u}{\partial y} = u(x, y)$$

est une EDP du premier ordre dont la solution générale est $u(x, y) = yf(x)$. Les fonctions $u(x, y) = ycos^2(x)$ et $u(x, y) = y(sin(x) + x^2)$ sont, par exemple, deux solutions particulières de cette EDP. La non-unicité de la fonction solution rend la résolution d'une EDP plus difficile que celle d'une équation différentielle. Une solution particulière est fixée par les conditions initiales et les conditions aux limites associées à l'EDP.

Une équation aux dérivées partielles linéaire est une équation dont l'expression est une combinaison linéaire de la fonction et de ses dérivées partielles. Nous appellerons, en particulier, EDP linéaire du second ordre à coefficients constants, en dimension N, toute équation de la forme :

$$\sum_{i,j=1}^{N} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{N} b_i \frac{\partial u}{\partial x_i} + cu = f$$

Les variables sont notées x_i, la fonction inconnue est u, les a_{ij}, les b_i et c sont des réels et f est une fonction donnée.

3.2 Quelques équations modèles

Considérons quelques exemples d'équations aux dérivées partielles linéaires et non-linéaires, modèles de problèmes physiques classiques.

1. $\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0$ est l'équation de transport linéaire du premier ordre dont la solution générale, pour une condition initiale au temps 0 donnée égale à u_0, s'écrit : $u(x,t) = u_0(x - ct)$.
2. Si $c = u$ on obtient $\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0$.
   Cette équation s'appelle équation de Burgers et est un exemple simple d'équation non-linéaire du premier ordre. Elle est utilisée comme modèle théorique en mécanique des fluides. On trouve également des versions de Burgers avec dissipation $\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} - \epsilon\frac{\partial^2 u}{\partial x^2} = 0$.
3. $\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0$ est l'équation de Korteweg-de Vries (KDV) qui modélise la propagation d'une onde solitaire ou soliton. C'est aussi une équation non-linéaire, mais cette fois d'ordre trois.
4. $\frac{\partial u}{\partial t} = \sigma \frac{\partial^2 u}{\partial x^2}$ est l'équation de la chaleur en dimension un (linéaire du second ordre). En dimensions supérieures, on obtient l'écriture générale
   $$\frac{\partial u}{\partial t} = div(\sigma\, grad(u)) + f$$
5. $\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0$ est l'équation des ondes en dimension un (linéaire du second ordre). Posons $\eta = x + ct$ et $\zeta = x - ct$. Il est facile de vérifier que
   $$\frac{\partial^2 u}{\partial \eta \partial \zeta} = 0$$

et donc que

$$u(x,t) = f(x+ct) + g(x-ct)$$

pour deux fonctions f et g déterminées par les conditions initiales et aux limites. On trouve ainsi la forme générale des solutions de l'équation des ondes en dimension un.

En dimensions supérieures, l'équation des ondes s'écrit

$$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u + f$$

6. l'équation de Poisson

$$-\Delta u = f$$

où Δ est l'opérateur laplacien : $\Delta = div(grad) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$ (en dimension deux)
est un modèle général de nombreux problèmes d'équilibre en physique.

7. l'équation biharmonique $-\Delta^2 u = f$ est, en particulier, un modèle bidimensionnel d'équilibre des plaques en flexion. C'est une EDP du quatrième ordre.

8. l'équation de Monge-Ampère

$$\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 u}{\partial y^2} - (\frac{\partial^2 u}{\partial x \partial y})^2 = f(x,y),$$

représente la recherche de la surface $z = u(x,y)$ ayant une courbure moyenne en chaque point donnée où la courbure moyenne est définie comme le déterminant du hessien de u. C'est une équation non-linéaire du second ordre.

3.3 Classification des EDP linéaires du second ordre

Les EDP linéaires du second ordre jouent un rôle fondamental dans les modèles physiques. Elles sont classées en trois types : elliptiques, paraboliques, hyperboliques (selon la même classification que les coniques). Nous présentons ici une façon simple de faire cette identification. Soit l'EDP du second ordre linéaire à coefficients constants, en dimension deux :

$$a\frac{\partial^2 u}{\partial x^2} + 2b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} + d\frac{\partial u}{\partial x} + e\frac{\partial u}{\partial y} + fu = 0.$$

L'équation caractéristique de cette EDP s'écrit :

$$ax^2 + 2bxy + cy^2 + dx + ey + f = 0.$$

C'est l'équation de la conique associée. La classification d'une EDP linéaire du second ordre à coefficients constants est donnée par le type de la conique associée :

- Elliptique si $ac - b^2 > 0$,
- Parabolique si $ac - b^2 = 0$,
- Hyperbolique si $ac - b^2 < 0$,
- si $a, b, c = 0$, l'EDP linéaire du premier ordre sera aussi appelée hyperbolique.

En particulier, l'équation des ondes

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0$$

est une équation hyperbolique. On peut observer que l'opérateur du second ordre

$$\frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2}$$

se factorise sous forme de produit de deux opérateurs du premier ordre

$$\frac{\partial^2}{\partial t^2} - c^2 \frac{\partial^2}{\partial x^2} = (\frac{\partial}{\partial t} - c\frac{\partial}{\partial x})(\frac{\partial}{\partial t} + c\frac{\partial}{\partial x})$$

ce qui a conduit à la solution générale trouvée ci-dessus et ce qui montre l'existence de deux familles de droites *caractéristiques*.

$$x - ct = cste \quad \text{et} \quad x + ct = cste$$

Cette classification, construite sur des critères mathématiques, a des conséquences sur les propriétés physiques des solutions. Les équations de type elliptique modélisent des problèmes d'équilibre dont la solution minimise une fonctionnelle d'énergie. Les équations paraboliques servent de modèles mathématiques des problèmes d'évolution dissipatifs, dans lesquels l'énergie décroit au cours du temps. Les équations hyperboliques représentent des phénomènes physiques conservatifs.

3.3.1 Cas de coefficients variables en espace

Il se peut que les coefficients soient fonctions de l'espace et du temps. Dans ce cas, le type de l'équation change selon le point où l'on se trouve. Ceci rend plus difficile la résolution de l'équation.

Exemple : Considérer l'EDP suivante :

$$y\frac{\partial^2 u}{\partial x^2} + 2x\frac{\partial^2 u}{\partial x \partial y} + y\frac{\partial^2 u}{\partial y^2} = 0,$$

Remarque 3.3.1 *On trouve d'autres façons de définir le type d'une EDP dans la littérature. Par exemple, un opérateur d'ordre 2 agissant sur* $T : \mathbb{R}^n \to \mathbb{R}$ *de la forme*

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\partial}{\partial x_j}(a_{ij}(x)\frac{\partial T}{\partial x_i}) = f, \quad a_{ij}(x) \in \mathbb{R},$$

est elliptique si,

$$\sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}(x)\zeta_i\zeta_j \geq c\sum_{i=1}^{n}\zeta_i^2, \quad \forall \zeta \in \mathbb{R}^n$$

avec la constante $c > 0$ *indépendante du point* x*. De la même façon, des caractérisations existent pour les EDP paraboliques et hyperboliques de dimensions supérieures.*

3.3.2 Cas non-linéaire

Dans ce cas, où l'on dispose de peu de résultats théoriques, on peut étudier l'EDP linéarisée correspondante. Cette analyse sera, au mieux, uniquement valable localement (i.e. au voisinage d'une solution particulière de l'équation).
Exemple : Considérons l'EDP non-linéaire suivante :

$$\frac{\partial^2 \phi}{\partial t^2} - \frac{\partial \phi}{\partial x}\frac{\partial^2 \phi}{\partial x^2} = 1.$$

L'équation linéarisée régit la variation d'un incrément au voisinage d'une solution ϕ de l'équation. Soit u cet incrément, en portant dans l'équation, on obtient :

$$\frac{\partial^2 \phi}{\partial t^2} - \frac{\partial \phi}{\partial x}\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 u}{\partial t^2} - \frac{\partial \phi}{\partial x}\frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial x}\frac{\partial^2 \phi}{\partial x^2} - \frac{\partial u}{\partial x}\frac{\partial^2 u}{\partial x^2} = 1,$$

Or, ϕ est solution et vérifie l'EDP, alors en négligeant les termes d'ordre supérieur à un en u, on obtient :

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial \phi}{\partial x}\frac{\partial^2 u}{\partial^2 x} - \frac{\partial u}{\partial x}\frac{\partial^2 \phi}{\partial x^2} = 0.$$

D'après le critère ci-dessus, cette équation en u est localement elliptique si $\dfrac{\partial \phi}{\partial x} < 0$ (on retrouve une équation d'advection-diffusion stationnaire dans le plan (x, t))

et localement hyperbolique sinon (on retrouve une équation des ondes avec un terme d'advection).

Si on considère l'équation de Monge-Ampère pour la variable ϕ

$$\frac{\partial^2\phi}{\partial x^2}\frac{\partial^2\phi}{\partial y^2} - (\frac{\partial^2\phi}{\partial x\partial y})^2 = f(x,y),$$

l'équation linéarisée en u est :

$$\frac{\partial^2 u}{\partial x^2}\frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial x^2}\frac{\partial^2 u}{\partial y^2} - 2\frac{\partial^2\phi}{\partial x\partial y}\frac{\partial^2 u}{\partial x\partial y} = 0.$$

D'après l'équation caractéristique, le type de l'équation dépend du déterminant, donc du signe du second membre f de l'équation ($ac - b^2 = f(x,y)$).

3.4 Équation elliptique linéaire

L'exemple le plus simple d'EDP elliptique linéaire est l'équation de Poisson.

$$-\Delta u = f$$

Cette équation modélise de nombreux problèmes d'équilibre stationnaire. Elle se généralise sous la forme

$$-\, div(kgrad)u = f \tag{3.1}$$

pour des coefficients variables. Voici quelques problèmes physiques classiques dont elle sert de modèle.

3.4.1 Conduction thermique

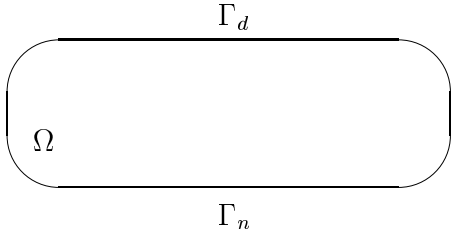


FIGURE 3.1 – Température sur une plaque

On considère une plaque plane Ω de frontière Γ dont une partie Γ_d est à température connue T_d et une partie Γ_n est isolée thermiquement. La température

T à l'équilibre vérifie l'équation

$$\begin{cases} -K\,\Delta\,T\,(x,y) = f\,(x,y) & \forall\ x,y \in \Omega \\ T_{|\,\Gamma_d}(x,y) = T_d\,(x,y) \\ \dfrac{\partial T}{\partial n}_{|\,\Gamma_n}(x,y) = 0 \end{cases} \tag{3.2}$$

K est le coefficient de conductivité thermique (la propagation de la chaleur augmente avec K) et f représente une source volumique de chaleur (sous forme d'une densité de puissance). Si K est variable, l'équation se généralise sous la forme

$$-div(Kgrad)T = f$$

3.4.2 Membrane élastique

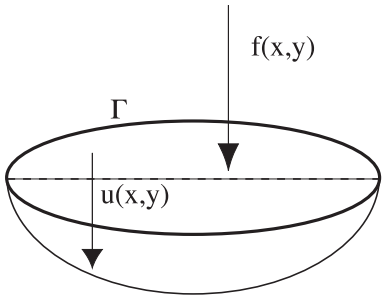


FIGURE 3.2 – Membrane élastique

On considère une membrane élastique plane Ω fixée sur son pourtour Γ. On suppose la membrane soumise en tout point (x,y) à une densité de forces f s'exerçant perpendiculairement au plan de la membrane. Sous l'action de f chaque point de la membrane subit un petit déplacement. Le déplacement transversal, perpendiculaire au plan de Ω est l'inconnue u de ce problème et vérifie l'équation :

$$\begin{cases} -\Delta\ u\,(x,y) = f\,(x,y) & \forall\ x,y \in \Omega \\ u_{\,|\Gamma}(x,y) = 0 \end{cases} \tag{3.3}$$

3.4.3 Mécanique des fluides parfaits

Soit Ω le domaine interne d'une tuyère, on considère l'écoulement bidimensionnel d'un fluide parfait, incompressible non-visqueux, à l'intérieur de cette

tuyère. L'écoulement étant incompressible, il est à divergence nulle et donc le vecteur vitesse $\mathbf{u}$ du fluide peut s'écrire comme le rotationnel d'une fonction ψ dite fonction de courant. On a :

$$\mathbf{u} = \mathbf{rot}\,\psi = (\,\frac{\partial \psi}{\partial y}\,,\,-\frac{\partial \psi}{\partial x}\,) \tag{3.4}$$

Les lignes iso-ψ sont les lignes de courant du fluide. La fonction de courant ψ

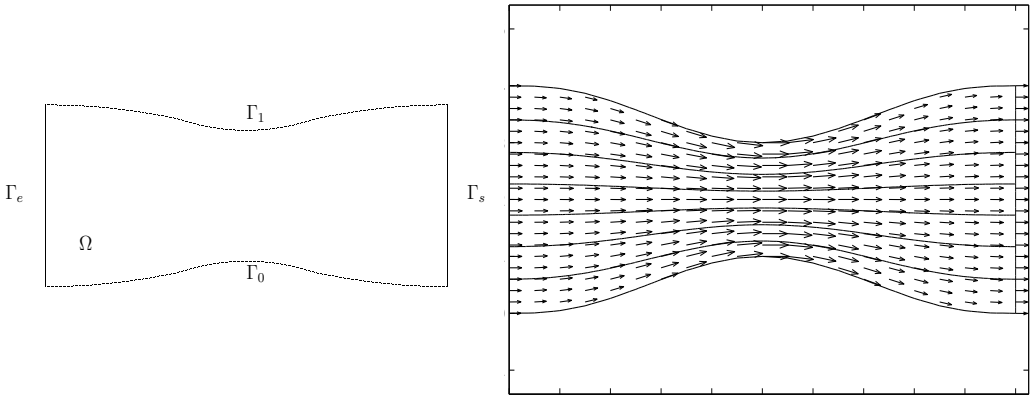


FIGURE 3.3 – Écoulement incompressible.

vérifie également une équation de Poisson, qui s'écrit dans le cas d'un écoulement irrotationnel (cas d'un profil de vitesse horizontal constant en entrée) :

$$\begin{cases} -\,\Delta\,\psi\,(x,y) = 0 \qquad \forall\; x,y \in \Omega \\ \psi_{|\,\Gamma_0}(x,y) = 0 \\ \psi_{|\,\Gamma_1}(x,y) = 1 \\ \psi_{|\,\Gamma_e}(x,y) = y \\ \dfrac{\partial \psi}{\partial n}_{|\,\Gamma_s}(x,y) = 0 \end{cases} \tag{3.5}$$

La même équation se retrouve dans de nombreux domaines de la physique : on l'a vu en thermique, en calcul de la déformation d'une membrane sous une charge, en écoulement potentiel en mécanique des fluides incompressibles. La figure 3.4 présente la distribution du potentiel électrostatique dans un milieu non-conducteur. En résumé, cette équation modélise les problèmes d'équilibre ou de minimisation d'énergie.

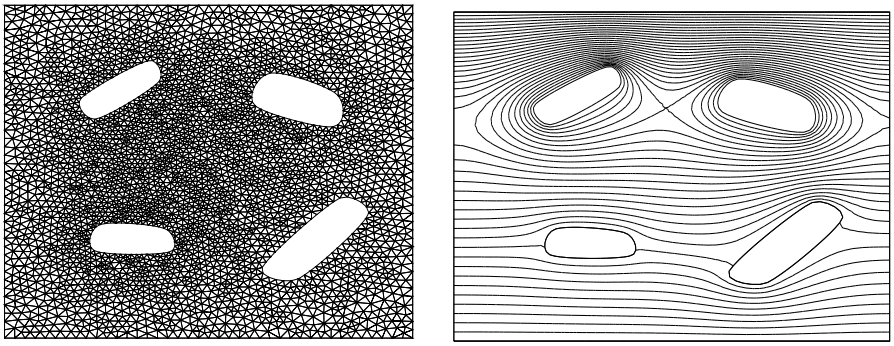


FIGURE 3.4 – Distribution du potentiel électrostatique autour de quatre conducteurs. Les conditions aux limites de Dirichlet se reconnaissent par des iso-valeurs parallèles à la frontière. Celles de Neumann homogènes le sont par des lignes de niveau normales à celle-ci, tandis que les conditions de Neumann non-homogène génèrent des iso-lignes obliques à la paroi.

3.4.4 Principe du maximum et unicité

La solution du problème de Dirichlet (Ω borné) :

$$-\Delta u = 0 \quad x \in \Omega, \quad u = g \quad x \in \partial\Omega.$$

est une fonction harmonique (son laplacien est nul). Les extrema de u sont atteints sur la frontière. Cela s'appelle le principe du maximum. En dimension un, il est facile de le vérifier car u est affine ($u = ax + b$). En dimensions supérieures, supposons que le maximum soit atteint en un point $x \in \overline{\Omega} - \partial\Omega$. Alors, en ce point, toutes les dérivées partielles secondes sont négatives, ce qui est incompatible avec l'équation :

$$\sum_{i=1}^{n} \frac{\partial^2 u}{\partial x_i^2} = 0.$$

Le principe du maximum permet d'obtenir l'unicité de la solution. En effet, considérons u_1 et u_2 deux solutions distinctes. Alors, $u = u_1 - u_2$ sera solution du problème de Dirichlet suivant :

$$-\Delta u = 0 \quad x \in \Omega, \quad u = 0 \quad x \in \partial\Omega.$$

Or, d'après le principe de maximum $u = u_1 - u_2 = 0$.

3.4.5 Propriété de la moyenne

La valeur d'une fonction harmonique en tout point M est la moyenne de ses valeurs sur la surface de toute sphère (cercle en 2D) de centre M. La

démonstration est basée sur la troisième formule de Green (voir annexe B). Cette propriété fondamentale explique pourquoi le laplacien se retrouve dans les modèles d'équilibre. Une grandeur en un point donné (température, potentiel électrique, déplacement) s'établira, à l'équilibre, en l'absence de sollicitations extérieures, à la moyenne des valeurs aux points qui l'entourent. On remarquera, plus loin, que les formules d'approximations, par différences finies par exemple, du laplacien, vérifient également cette propriété de la moyenne.

3.5 Équation parabolique linéaire

L'équation de la chaleur s'obtient par l'ajout d'une dérivée première en temps à l'équation de Poisson vue plus haut.

$$\frac{\partial T(x,t)}{\partial t} - \nabla.(K\nabla T(x,t)) = f(x,t)$$

où K est le coefficient de conductivité thermique et f représente une source de chaleur.

Remarque 3.5.1 *La solution de l'équation de Poisson peut s'obtenir comme limite stationnaire de solutions de l'équation de la chaleur. Cette remarque est très importante en pratique et nous nous en servirons pour d'autres équations.*

Pour que cette équation soit bien posée, il faut ajouter une condition initiale (le champ de température au temps 0, par exemple) en plus des conditions aux limites en espace. L'étude détaillée des schémas numériques pour cette équation est présentée au chapitre 10.

3.6 Équation hyperbolique linéaire

Deux exemples importants d'équations hyperboliques sont, pour le premier ordre, l'équation de transport (chapitre 12) et, pour le second ordre, l'équation des ondes (chapitre 11).

3.6.1 Équation de transport

L'équation de transport (dite également d'advection ou de convection) modélise le transport d'un scalaire passif par un champ de vecteur $\vec{V} \in I\!R^n$:

$$\frac{\partial u}{\partial t} + \vec{V}.\nabla u = f,$$

$u(x,t)$ représente le taux de présence du scalaire transporté en un point et à un instant donnés : par exemple le taux local de présence d'un polluant dans une rivière coulant à une vitesse $\vec{V}(x,t)$. f représente une source (production) ou un puits (dissipation). Cette équation peut se résoudre de façon exacte si l'on connaît les lignes caractéristiques définies par :

$$\frac{d\vec{X}}{dt} = -\vec{V}, \quad X(\vec{T}) = \vec{X_T}.$$

Cette équation différentielle permet de savoir d'où provient une particule se trouvant à $t = T$ en $\vec{x} = \vec{X_T}$. Ainsi, en connaissant les caractéristiques $\vec{X}$, on sait que $u(\vec{x} = \vec{X_T}, t = T) = u(\vec{x} = \vec{X_0}, t = 0)$. En d'autres termes, le scalaire u est convecté le long des caractéristiques. Cette notion peut être mieux comprise en 1D :

$$\frac{\partial u}{\partial T} + c\frac{\partial u}{\partial x} = 0, \quad u(x, t = 0) = u_0(x)$$

ou c représente la vitesse.
Cette équation modélise la propagation du signal u_0 vers la droite si c est positif et vers la gauche si c est négatif. La solution de cette équation, comme on l'a indiqué plus haut, est $u(x,t) = u_0(x - ct)$. Ici les caractéristiques sont les droites $x - ct = cste$.

3.6.2 Advection-diffusion

La convection d'un polluant par une rivière ou un courant marin est en général accompagnée d'un phénomène de diffusion dû à la viscosité des liquides, mais aussi à l'aspect plus ou moins perturbé de l'écoulement (penser à l'encre qui se dilue mieux si l'on agite l'eau). En d'autres termes, la diffusion augmente si $V - \bar{V}$ est grand où $\bar{V}$ est défini par :

$$\bar{V}(x,t) = \frac{1}{T}\int_{t-T/2}^{t+T/2} V(x,\tau)d\tau.$$

On retrouvera l'application de cette idée en modélisation de la turbulence et pour les applications en finance (voir chapitre 12 pour des calculs sur le modèle de Black et Scholes) où cela caractérise la notion de volatilité des marchés. La prise en compte des phénomènes de diffusion se fait par l'ajout d'un laplacien généralisé à l'équation d'advection :

$$\frac{\partial u}{\partial t} + \vec{V}.\nabla u - \nabla.(\nu(x,t)\nabla u) = f, \quad \nu(x,t) > 0, \forall (x,t).$$

Supposons que l'on s'intéresse à la dérive d'une pollution. On connaît la source de la pollution $f = 1$ (f peut être aussi une fonction du temps si la source n'est pas uniforme en temps). Pour fermer le modèle ci-dessus, nous devons donc fournir $\vec{V}$ et ν. Ces données peuvent être le résultat d'autres calculs où bien elles peuvent provenir d'observations. On peut par exemple modéliser ν comme une fonction du gradient de la vitesse.

Un exemple de solution d'équation d'advection-diffusion avec adaptation de maillage est donné au chapitre 18.

Quelques remarques sur le type de l'équation d'advection-diffusion

Considérons la forme monodimensionnelle de l'équation d'advection-diffusion sans terme source :

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = 0, \quad \text{pour}(x,t) \in I\!R \times I\!R^+,$$

On peut montrer par adimensionnement (voir plus loin), que cette équation vérifie un principe de similitude avec un nombre caractéristique appelé Péclet $Pe = cL/\nu > 0$ où L désigne une longueur de référence. Les combinaisons de c, L et ν donnant le même nombre de Péclet ont alors la même solution.

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} - \frac{1}{Pe}\frac{\partial^2 u}{\partial x^2} = 0, \quad \text{pour}(x,t) \in]0,1[\times]0,T[.$$

On voit que si $Pe \to 0$, l'équation tend vers une équation de la chaleur, tandis que si $Pe \to \infty$, on retrouve une équation d'advection pure. Ainsi, l'EDP peut avoir un comportement parabolique ou hyperbolique selon les valeurs de Pe. Par exemple, pour $\nu = 0$ elle est hyperbolique, tandis que pour $c = 0$ elle devient parabolique, et même elliptique si la solution devient stationnaire ($\frac{\partial u}{\partial t} = 0$). Ce changement de nature représente parfois une difficulté lors de la résolution, car les mêmes méthodes numériques ne sont pas nécessairement efficaces pour deux types d'EDP différents.

3.6.3 Équation des ondes

Le deuxième exemple concerne la modélisation de la propagation des ondes dans un milieu continu ou bien les vibrations d'une membrane (un tambour) :

$$\frac{\partial^2 u}{\partial t^2} - c^2\Delta u = 0, \quad (x,t) \in I\!R^n \times I\!R^+,$$

où $u(\vec{x},t)$ représente ici le déplacement du point $\vec{x}$ en t. En une dimension d'espace, cette équation s'écrit :

$$\frac{\partial^2 u}{\partial t^2} - c^2\frac{\partial^2 u}{\partial x^2} = 0, \quad (x,t) \in R \times R^+.$$

C'est l'équation de la corde vibrante. La solution de cette équation s'écrit :

$$u(x,t) = v(x-ct) + w(x+ct).$$

Il faut ajouter, dans le cas de cette équation du second ordre en temps, deux conditions initiales. On peut fixer la position de la corde et sa vitesse au temps initial.

Remarque 3.6.1 *On retrouve ici les deux façons de produire un son à l'aide d'un instrument à cordes, soit en déplaçant la corde à partir de sa position d'équilibre (instruments à cordes pincées), soit en lui impulsant une vitesse initiale (instruments à cordes frappées).*

Équation d'Helmholtz

Le caractère transitoire et non-amorti (absence de dérivée première en temps) de l'équation des ondes rend difficile la recherche des mouvements stationnaires. Cependant, en remarquant la linéarité de l'équation des ondes, on peut utiliser le principe de superposition des solutions pour chercher les mouvements stationnaires. En effet, les solutions monochromatiques (i.e. $u(x,t) = u(x)exp(i\omega t)$) de cette équation vérifient l'équation de Helmholtz :

$$u + \frac{c^2}{\omega^2}\Delta u = 0.$$

La résolution de cette équation nécessite, à cause du signe du laplacien, quelques précautions, notamment de travailler en variables complexes. Une difficulté lors de la résolution de l'équation des ondes (ou d'Helmholtz) est la dépendance du maillage par rapport à la fréquence envisagée (il faut compter plusieurs mailles par longueur d'onde dans toutes les directions, voir Fig.3.5). En dimension trois, cela implique un facteur 8 en nombre de points chaque fois que l'on double la fréquence.

Système équivalent du premier ordre

On peut réécrire l'équation des ondes comme un système du premier ordre. En particulier, dans le cas de coefficients non-constants, l'écriture sous forme de système du premier ordre permet d'utiliser les techniques mises au point pour l'advection (voir 12).

En effet, il est facile de vérifier que le système suivant :

$$\frac{\partial U}{\partial t} + A\frac{\partial U}{\partial x} = 0$$

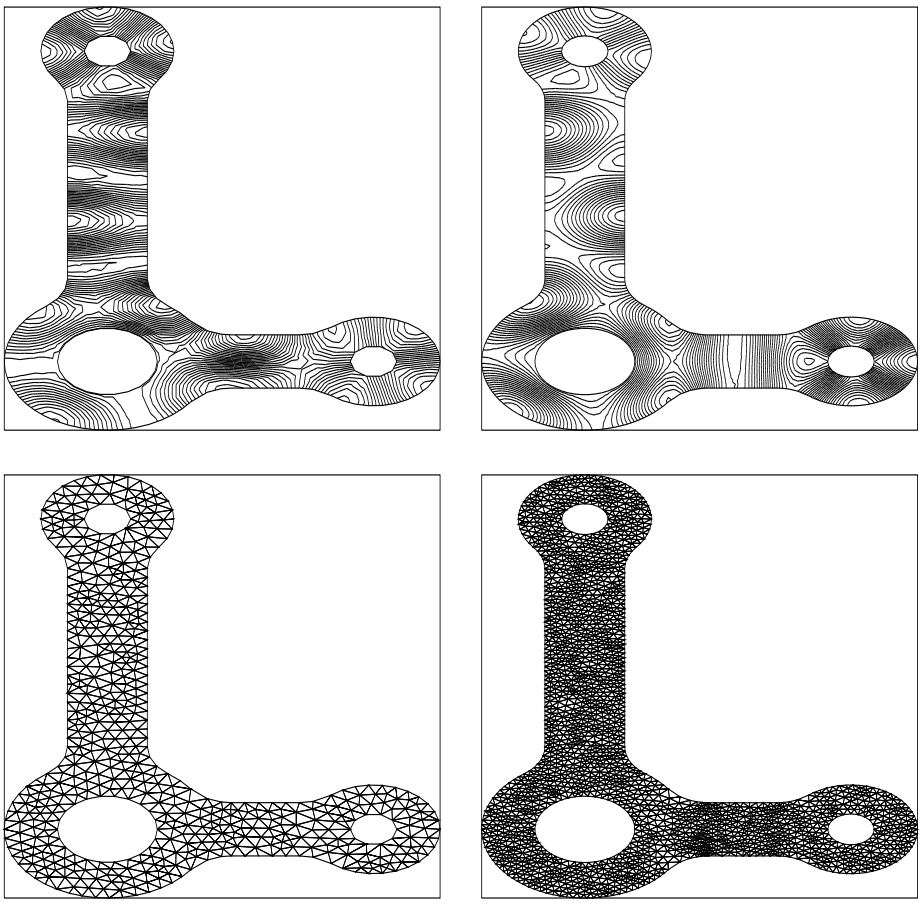


FIGURE 3.5 – Effet du raffinement du maillage pour le calcul de la propagation des vibrations dans une pièce mécanique. La solution calculée à gauche correspond à un maillage trop grossier. Le maillage doit être compatible avec la fréquence du phénomène que l'on souhaite capter.

où

$$U = (u, v)^t, \quad \text{et} \quad A = \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix},$$

est équivalent à $\dfrac{\partial^2 u}{\partial t^2} - c^2 \dfrac{\partial^2 u}{\partial x^2} = 0$. On a en effet,

$$\frac{\partial^2 u}{\partial t^2} + c\frac{\partial^2 v}{\partial t \partial x} = 0 \quad \text{et} \quad \frac{\partial^2 v}{\partial t \partial x} + c\frac{\partial^2 u}{\partial x^2} = 0$$

La matrice A a deux valeurs propres réelles $\lambda_1 = -c$ et $\lambda_2 = c$ (système hyperbolique). Les vecteurs propres associés sont $r_1 = (1, -1)^t$ et $r_2 = (1, 1)^t$. U se décompose, sur la base (r_1, r_2), en $U = a_1 r_1 + a_2 r_2$, où $a_1 = (u - v)/2$ et $a_2 = (u + v)/2$ sont appelés "invariants de Riemann" et vérifient deux équations de transport scalaires :

$$\frac{\partial a_1}{\partial t} - c\frac{\partial a_1}{\partial x} = \frac{\partial a_2}{\partial t} + c\frac{\partial a_2}{\partial x} = 0,$$

qui représentent la convection sur la caractéristique "axe réel" vers la gauche et la droite à la vitesse $\pm c$. Cette formulation d'ordre un est parfois utile, en contrôle en particulier.

3.7 Systèmes d'EDP

Un système d'EDP est un ensemble couplé d'EDP pour les composantes d'un vecteur représentant différentes grandeurs physiques. On présente deux exemples de systèmes pour les cas elliptique et parabolique.

Le système d'élasticité linéaire stationnaire (voir chapitre 9) est une extension de l'équation (3.1) :

$$\nabla.\sigma = f,$$

où $\nabla. = div$ et σ est le tenseur des contraintes (i.e. une relation matricielle entre les déformations et les efforts) et f désigne une densité volumique de force.
Cette équation modélise l'équilibre d'un matériau élastique soumis à un champ de force. Un exemple simple de tenseur des contraintes est le tenseur des contraintes linéaires Newtonien où $\sigma = \mu(\nabla u + \nabla u^t)$ avec $\nabla = grad$ et $\mu > 0$.

Le système des équations de Navier-Stokes

$$\begin{cases} \dfrac{\partial u}{\partial t} + u\, grad(u) - \nu \Delta u = -\dfrac{1}{\rho} grad(P) \\ div(u) = 0 \end{cases}$$

où u désigne le vecteur vitesse, P la pression, ν la viscosité, modélise un écoulement incompressible (voir annexe B pour la définition des opérateurs gradients et divergence).

3.8 Conditions aux limites et conditions initiales

Nous rencontrerons principalement les types suivants de conditions aux limites :

1. Les conditions aux limites de Dirichlet où la valeur de la variable est prescrite sur la frontière ou sur une partie de celle-ci :

$$u(x \in \partial\Omega, t) = g(x, t).$$

   Si $g = 0$, on obtient une condition de Dirichlet homogène. Les conditions aux limites de Dirichlet permettent par exemple de fixer un déplacement ou d'imposer la température T à la surface d'un corps à une valeur $T_0(x, t)$. On précise que T_0 peut être fonction du temps et de l 'espace.

2. Les conditions aux limites de Neumann où l'on fixe la valeur du gradient de la variable dans la direction normale à la frontière :

$$\nabla u.n = \frac{\partial u}{\partial n} = g.$$

   Cette condition permet par exemple d'imposer un flux de chaleur sur une partie du domaine. Si $g = 0$, on obtient la condition de Neumann homogène. Cette condition représente un flux nul, mais est également utile pour représenter une symétrie de la solution ou une condition de sortie libre. g peut être également fonction du temps et du point du domaine de calcul. On peut étendre cette condition au cas vectoriel et par exemple imposer des contraintes sur la frontière pour un corps élastique :

$$\sigma(x \in \partial\Omega, t).n(x) = F(x).$$

3. Les conditions limites de Robin ou Fourier ou mixtes qui font intervenir une combinaison des deux conditions précédentes :

$$\alpha\frac{\partial u}{\partial n} + \beta u = g.$$

   Elles les généralisent. Les conditions de Dirichlet et de Neumann en sont des cas particuliers. Ces conditions de Fourier permettent de représenter les flux en fonction des valeurs de la variable sur la frontière. Elles sont très utiles en modélisation. Elles permettent, par exemple en thermique, de modéliser un échange convectif sur une paroi, en mécanique des fluides, de modéliser l'effet d'une paroi sur l'écoulement ou bien en propagation d'ondes de modéliser les effets d'un matériau absorbant.

4. Les conditions de rayonnement de type corps noir qui s'expriment sous la forme générale :
$$\frac{\partial u}{\partial n} + \alpha u^{\beta} = g,$$
voir chapitre 13.

3.8.1 Fonctions de paroi

Les conditions de type Robin sont très utiles en mécanique des fluides. Elles permettent de modéliser l'effet de l'adhérence à la paroi sur la vitesse au voisinage de celle-ci. Considérons le développement de Taylor suivant pour la solution au voisinage d'une paroi située en $x = 0$, dans la direction normale à la paroi (la normale est notée n) :

$$u(0) = u(\delta) + \partial_n u(\delta)\delta + \partial^2_{nn}(\delta)\frac{\delta^2}{2} + \delta^2 o(1),$$

où l'on a noté : $\partial_n u = \frac{\partial u}{\partial n}$ et $\partial^2_{nn} u = \frac{\partial^2 u}{\partial n^2}$ et supposons que l'on connaisse la valeur $u(0)$ à imposer en $x = 0$. On peut ainsi déduire une condition de Fourier à prescrire en $x = \delta$:

$$u(0) - u(\delta) = \partial_n u(\delta)\delta, \quad \text{à l'ordre 1.}$$

L'avantage de ce type de condition est de permettre d'éviter un maillage très fin de la région entre $x = 0$ et $x = \delta$ où la variable peut varier rapidement. Supposons que l'on soit intéressé par l'évaluation de l'effet d'une peinture absorbante pour améliorer (diminuer ou pourquoi pas augmenter pour un avion civil) la signature radar d'un avion. Bien sûr, il est hors de question de mailler la peinture. On va essayer de modéliser sa présence par l'introduction d'une condition aux limites de Fourier.

Il arrive de plus que l'on puisse avoir par un calcul supplémentaire des informations sur les dérivées supérieures (par exemple la dérivée seconde). Considérons l'EDP :

$$-\Delta u = f \quad \forall x \in \Omega \in I\!R^2, \quad u(x \in \partial\Omega) = g.$$

Au voisinage de la frontière $\partial\Omega$, si l'on peut définir la normale locale à la paroi de façon unique, on a

$$\partial^2_{nn} u = -f + \partial^2_{\tau\tau} u, \quad (n, \tau) \text{ repère normal local, avec } n \text{ normale et } \tau \text{ tangente.}$$

Ainsi, on peut prescrire une condition d'ordre plus élevé en utilisant $\partial^2_{\tau\tau} u$ provenant du calcul. De plus, il arrive que les variations soient négligeables dans

certaines directions (si par exemple $\partial^2_{\tau\tau} u << \partial^2_{nn} u$). Ceci permet d'écrire :

$$u(0) - u(\delta) = \partial_n u(\delta)\delta + f(\delta)\frac{\delta^2}{2}, \quad \text{à l'ordre 2.}$$

Nous verrons au chapitre 13 des exemples d'utilisation de ces conditions équivalentes pour la prise en compte des déformations de domaines.

3.8.2 Conditions aux limites à l'infini

Dans les problèmes en milieu infini, que l'on rencontre en particulier en propagation d'ondes ou en calcul d'écoulements externes en milieu ouvert, se pose la question du choix des conditions aux limites adéquates à imposer aux bornes réelles du domaine de calcul. En effet, sauf si l'on utilise des méthodes intégrales, on sera obligé pratiquement de limiter le domaine de calcul à l'intérieur d'une boîte artificielle. Le choix des conditions aux limites sur les bords de cette boîte s'avère crucial pour la qualité des résultats. En propagation d'ondes, il faut éviter les réflexions parasites sur les bords "infinis". De même, en mécanique des fluides, le fluide est supposé s'écouler librement à l'aval. Les conditions de sorties doivent être appropriées. Ce problème des conditions aux limites en "sortie" est l'un des plus délicats de la modélisation numérique. Parmi les techniques simples, on peut citer l'utilisation de conditions de Neumann homogènes suffisantes dans certaines applications et, surtout, les méthodes de caractéristiques dans lesquelles les conditions en sortie sont déterminées en remontant le courant ou l'onde. Á l'inverse, les conditions de Dirichlet produisent des réflexions parasites.

3.8.3 Approche pseudo-instationnaire

On peut résoudre les EDP elliptiques comme limites stationnaires d'EDP paraboliques. Ceci permet, en particulier, de prendre en compte facilement les conditions de Dirichlet non-homogènes. Plus généralement, considérons le problème modèle suivant :

$$\left\{ \begin{array}{l} -\nabla.(F(u)) = f, \quad \text{dans} \quad \Omega \\ u = g_1 \quad \text{sur} \quad \Gamma_1, \quad F(u).n = g_2 \quad \text{sur} \quad \Gamma_2 \end{array} \right. \tag{3.6}$$

où F est une fonction vectorielle donnée. La solution de ce problème peut être recherchée comme limite stationnaire du modèle pseudo-instationnaire suivant :

$$\left\{ \begin{array}{l} \dfrac{\partial u}{\partial t} - \nabla.(F(u)) = f, \quad \text{dans} \quad \Omega \\ u = g_1 \quad \text{sur} \quad \Gamma_1, \quad F(u).n = g_2 \quad \text{sur} \quad \Gamma_2 \\ u(t=0) = u^0 \quad \text{dans} \quad \Omega \end{array} \right. \tag{3.7}$$

Le temps n'a pas de signification physique. On construit la suite u^n des solutions approchées de la façon suivante.

Soit u^0 donné telle que $u^0 = g_1$ sur Γ_1 et $F(u^0).n = g_2$ sur Γ_2 (ceci s'appelle une condition initiale admissible), on considère l'itération :

$$\left\{ \begin{array}{l} \dfrac{\delta u}{\delta t} + \nabla.(F'_u(u^n)\delta u) = f - \nabla.(F(u^n)), \quad \text{dans} \quad \Omega \\ \delta u = 0 \quad \text{sur} \quad \Gamma_1 \quad (F'_u(u^n)\delta u).n = g_2 - F(u^n).n = 0 \quad \text{sur} \quad \Gamma_2 \\ u^{n+1} = u^n + \delta u \end{array} \right. \tag{3.8}$$

dans laquelle, δt est le pas de temps et F'_u représente la matrice Jacobienne de F.

On n'a plus, ainsi, que des conditions homogènes à prendre en compte.

Dans le cas $F = \nabla u$, on retrouve le problème de Poisson :

$$\left\{ \begin{array}{l} -\Delta u = f \quad \text{dans} \quad \Omega \\ u = g_1 \quad \text{sur} \quad \Gamma_1, \quad \dfrac{\partial u}{\partial n} = g_2 \quad \text{sur} \quad \Gamma_2 \end{array} \right. \tag{3.9}$$

dont la forme pseudo-instationnaire s'écrit :

$$\left\{ \begin{array}{l} \dfrac{\delta u}{\delta t} - \Delta \delta u = f + \Delta u^n, \quad \text{dans} \quad \Omega \\ \delta u = 0, \quad \text{sur} \quad \Gamma_1, \quad \dfrac{\partial \delta u}{\partial n} = 0, \quad \text{sur} \quad \Gamma_2 \\ u^{n+1} = u^n + \delta u \end{array} \right. \tag{3.10}$$

On verra plus loin, pour chaque classe de méthodes de discrétisation numérique, comment prendre en compte les différents types de conditions aux limites et leurs incidences sur les systèmes matriciels obtenus.

3.9 L'adimensionnement et la similitude

Considérons l'équation suivante modélisant la convection-diffusion de la quantité ρ par le champ de vitesse u :

$$\partial_t \rho + \nabla_x.(\rho u) - \nabla_x.(\mu \nabla_x \rho) = 0, \quad (x,t) \in \Omega \times (0,T),$$

où ∇_x indique que les dérivées sont prises par rapport aux variables d'espace $x \in \Omega$. Soient ρ_0, u_0, μ_0 des quantités de référence pour la quantité advectée ρ, la vitesse et la viscosité, L_0 une longueur caractéristique et t_0 le temps

caractéristique défini par $t_0 = L_0/u_0$. On notera $\overline{r}$, $\overline{u}$, $\overline{\mu}$, $\overline{t}$, $\overline{x}$ les variables adimensionnées correspondant à ρ, u, μ, t et x définies par :

$$\begin{aligned}
\overline{r} &= \frac{\rho}{\rho_0}\,, \\
\overline{u} &= \frac{u}{u_0}\,, \\
\overline{\mu} &= \frac{\mu}{\mu_0} \\
\overline{t} &= \frac{t}{t_0}\,, \\
\overline{x} &= \frac{x}{L_0}\,.
\end{aligned}$$

Le changement de variables ci-dessus conduit à l'équation :

$$\frac{L_0}{t_0 u_0}\partial_{\overline{t}}\overline{r} + \nabla_{\overline{x}}(\overline{r}\,\overline{u}) - \nabla_{\overline{x}}.(\frac{1}{Pe}\nabla_{\overline{x}}\overline{r}) = 0,$$

qui, par définition de t_0, se réduit à :

$$\partial_{\overline{t}}\overline{r} + \nabla_{\overline{x}}(\overline{r}\,\overline{u}) - \nabla_{\overline{x}}.(\frac{1}{Pe}\nabla_{\overline{x}}\overline{r}) = 0,$$

avec $Pe = u_0 L_0/\mu_0$. Ici $\nabla_{\overline{x}}$ représente les dérivées par rapport aux variables d'espace adimensionnées $\overline{x} \in \overline{\Omega}$. Ainsi, il y a similitude entre les solutions des configurations ayant le même nombre de Péclet. Ceci permet en particulier d'augmenter la précision des calculs en minimisant les erreurs d'arrondis numériques dans la mesure où l'on peut faire les opérations sur des grandeurs adimensionnées voisines de l'unité.

Remarque 3.9.1 *Travailler complètement en variables adimensionnées n'est pas toujours possible ni très pratique lors du couplage de plusieurs modèles, comme on le verra plus loin. En effet, chaque modèle a son propre adimensionnement. Les outils boîte-noire de simulation intègrent souvent un adimensionnement, caché à l'utilisateur. On préfèrera alors rester en variables dimensionnées au niveau des interfaces des outils de simulation et donc entre les modèles.*

Chapitre 4

Introduction aux méthodes de discrétisation des équations aux dérivées partielles

4.1 Présentation générale

En vue du passage d'un problème exact (continu) au problème approché (discret), on dispose de plusieurs techniques concurrentes et complémentaires : les différences finies, les éléments finis et les volumes finis. Chacune de ces trois méthodes correspond à une formulation différente des équations de la physique :

- équilibre des forces en chaque point pour les différences finies
- minimisation de l'énergie ou principe des travaux virtuels pour les éléments finis
- loi de conservation et calcul des flux pour la méthode des volumes finis.

Examinons rapidement les avantages et les inconvénients de chacune de ces trois méthodes.

4.1.1 Différences finies

La méthode des différences finies consiste à remplacer les dérivées apparaissant dans le problème continu par des différences divisées ou combinaisons de valeurs ponctuelles de la fonction en un nombre fini de points discrets ou noeuds du maillage.

Avantages : grande simplicité d'écriture et faible coût de calcul.

Inconvénients : Limitation de la géométrie des domaines de calculs, difficultés de prise en compte des conditions aux limites portant sur les dérivées ou les gradients de l'inconnue et en général absence de résultats de majoration d'erreurs.

4.1.2 Éléments finis

La méthode des éléments finis consiste à approcher, dans un sous-espace de dimension finie, un problème écrit sous forme variationnelle (comme minimisation de l'énergie, en général) dans un espace de dimension infinie. La solution approchée est dans ce cas une fonction déterminée par un nombre fini de paramètres comme, par exemple, ses valeurs en certains points (les noeuds du maillage).
Avantages : Traitement possible de géométries complexes, détermination plus naturelle des conditions aux limites, possibilité de démonstrations mathématiques de convergence et de majoration d'erreurs.
Inconvénients : Complexité de mise en oeuvre et coût en temps de calcul et en mémoire.

4.1.3 Volumes finis

La méthode des volumes finis intègre, sur des volumes élémentaires de forme simple, les équations écrites sous forme de loi de conservation. Elle fournit ainsi de manière naturelle des approximations discrètes conservatives et est donc particulièrement bien adaptée aux équations de la mécanique des fluides : équation de conservation de la masse, équation de conservation de la quantité de mouvement, équation de conservation de l'énergie.
Sa mise en oeuvre est simple si les volumes élémentaires sont des rectangles (ou des parallélépipèdes rectangles en dimension 3). Cependant la méthode des volumes finis permet d'utiliser des volumes élémentaires de forme quelconque, donc de traiter des géométries complexes, ce qui est un avantage sur les différences finies. Il existe une grande variété de méthodes selon le choix de la géométrie des volumes élémentaires et des formules de calcul des flux. Par contre, on dispose de peu de résultats théoriques de convergence.

Nous avons choisi, par souci pédagogique, de commencer par présenter ces trois méthodes dans le cas simple de problèmes en dimension un d'espace.

4.2 L'approche différences finies en dimension un

Toutes les méthodes numériques présupposent la discrétisation du domaine géométrique afin de passer d'un problème continu à une infinité d'inconnues à un problème discret ne comptant qu'un nombre fini d'inconnues.
Dans le cas des différences finies en dimension un, on discrétise l'intervalle continu $[a, b]$ en un nombre fini de points x_i.

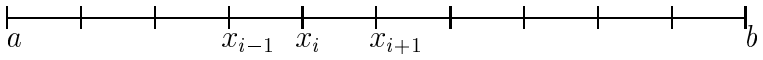


FIGURE 4.1 – Discrétisation en différences finies d'un segment [a,b].

On remplace ainsi le problème continu par celui de la recherche de valeurs approchées u_i des solutions exactes $u(x_i)$ aux points x_i de la discrétisation.

Mais on ne peut plus, dans ce cas, conserver les opérateurs de dérivation qui s'appliquent à des fonctions continues. On les remplace par des analogues discrets, les différences divisées ou différences finies.

Le type de conditions aux limites conditionne le nombre d'inconnues du problème discret. Dans le cas de conditions de Dirichlet, la solution est fixée, et donc en ces points, les valeurs sont connues. Dans tous les autres cas de conditions aux limites, la valeur de la solution reste inconnue et fait donc partie du vecteur inconnu.

4.2.1 Quelques formules simples d'approximation des dérivées par des différences divisées.

Pour la dérivée première :

— différence divisée progressive d'ordre un :
Le développement limité :

$$u'(x_i) = \frac{u(x_i + h) - u(x_i)}{h} - \frac{h}{2}u''(\xi_i)$$

conduit à l'approximation suivante :

$$u'(x_i) = \frac{du}{dx}(x_i) \simeq \frac{u_{i+1} - u_i}{x_{i+1} - x_i}$$

— différence divisée progressive d'ordre deux :
Pour des pas réguliers de longueur h, le développement limité :

$$u'(x_i) = \frac{-u(x_i + 2h) + 4u(x_i + h) - 3u(x_i)}{2h} + \frac{h^2}{3}u'''(\xi_i)$$

donne la formule d'ordre deux progressive suivante :

$$u'(x_i) = \frac{du}{dx}(x_i) \simeq \frac{-u_{i+2} + 4u_{i+1} - 3u_i}{2h}$$

— différence divisée régressive d'ordre un :

De même le développement limité :

$$u'(x_i) = \frac{u(x_i) - u(x_i - h)}{h} + \frac{h}{2}u''(\eta_i)$$

donne :

$$u'(x_i) = \frac{du}{dx}(x_i) \simeq \frac{u_i - u_{i-1}}{x_i - x_{i-1}}$$

— différence divisée régressive d'ordre deux :
On obtient également une formule régressive d'ordre deux :

$$u'(x_i) = \frac{du}{dx}(x_i) \simeq \frac{3u_i - 4u_{i-1} + u_{i-2}}{2h}$$

— différence divisée centrée : on a

$$u'(x_i) = \frac{u(x_i + h) - u(x_i - h)}{2h} - \frac{h^2}{6}u'''(\theta_i)$$

ou également

$$u'(x_i) = \frac{u(x_i + \frac{h}{2}) - u(x_i - \frac{h}{2})}{h} - \frac{h^2}{24}u'''(\theta_i) \tag{4.1}$$

Ce qui conduit, dans le cas de discrétisations uniformes de pas constant h, à :

$$u'(x_i) = \frac{du}{dx}(x_i) \simeq \frac{u_{i+1} - u_{i-1}}{2h} \quad ou \quad \frac{u_{i+1/2} - u_{i-1/2}}{h}$$

On a noté $u_{i+1/2}$ et $u_{i-1/2}$ les valeurs approchées de u aux points $x_i + \frac{h}{2}$ et $x_i - \frac{h}{2}$ respectivement.

Pour la dérivée seconde :

— différence divisée centrée
Dans le cas particulier de points x_i régulièrement espacés d'un pas h uniforme, on retrouve en utilisant :

$$u''(x_i) = \frac{u'(x_i + \frac{h}{2}) - u'(x_i - \frac{h}{2})}{h} - \frac{h^2}{24}u^{(4)}(\theta_i)$$

et 4.1 :

$$u''(x_i) = \frac{u(x_i + h) - 2\,u(x_i) + u(x_i - h)}{h^2} - \frac{h^2}{12}u^{(4)}(\theta_i)$$

d'où : la discrétisation centrée classique de la dérivée seconde

$$u''(x_i) = \frac{d^2u}{dx^2}(x_i) \simeq \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} \tag{4.2}$$

Pour la dérivée troisième :

— différence divisée progressive

Dans le cas particulier de points x_i régulièrement espacés d'un pas h uniforme, on obtient la formule décentrée progressive (d'ordre un) :

$$u'''(x_i) = \frac{d^3u}{dx^3}(x_i) \simeq \frac{u_{i+3} - 3u_{i+2} + 3u_{i+1} - u_i}{h^3}$$

On obtiendrait de même une différence divisée régressive.

— différences divisées centrées

On a deux formules possibles. L'une utilisant des points milieux de segments

$$u'''(x_i) = \frac{d^3u}{dx^3}(x_i) \simeq \frac{u_{i+3/2} - 3u_{i+1/2} + 3u_{i-1/2} - u_{i-3/2}}{h^3}$$

L'autre utilisant les noeuds du maillages.

$$u'''(x_i) = \frac{d^3u}{dx^3}(x_i) \simeq \frac{u_{i+2} - 2u_{i+1} + 2u_{i-1} - u_{i-2}}{h^3}$$

Ces deux formules centrées sont d'ordre deux.

Pour la dérivée quatrième :

— différence divisée centrée

Dans le cas particulier de points x_i régulièrement espacés d'un pas h uniforme, on retrouve en utilisant :

$$u''(x_i) = \frac{u'(x_i + \frac{h}{2}) - u'(x_i - \frac{h}{2})}{h} - \frac{h^2}{24}u^{(4)}(\theta_i)$$

deux fois la discrétisation centrée classique d'ordre deux de la dérivée quatrième

$$u^{IV}(x_i) = \frac{d^4u}{dx^4}(x_i) \simeq \frac{u_{i+2} - 4u_{i+1} + 6u_i - 4u_{i-1} + u_{i-2}}{h^4}$$

On peut observer que l'on retrouve dans ces formules les coefficients du développement du binôme. Il existe un grand nombre de formules de différences divisées, le bon choix n'est pas toujours évident. On peut également obtenir des approximations des dérivées par des techniques de différences finis dites implicites, ou de Padé (voir un exemple au chapitre 12)

4.2.2 Applications en dimension un

Problème de Dirichlet

$$\begin{cases} -u''(x) = f(x) & a < x < b \\ u(a) = \alpha \quad u(b) = \beta \end{cases} \tag{4.3}$$

Interprétations physiques

— barre élastique sous un chargement axial.
— corde élastique soumise à un chargement transverse.
— conduction thermique dans une barre.

On utilise la discrétisation centrée classique de la dérivée seconde

$$u''(x_i) = \frac{d^2u}{dx^2}(x_i) \simeq \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}$$

On obtient ainsi le système d'équations linéaires suivant dont la résolution donne les valeurs u_i de la solution approchée du problème 4.3

$$\begin{cases} -\dfrac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = f_i \qquad \text{pour} \quad i = 1, N-1 \\ \text{avec} \quad u_0 = \alpha \quad u_N = \beta \end{cases} \tag{4.4}$$

Ce qui s'écrit sous forme matricielle :

$$\frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & \cdots & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \\ u_i \\ \\ u_{N-2} \\ u_{N-1} \end{pmatrix} = \begin{pmatrix} f_1 + \alpha/h^2 \\ f_2 \\ \\ f_i \\ \\ f_{N-2} \\ f_{N-1} + \beta/h^2 \end{pmatrix} \tag{4.5}$$

Il ne reste alors plus qu'à résoudre ce système linéaire par des techniques standard de factorisation (méthodes de Gauss LU ou méthode de Choleski LL^T, voir chapitre 2, paragraphe 2.7).

Problème mixte Dirichlet-Neumann

Soit, maintenant, le problème

$$\begin{cases} -u''(x) = f(x) \qquad a < x < b \\ u(a) = \alpha \quad u'(b) = \beta \end{cases} \tag{4.6}$$

où l'on a, cette fois, une condition de Neumann en b.
Les modifications du problème discrétisé sont les suivantes. Tout d'abord, le nombre d'inconnues a changé. Il y a une inconnue au point b. En effet, la donnée de $u'(b)$ ne dit rien de la valeur de $u(b)$. C'est donc une nouvelle inconnue du problème. Le problème discret a donc maintenant, sur la base du même maillage,

N inconnues u_i pour $i = 1$ à N. D'autre part, il faut proposer une formule discrétisée de la condition de Neumann $u'(b) = \beta$. Or, on l'a vu, plusieurs choix sont possibles pour approcher une dérivée première. C'est un des écueils des méthodes de différences finies qu'elles ne donnent pas de façon naturelle une bonne approximation des conditions de Neumann. Ici, il y a deux choix possibles :

— On peut remplacer au point b la dérivée $u'(b)$ par

$$u'(b) \simeq \frac{u_{N+1} - u_N}{h}$$

formule qui, bien que seulement d'ordre un, est cohérente avec la modélisation d'un flux ou d'un effort externe, et qui maintient la symétrie du système matriciel. C'est en particulier la bonne modélisation par différences finies de l'effet d'un effort concentré au point L.

— On peut remplacer au point b la dérivée $u'(b)$ par

$$u'(b) \simeq \frac{u_{N+1} - u_{N-1}}{2h}$$

Cette formule est d'ordre 2. Donc elle permettra de conserver globalement l'ordre 2 de l'approximation. Elle est cohérente avec la modélisation d'une condition de symétrie par une condition de Neumann. Pour retrouver la symétrie de la matrice, il faudra diviser la dernière ligne du système par 2. Ceci conduit d'ailleurs précisément à la formule que l'on obtiendra par une méthode d'éléments finis P1.

On obtient ainsi dans ce cas :

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & \cdots & -1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_i \\ u_{N-1} \\ u_N \end{pmatrix} = \begin{pmatrix} f_1 + \alpha/h^2 \\ f_2 \\ f_i \\ f_{N-1} \\ f_N/2 + \beta/h \end{pmatrix}$$

4.3 Approximation par différences finies en dimension supérieures

La méthode des différences finies a l'avantage d'être facile à exposer. Nous pouvons donc présenter son application aux problèmes multidimensionnels. Cela ne sera pas le cas des méthodes d'éléments finis pour lesquelles un exposé plus long, réparti sur plusieurs chapitres, sera nécessaire.

Comme en dimension un, la première étape consiste à discrétiser le domaine. C'est dans l'application aux problèmes bidimensionnels et tridimensionnels que la méthode des différences finies présente sa plus sévère limitation. En effet elle n'est bien adaptée qu'à la discrétisation de domaines rectangulaires ou parallélépipèdiques par des maillages formées de grilles perpendiculaires. Les dérivées partielles dans chaque direction d'axe étant approchées comme les dérivées en dimension un.

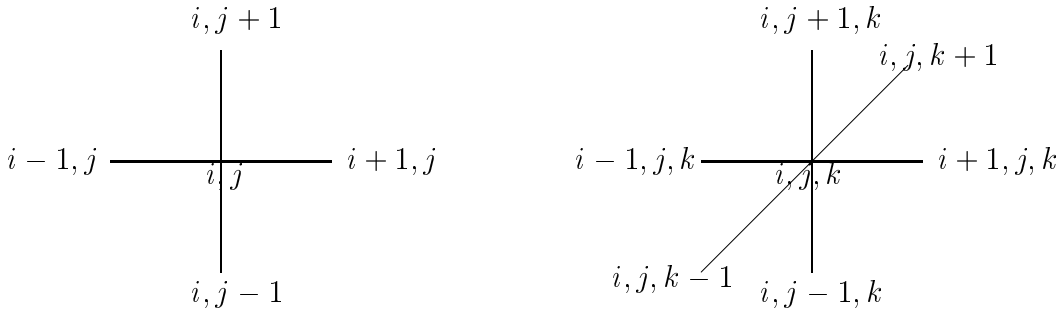


FIGURE 4.2 – Grilles différences finies bidimensionnelles et tridimensionnelles.

4.3.1 Discrétisation géométrique

Dans le cas de domaines rectangulaires (ou parallélépipédiques en dimension 3) de côtés parallèles aux axes, on construit une grille de discrétisation en différences finies par quadrillage selon les deux (ou trois) directions d'axes. On notera Δx le pas de discrétisation selon x et de même Δy et Δz les pas de discrétisation en y et z. On obtient ainsi aux intersections des lignes du quadrillage les noeuds de coordonnées (x_i, y_j, z_k) du maillage en différences finies. Cette technique de maillage est généralisable aux assemblages de rectangles (ou de parallélépipèdes) ainsi qu'aux domaines se ramenant par bijection régulière à un rectangle (ou un parallélépipède). Par contre dans le cas de géométries complexes les discrétisations par éléments finis sont mieux adaptées.

4.3.2 Quelques formules simples d'approximation des dérivées partielles par différences finies

Notons, en dimension deux, $u_{i,j}$ l'approximation de la valeur exacte $u(x_i, y_j)$ pour le point d'indice i, j de la grille.

Pour les dérivées partielles premières :

— différences divisées progressives : on a les approximations suivantes :

$$\frac{\partial u}{\partial x}(x_i, y_j) \simeq \frac{u_{i+1,j} - u_{i,j}}{\Delta x} \qquad \frac{\partial u}{\partial y}(x_i, y_j) \simeq \frac{u_{i,j+1} - u_{i,j}}{\Delta y}$$

— différences divisées régressives : on considère cette fois les approximations :

$$\frac{\partial u}{\partial x}(x_i, y_j) \simeq \frac{u_{i,j} - u_{i-1,j}}{\Delta x} \qquad \frac{\partial u}{\partial y}(x_i, y_j) \simeq \frac{u_{i,j} - u_{i,j-1}}{\Delta y}$$

— différences divisées centrées : on obtient (comme en dimension un) une approximation du second ordre :

$$\frac{\partial u}{\partial x}(x_i, y_j) \simeq \frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x} \quad \frac{\partial u}{\partial y}(x_i, y_j) \simeq \frac{u_{i,j+1} - u_{i,j-1}}{2\Delta y}$$

et avec un demi-pas :

$$\frac{\partial u}{\partial x}(x_i, y_j) \simeq \frac{u_{i+1/2,j} - u_{i-1/2,j}}{\Delta x} \quad \frac{\partial u}{\partial y}(x_i, y_j) \simeq \frac{u_{i,j+1/2} - u_{i,j-1/2}}{\Delta y}$$

On en déduit, par double différentiation, l'approximation centrée d'ordre deux du laplacien

$$-\Delta u(x_i, y_j) \simeq \frac{-u_{i+1,j} + 2\,u_{i,j} - u_{i-1,j}}{\Delta x^2} + \frac{-u_{i,j+1} + 2\,u_{i,j} - u_{i,j-1}}{\Delta y^2} \tag{4.7}$$

Et plus généralement l'approximation de l'équation l'opérateur $div(\sigma grad)$ pour σ variable :

$$\begin{aligned} -div(\sigma gradu)(x_i, y_j) \simeq & -\frac{\sigma(x_{i+1/2}, y_j)\,(u_{i+1,j} - u_{i,j}) + \sigma(x_{i-1/2}, y_j)\,(u_{i,j} - u_{i-1,j})}{\Delta x^2} \\ & -\frac{\sigma(x_i, y_{j+1/2})(u_{i,j+1} - u_{i,j}) + \sigma(x_i, y_{j-1/2})(u_{i,j} - u_{i,j-1})}{\Delta y^2} \end{aligned}$$

Conditions aux limites de Neumann

On doit ajouter la prise en compte des conditions aux limites. Pour les conditions de Dirichlet, il suffit de fixer les valeurs de $u_{i,j}$ correspondant aux valeurs données sur la frontière Γ_d. Pour les conditions de Neumann, on doit discrétiser

$$\frac{\partial u}{\partial n}_{|\,\Gamma_n} = g$$

Il y a, comme pour la dérivée en dimension un, plusieurs choix possibles pour approcher la dérivée normale en différences finies.

— Le choix décentré d'ordre un conserve la symétrie de la matrice du système linéaire global, et s'interprète de manière naturelle en terme de flux. Il consiste à remplacer, selon le côté de frontière concerné, la dérivée normale $\dfrac{\partial u}{\partial n}_{|\Gamma_n}$ par l'une des quatre expressions

$$\frac{u_{i+1,j}-u_{i,j}}{\Delta x}, \quad \frac{u_{i,j}-u_{i-1,j}}{\Delta x}, \quad \frac{u_{i,j+1}-u_{i,j}}{\Delta y}, \quad \frac{u_{i,j}-u_{i,j-1}}{\Delta y}$$

— Le choix centré du second ordre conservera globalement l'ordre 2 de l'approximation. Ce choix est cohérent avec la modélisation d'une condition de symétrie par une condition de Neumann. Il consiste à remplacer, selon le côté de frontière concerné, la dérivée normale $\dfrac{\partial u}{\partial n}_{|\Gamma_n}$ par l'une des quatre expressions

$$\frac{u_{i+1,j}-u_{i-1,j}}{2\Delta x}, \quad \frac{u_{i+1,j}-u_{i-1,j}}{2\Delta x}, \quad \frac{u_{i,j+1}-u_{i,j-1}}{2\Delta y}, \quad \frac{u_{i,j+1}-u_{i,j-1}}{2\Delta y}$$

Il est ensuite nécessaire de numéroter les $u_{i,j}$ pour qu'elles constituent les composantes U_I d'un vecteur inconnu U. La numérotation influe sur la structure de la matrice. On utilise des algorithmes de numérotation optimale afin de minimiser le stockage ("profil") de la matrice. Il ne reste alors plus qu'à résoudre le système linéaire obtenu par des méthodes directes de factorisation (méthodes de Gauss LU ou méthode de Choleski LL^T) ou par des méthodes itératives, voir chapitre 2, paragraphe 2.7.

Remarque 4.3.1 *Les formules proposées ci-dessus ne représentent qu'une petite partie des formules possibles. Ce sont les plus fréquemment utilisées. Il existe un grand nombre de formules à tous les ordres de précision.*

4.4 L'approche éléments finis en dimension un

La présentation très succincte faite ici n'est qu'une première introduction à la méthode des éléments finis. Elle a pour but de donner les idées de base dans un cas extrêmement simple.

Remarquons tout d'abord que la formulation différentielle (4.3) du problème suppose l'existence de dérivées secondes de la solution u. Or, physiquement, on peut trouver des exemples très simples de problèmes pour lesquels u' est discontinue et donc u'' n'a pas de sens classique (il faut alors considérer u'' au sens des distributions). Prenons, par exemple, une corde soumise à un chargement discontinu. La solution prendra la forme d'une ligne brisée.

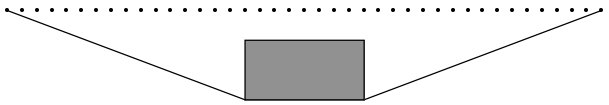


FIGURE 4.3 – Corde élastique sous chargement discontinu

En réalité, la bonne formulation mathématique du problème consiste à écrire que la déformée de la corde réalise le minimum de l'énergie à sa position d'équilibre. L'énergie de ce problème s'écrit

$$J(u) = \frac{1}{2}\int_a^b [u'(x)]^2 dx - \int_a^b f(x)u(x)dx$$

C'est à partir de cette formulation en minimisation d'énergie que se déduit la formulation variationnelle (ou principe des puissances virtuelles) qui est à la base de la méthode des éléments finis. On comprend que les éléments finis soient particulièrement adaptés aux problèmes d'équilibre.

On introduit tout d'abord un produit scalaire de 2 fonctions selon :

$$(v, w) = \int_a^b v(x)\, w(x)\, dx$$

et l'espace $L^2[a, b]$ des fonctions de carré sommable sur $[a, b]$, c'est à dire telles que l'intégrale suivante existe :

$$\int_a^b v^2(x)\, dx$$

L'espace $L^2[a, b]$, muni du produit scalaire ci-dessus est un espace de Hilbert.
Soit $H^1[a, b]$ l'espace de Hilbert des fonctions v, de carré sommable et dont la dérivée est également de carré sommable.

$$v \in L^2[a, b], \quad v' \in L^2[a, b]$$

et soit $H_0^1[a, b]$ l'espace de Hilbert des fonctions v de $H^1[a, b]$ nulles en a et b.

$$v \in L^2[a, b], \quad v' \in L^2[a, b], \quad v(a) = v(b) = 0$$

Il est facile de vérifier que la forme bilinéaire

$$a(u, v) = \int_a^b u'(x)\, v'(x)\, dx$$

est un produit scalaire sur $H_0^1[a,b]$. Voir l'annexe B pour la définition de ces espaces fonctionnels. Considérons donc le problème de minimisation de la forme quadratique représentant l'énergie du système :

$$\left\{\begin{array}{l} \text{Chercher la fonction } u \text{ vérifiant } u(a)=\alpha \quad u(b)=\beta \\ \text{qui réalise le minimum de la forme } J \text{ définie par} \\ J(v)=\dfrac{1}{2}\displaystyle\int_a^b v'^2 dx \; - \; \int_a^b f\, v\, dx \end{array}\right. \tag{4.8}$$

Ce problème de minimisation est équivalent au problème variationnel suivant

$$\left\{\begin{array}{l} \text{Chercher la fonction } u \text{ vérifiant } u(a)=\alpha \quad u(b)=\beta \text{ telle que :} \\ \displaystyle\int_a^b u'(x)\, v'(x)\, dx = \int_a^b f(x)\, v(x)\, dx \qquad \forall v \in H_0^1[a,b] \end{array}\right. \tag{4.9}$$

Pour s'en convaincre, il suffit de calculer $J(u+\lambda v)$ avec u solution du problème variationnel, λ réel quelconque et $v \in H_0^1[a,b]$ quelconque. Ainsi la fonction $u+\lambda v$ est une fonction quelconque de $H^1[a,b]$ vérifiant les conditions aux limites imposées en a et b. Remarquons que ce sont ces conditions aux limites de Dirichlet qui imposent le choix de fonctions v nulles en a et b.

$$J(u+\lambda v) = J(u) + \lambda[\int_a^b u'(x)v'(x)dx - \int_a^b f(x)v(x)dx] + \frac{\lambda^2}{2}\int_a^b [v'(x)]^2 dx$$

Si u minimise J, l'expression

$$\lambda[\int_a^b u'(x)v'(x)dx - \int_a^b f(x)v(x)dx] + \frac{\lambda^2}{2}\int_a^b [v'(x)]^2 dx$$

doit être positive pour tout λ réel et tout $v \in H_0^1[a,b]$ ce qui nécessite que

$$[\int_a^b u'(x)v'(x)dx - \int_a^b f(x)v(x)dx] = 0 \quad \forall v \in H_0^1[a,b].$$

Inversement si le crochet

$$[\int_a^b u'(x)v'(x)dx - \int_a^b f(x)v(x)dx]$$

est nul, on voit que $J(u) \leq J(u+\lambda v) \quad \forall \lambda$ et $\forall v$. On obtient donc , pour tout $\lambda \in I\!R$, et pour tout $v \in H_0^1[a,b]$, l'équivalence

$$J(u+\lambda v) \geq J(u) \Longleftrightarrow \int_a^b u'(x)\, v'(x)\, dx = \int_a^b f(x)\, v(x)\, dx$$

Remarque 4.4.1 *Observons que la formulation variationnelle revient à écrire qu'à l'équilibre, donc au minimum de l'énergie, les dérivées de J au point u dans toutes les directions v sont nulles. En effet on a, par définition de la dérivée directionnelle :*

$$J'(u;v) = \lim_{\lambda \to 0} \frac{J(u+\lambda v) - J(u)}{\lambda} = \int_a^b u'(x)\, v'(x)\, dx - \int_a^b f(x)\, v(x)\, dx$$

Une fois la formulation variationnelle obtenue, on peut retrouver l'équation différentielle initiale (4.3) de la façon suivante :

On intègre par parties le terme

$$\int_a^b u'(x)\, v'(x)\, dx = -\int_a^b u''(x)\, v(x)\, dx + u'(b)v(b) - u'(a)v(a)$$

En prenant en compte les conditions sur v ($v(a) = v(b) = 0$), on en déduit

$$-\int_a^b u''(x)v(x)dx - \int_a^b f(x)v(x)dx = 0 \quad \forall v \in H_0^1[a,b].$$

D'où l'égalité

$$-u''(x) = f(x)$$

au sens de $L^2(a,b)$.

On peut faire la démarche inverse. C'est celle que l'on rencontre dans la plupart des ouvrages. Le point de départ est alors le problème différentiel (4.3). On multiplie l'équation par $v(x)$ et on intègre sur [a,b] :

$$-\int_a^b u''(x)\, v(x)\, dx = \int_a^b f(x)\, v(x)\, dx$$

Par intégration par parties :

$$-\int_a^b u''(x)\, v(x)\, dx = \int_a^b u'(x)\, v'(x)\, dx + u'(a)\, v(a) - u'(b)\, v(b)$$

On obtient la formulation variationnelle, qui, en prenant en compte les conditions sur les fonctions v appartenant à l'espace $H_0^1[a,b]$ [1]

$$v(a) = v(b) = 0$$

1. C'est ce choix des fonctions tests v nulles pour les points où la solution est fixée qui n'est pas évident dans cette seconde approche, alors que sa nécessité est très naturelle dans l'approche minimisation

s'écrit :

$$\begin{cases} \text{Chercher la fonction } u \in H^1[a,b] \text{ vérifiant } u(a)=\alpha \quad u(b)=\beta \text{ telle que :} \\ \displaystyle\int_a^b u'(x)\,v'(x)\,dx = \int_a^b f(x)\,v(x)\,dx \qquad \forall v \in H_0^1[a,b] \end{cases}$$

Cette formulation est équivalente à la minimisation d'une forme quadratique représentant l'énergie du système qui s'écrit :

$$\begin{cases} \text{Chercher la fonction } u \text{ vérifiant } u(a)=\alpha \quad u(b)=\beta \\ \text{qui réalise le minimum de la forme } J \text{ définie par} \\ \displaystyle J(v) = \frac{1}{2}\int_a^b {v'}^2 dx \; - \; \int_a^b f\,v\,dx \end{cases}$$

On obtient ainsi 3 formes équivalentes du problème
- — une forme différentielle
- — une forme variationnelle (principe des travaux virtuels)
- — une forme minimisation de l'énergie

4.4.1 Un premier exemple simple : les éléments P1

Considérons le problème de Dirichlet homogène :

$$\begin{cases} -u''(x) = f(x) \qquad a < x < b \\ u(a) = 0 \quad u(b) = 0 \end{cases} \tag{4.10}$$

dont on écrira la formulation variationnelle dans l'espace $H_0^1]a,b[$. On approche l'espace $H_0^1[a,b]$ par l'espace $V_{0,h} \subset H_0^1[a,b]$ construit de la manière suivante.

On choisit une discrétisation de l'intervalle $[a,b]$ en N sous-intervalles ou éléments $K_i = [x_{i-1}, x_i]$. Les éléments K_i n'ont pas forcément même longueur. $V_{0,h}$ est alors

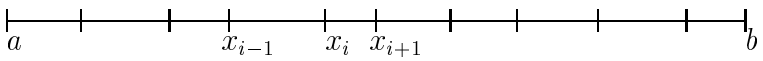


FIGURE 4.4 – Discrétisation (maillage) du segment [a,b] en éléments finis.

l'espace des fonctions continues affines par morceaux (c'est à dire affines sur les segments K_i) et nulles aux extrémités a et b.
L'utilisation de fonctions affines, fonctions polynomiales de degré un, justifie la dénomination d'éléments P1. Chaque fonction $v_h \in V_{0,h}$ est déterminée de manière unique par la donnée de ses valeurs aux points x_i pour $i = 1, ...N-1$. L'espace $V_{0,h}$ est de dimension $N-1$.

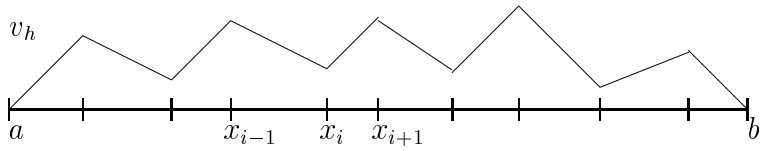


FIGURE 4.5 – Une fonction affine par morceaux.

4.4.2 Base de Lagrange

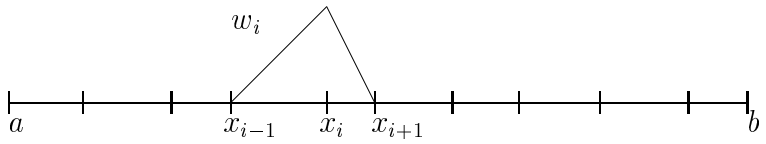


FIGURE 4.6 – Fonction de base de Lagrange

Considérons les N-1 fonctions $w_i \in V_{0,h}$ définies par les N-1 conditions suivantes :

$$w_i(x_j) = \delta_{ij} \quad \forall i = 1, ...N-1 \quad et \quad \forall j = 1, ...N-1 \tag{4.11}$$

Ces N-1 fonctions forment une base de $V_{0,h}$ et une fonction v_h quelconque s'écrit dans cette base :

$$v_h(x) = \sum_{i=1}^{i=N-1} v_i \, w_i(x)$$

avec $v_i = v_h(x_i)$. Les coefficients v_i sont donc les valeurs de v_h aux points (x_i)

4.4.3 Écriture du problème approché

Ecrivons le problème approché dans $V_{0,h}$

$$\int_a^b u_h'(x) \, v_h'(x) \, dx = \int_a^b f(x) \, v_h(x) \, dx \qquad \forall v_h \in V_{0,h} \tag{4.12}$$

Le problème étant linéaire, l'égalité est vraie pour tout v_h si elle est vraie pour une base de l'espace vectoriel $V_{0,h}$.
On peut donc remplacer dans (4.12) $\forall v_h \in V_{0,h}$ par $\forall w_i \quad$ pour $\quad i = 1, ...N-1$.
D'autre part, écrivons u_h, solution du problème approché dans $V_{0,h}$, dans la base des w_i

$$u_h(x) = \sum_{j=1}^{j=N-1} u_j \, w_j(x)$$

avec $u_j = u_h(x_j)$ valeur approchée de la solution exacte au point (x_j)
On obtient l'écriture suivante du problème approché :

Trouver $u_1,\ u_2, ...u_{N-1}$ tels que

$$\int_a^b \Big(\sum_{j=1}^{j=N-1} u_j\, w_j'(x) \Big)\, w_i'(x)\, dx = \int_a^b f(x)\, w_i(x)\, dx \qquad \forall i = 1, ...N-1$$

Ce que l'on peut récrire

$$\sum_{j=1}^{j=N-1} \Big(\int_a^b w_j'(x)\, w_i'(x)\, dx \Big)\, u_j = \int_a^b f(x)\, w_i(x)\, dx \qquad \forall i = 1, ...N-1$$

Soit en posant

$$\int_a^b f(x)\, w_i(x)\, dx = F_i$$

et

$$\int_a^b w_j'(x)\, w_i'(x)\, dx = A_{ij}$$

$$\sum_{j=1}^{j=N-1} A_{ij}\ u_j = F_i \qquad \forall i = 1, ...N-1$$

On a ainsi obtenu un système linéaire de $N-1$ équations à $N-1$ inconnues, qui peut s'écrire sous la forme matricielle suivante :

$$A\ U \ = \ F$$

Chaque ligne d'indice i du système linéaire correspond au choix d'une fonction de base d'indice i de l'espace des fonctions tests $V_{0,h}$. Il y a autant d'équations que d'inconnues puisque la partie inconnue de la solution approchée s'exprime elle aussi dans la base des w_i.

Les calculs détaillés des coefficients des matrices et seconds membres seront présentés chapitre 7.

4.5 L'approche volumes finis

La méthode des volumes finis, comme celle des éléments finis utilise une formulation intégrale des équations. Mais au lieu d'utiliser un produit scalaire de L^2 par des fonctions tests, on se contente d'intégrer les équations différentielles sur des volumes élémentaires ou volumes de contrôle. Ceci peut s'interpréter comme l'utilisation de fonctions tests, fonctions indicatrices des volumes élémentaires.

4.5.1 En dimension un

À partir d'un maillage en volumes finis, où l'on prend les inconnues au centre x_i des volumes de contrôles qui sont dans ce cas monodimensionnel, les segments $[x_{i-1/2}, x_{i+1/2}]$, on passe ainsi de

$$-u''(x) = f(x) \quad \Longleftrightarrow \quad -(u'(x))' = f(x)$$

à

$$-\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (u'(x))' dx = u'(x_{i-\frac{1}{2}}) - u'(x_{i+\frac{1}{2}}) = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x) dx$$

Les $u'(x_{i-\frac{1}{2}})$ et $u'(x_{i+\frac{1}{2}})$ apparaissent comme des flux aux interfaces des volumes élémentaires, les segments $[x_{i-1/2}, x_{i+1/2}]$.

Remarque 4.5.1 *Ceci explique, qu'à l'inverse des différences finies, ce sont, dans le cas des volumes finis, les conditions de Neumann qui sont faciles à prendre en compte et les conditions de Dirichlet qui sont plus délicates.*

Si on approche les dérivées selon :

$$u'(x_{i+\frac{1}{2}}) = \frac{u(x_i + h) - u(x_i)}{h} \quad \text{et} \quad u'(x_{i-\frac{1}{2}}) = \frac{u(x_i) - u(x_{i-h})}{h}$$

dans le cas particulier de points x_i régulièrement espacés d'un pas h uniforme, on retrouve la discrétisation centrée classique de la dérivée seconde

$$u''(x_i) = \frac{d^2u}{dx^2}(x_i) \simeq \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}$$

L'intérêt de cette technique de volumes finis n'apparaît que dans les applications aux problèmes en dimension 2 et 3. De nombreuses variantes existent, selon la géométrie des volumes élémentaires, le choix du positionnement des inconnues aux centres ou aux sommets des volumes de contrôle, et les diverses formules de calcul des flux aux interfaces.

4.5.2 En dimension deux

La méthode des volumes finis, conservative par essence, est particulièrement bien adaptée à la résolution numérique de problèmes conservatifs. On l'utilise donc avec succès pour modéliser des phénomènes de transport (ou convection, voir chapitre 12). Il existe un grand nombre de techniques de type volumes finis qui se distinguent

— par la situation des inconnues : au noeuds du maillage ou au centre des mailles

— par la forme des volumes de contrôle : carrés, quadrilatères quelconques, polygones quelconques réunions de triangles etc
— par le choix des formules de calcul approché des flux et en particulier leur ordre de précision.

Nous choisissons ici de nous limiter au choix originel historiquement de mailles carrées ou quadrangulaires avec les inconnues représentées au centre des mailles. Considérons l'équation de convection

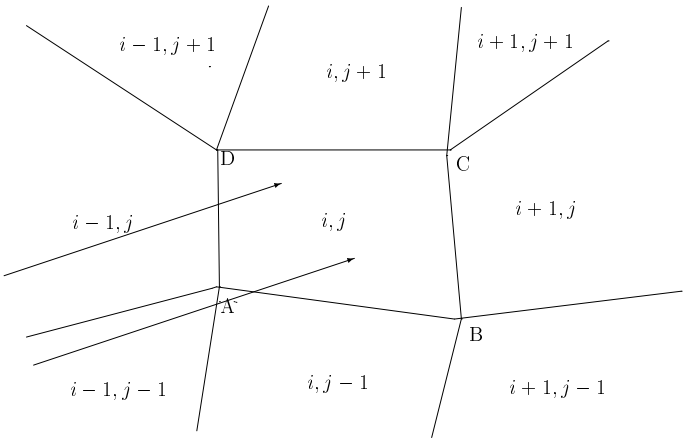


FIGURE 4.7 – Grille volumes finis.

$$\frac{\partial u}{\partial t} + div(F(u)) = Q \tag{4.13}$$

En intégrant l'équation dans le volume de contrôle $ABCD$ on obtient

$$mes(ABCD)\,\frac{du_{ij}}{dt} + \int_{ABCD} F(u).n\,ds = mes(ABCD)\,Q_{ij}$$

L'intégrale $\int_{ABCD} F(u).n\,ds$ représente la somme des flux sortants de F à travers les faces AB, BC, CD et DA.

Les flux doivent être calculés par des formules indépendantes de la cellule, afin d'assurer la conservation globale. En particulier, le flux sortant du volume élémentaire $\Omega_{i,j}$ à travers la face AB doit être équilibré par celui sortant de $\Omega_{i,j-1}$ à travers la même face AB.

Schémas centrés

Toujours en se plaçant dans le cas d'inconnues au centre des mailles on a les choix suivants

1.
$$F_{AB} = \frac{1}{2}[F_{i,j} + F_{i,j-1}] \quad \text{avec} \quad F_{i,j} = F(u_{i,j})$$

2.
$$F_{AB} = F(\frac{u_{i,j} + u_{i,j-1}}{2})$$

3.
$$F_{AB} = \tfrac{1}{2}[F_A + F_B]$$
$$\text{avec} \quad F_A = \tfrac{1}{4}[F(u_{i,j}) + F(u_{i-1,j}) + F(u_{i,j-1}) + F(u_{i-1,j-1})]$$
$$\text{et} \quad F_B = \tfrac{1}{4}[F(u_{i,j}) + F(u_{i+1,j}) + F(u_{i,j-1}) + F(u_{i+1,j-1})]$$

4.
$$F_{AB} = \tfrac{1}{2}[F_A + F_B]$$
$$\text{avec} \quad F_A = F(\tfrac{u_{i,j}+u_{i-1,j}+u_{i,j-1}+u_{i-1,j-1}}{4})$$
$$\text{et} \quad F_B = F(\tfrac{u_{i,j}+u_{i+1,j}+u_{i,j-1}+u_{i+1,j-1}}{4})$$

Schémas décentrés

Prenons le cas d'un flux dirigé comme sur la figure ci-dessus. Dans le cas de schémas décentrés, on calcule le flux à travers AB en prenant la valeur $F(u_{i,j-1})$ et ainsi de suite pour les flux à travers les faces BC , CD, et DA. On prend donc

$$F_{AB} = F(u_{i,j-1}) \quad F_{BC} = F(u_{i,j}) \quad F_{CD} = F(u_{i,j}) \quad F_{DA} = F(u_{i-1,j})$$

Ce choix conduit à des approximations du premier ordre. D'autres formules plus complexes permettent d'obtenir des flux décentrés au second ordre. Voir l'exposé des techniques de décentrage au chapitre 12.

Chapitre 5

Introduction aux méthodes variationnelles

La méthode variationnelle, dont voici une introduction, place la recherche de solutions des équations elliptiques dans le cadre des espaces de Hilbert. Ces solutions apparaissent alors souvent comme les fonctions qui minimisent, sur un sous-espace ou un convexe donné, une forme quadratique représentant l'énergie. La méthode variationnelle présente de nombreux avantages

- d'un point de vue théorique, elle permet d'obtenir l'existence d'une solution faible.
- du point de vue du modèle, cette solution faible représente mieux la réalité physique.
- enfin, elle place le problème de la recherche des solutions approchées, dans le cadre général très commode de la meilleure approximation au sens d'une norme associée à un produit scalaire.

5.1 Un problème elliptique modèle en dimension un

Reprenons un problème de Dirichlet homogène voisin du problème présenté dans le chapitre précédent.

$$\begin{cases} -u''(x) + \alpha(x)\, u(x) = f(x) & \forall x \in [a,b] \\ u(a) = u(b) = 0 \end{cases} \tag{5.1}$$

où f et α sont deux fonctions données sur $[a,b]$. On suppose α continue et positive sur $[a,b]$: $\alpha(x) \geq 0 \quad \forall x \in [a,b]$.

Si la fonction au second membre f est continue sur $[a,b]$, la solution u de ce problème est une fonction continue et deux fois continûment dérivable sur $[a,b]$,

l'égalité $-u''(x)+\alpha(x)\,u(x) = f(x)$ est vérifiée en tout point $x \in [a,b]$ au sens fort. On dit alors que u est solution classique ou solution forte du problème différentiel.

On a vu, lors de la courte introduction aux éléments finis du chapitre précédent, que, pour des raisons d'approximation numérique, mais également pour des raisons de modélisation physique, ce problème doit être reformulé en termes de minimisation d'énergie. On va choisir de rechercher la solution u, non plus comme solution classique de 5.1, mais comme fonction minimisant l'énergie, ici :

$$J(v) = \frac{1}{2}\int_a^b [v'(x)^2 + \alpha(x)v(x)^2]dx - \int_a^b f(x)v(x)dx$$

sur l'espace $H_0^1[a,b]$ des fonctions de carré sommable à dérivées de carré sommable (voir annexeB) et nulles en a et b (pour ce problème de Dirichlet homogène).

- *En effet le second membre f n'est pas toujours une fonction continue dans la pratique et donc la solution u n'est pas toujours de classe C^2. Par exemple dans le cas d'une corde élastique, on peut avoir un chargement concentré sur une partie, voire sur un seul point de $[a,b]$. De même, on peut concevoir une barre chauffée par une résistance sur une partie limitée ou en un point par un laser.*

On obtient ainsi le problème

$$\left\{\begin{array}{l} \text{Chercher la fonction } u \in H_0^1[a,b] \text{ telle que, } \forall v \in H_0^1[a,b] \\ \displaystyle\int_a^b u'(x)v'(x)dx \,+\, \int_a^b \alpha(x)u(x)v(x)dx \,=\, \int_a^b f(x)v(x)dx \end{array}\right. \tag{5.2}$$

Cette nouvelle formulation, dite formulation variationnelle ou formulation faible peut aussi être obtenue après multiplication par $v(x)$ et intégration sur $[a,b]$ de 5.1 :

$$-\int_a^b u''(x)v(x)\,dx \,+\, \int_a^b \alpha(x)u(x)v(x)\,dx = \int_a^b f(x)v(x)\,dx$$

puis par intégration par parties :

$$-\int_a^b u''(x)\,v(x)\,dx = \int_a^b u'(x)\,v'(x)\,dx + u'(a)\,v(a) - u'(b)\,v(b)$$

On peut interpréter cette démarche de la façon suivante : on a remplacé l'égalité $-u''+\alpha u = f$ écrite initialement au sens fort d'une égalité de fonctions continues, par une égalité au sens faible de fonctions de carré sommable qui s'exprime précisément par produit scalaire dans $L^2[a,b]$. On doit alors considérer la dérivée u'' au sens des distributions, et c'est ce que l'on fait en intégrant par parties.

Équivalence des formulations dans le cas de solutions régulières

Théorème 5.1.1 *Toute solution classique est évidemment solution faible puisqu'elle vérifie (5.2) après intégration par parties.*

Inversement si f est continue, u est une solution régulière (de classe $C^2[a,b]$) et u est une solution classique.

Nous renvoyons à la bibliographie pour l'analyse fonctionnelle de ces problèmes. Nous allons maintenant montrer l'existence et l'unicité dans $H_0^1[a,b]$ d'une solution faible du problème variationnel (5.2). Pour cela nous allons utiliser le théorème fondamental de **Lax -Milgram** présenté dans un cadre général abstrait.

En effet le problème variationnel peut s'écrire sous la forme générale :

$$\left\{ \begin{array}{c} \text{Chercher } u \text{ appartenant à l'espace de Hilbert } V \text{ telle que :} \\ a(u,v) = l(v) \quad \forall v \in V \end{array} \right. \tag{5.3}$$

en posant

$$V = H_0^1[a,b]$$
$$a(u,v) = \int_a^b u'(x)v'(x)dx \; + \; \int_a^b \alpha(x)u(x)v(x)dx$$
$$l(v) = \int_a^b f(x)v(x)dx$$

5.2 Théorème de Lax-Milgram

Définition 5.2.1 *La forme bilinéaire a est continue dans V s'il existe une constante positive M telle que*

$$|a(u,v)| \leq M\|u\| \, \|v\| \quad \forall\, u, v \in V$$

Définition 5.2.2 *La forme bilinéaire a est dite* **elliptique** *(on trouve également coercive) dans V s'il existe une constante* **strictement** *positive m telle que*

$$a(v,v) \geq m \, \|v\|^2 \quad \forall\, v \in V$$

Remarque 5.2.1 *Lorsqu'elle est elliptique et symétrique, la forme bilinéaire* $a : u, v \longrightarrow a(u,v)$ *est un produit scalaire sur* V*, et sa norme associée* $v \longrightarrow a(v,v)^{\frac{1}{2}}$ *est une norme équivalente à la norme initialement choisie sur* V*. On a en effet :*

$$\sqrt{m}\ \|v\| \leq a(v,v)^{\frac{1}{2}} \leq \sqrt{M}\ \|v\|$$

On a alors le résultat suivant :

Théorème 5.2.1 (Lax-Milgram) *Soit* $a(u,v)$ *une forme bilinéaire continue elliptique sur l'espace de Hilbert* V*, soit* $l(v)$ *une forme linéaire continue sur* V*, le problème variationnel*

$$\begin{cases} \text{Chercher la fonction } u \text{ appartenant à l'espace de Hilbert } V \text{ telle que :} \\ a(u,v) = l(v) \quad \forall v \in V \end{cases}$$

admet une solution unique dans V*.*

De plus, si a *est symétrique, ce problème est équivalent au problème de minimisation :*

$$\begin{cases} \text{Chercher la fonction } u \text{ qui réalise le minimum dans } V \text{ de :} \\ J(v) = \frac{1}{2}a(v,v) - l(v) \end{cases}$$

Démonstration. L'application $v \longrightarrow a(u,v)$ est une forme linéaire continue sur V. Donc, en appliquant le théorème de Riesz, il existe un élément $Au \in V$ tel que

$$a(u,v) = (Au, v)_V \quad \forall v \in V$$

De même $l(v)$ est une forme linéaire continue sur V, donc il existe $L \in V$ tel que

$$l(v) = (L, v)_V \quad \forall v \in V$$

Considérons l'application de V dans V définie par $v \longrightarrow v - \rho\,(Av - L)$ où ρ est un réel strictement positif. Montrons qu'avec les hypothèses faites sur a et l, cette application est une contraction dans V pour ρ assez petit.

En effet

$$\|v - \rho Av\|^2 = \|v\|^2 - 2\rho(Av, v) + \rho^2\|Av\|^2 \leq (1 - 2\rho m + \rho^2 M^2)\,\|v\|^2$$

où m est la constante d'ellipticité et M la constante de continuité de a.

On aura contraction dès que $(1 - 2\rho m + \rho^2 M^2) < 1$ donc pour $\rho \in]0, \dfrac{2m}{M^2}[$. En choisissant donc ρ dans cet intervalle on a construit une application contractante dans V qui admet un point fixe unique $u \in V$ tel que

$$u = u - \rho(Au - L) \quad \text{donc tel que } Au = L$$

u est la solution du problème variationnel.

Considérons maintenant le cas a symétrique. Calculons, pour tout réel λ et tout $v \in V$ l'expression $J(u + \lambda v)$

$$J(u + \lambda v) = J(u) + \lambda\big[a(u, v) - l(v)\big] + \frac{\lambda^2}{2}a(v, v)$$

Donc, si u est solution du problème variationnel

$$a(u, v) - l(v) = 0 \quad \forall v \in V$$

u minimise J.

Inversement si u minimise J, alors

$$\lambda\big[a(u, v) - l(v)\big] + \frac{\lambda^2}{2}a(v, v) \geq 0 \quad \forall \lambda \text{ et } v$$

Donc le trinôme en λ ci-dessus étant toujours positif, on doit avoir

$$a(u, v) - l(v) = 0 \quad \forall v \in V$$

5.3 Problèmes de Dirichlet en dimension un

Appliquons ce théorème au problème (5.1).

5.3.1 Problème de Dirichlet homogène

On appelle problème de Dirichlet homogène, dans le cas d'un problème différentiel du second ordre, en dimension un, sur un intervalle $[a, b]$, un problème dans lequel les conditions aux limites aux extrémités a et b de l'intervalle sont

$$u(a) = 0 \quad u(b) = 0$$

On parle aussi de conditions aux limites fixées ou de conditions aux limites essentielles.

La formulation variationnelle du problème (5.1) s'écrit :

$$\left\{ \begin{array}{l} \text{Chercher la fonction } u \in H_0^1[a, b] \text{ telle que : } \quad \forall v \in H_0^1[a, b] \\ \displaystyle\int_a^b u'(x)v'(x)dx + \int_a^b \alpha(x)u(x)v(x)dx = \int_a^b f(x)v(x)dx \end{array} \right. \tag{5.4}$$

Vérifions, pour ce problème, les hypothèses du théorème de Lax-Milgram.

1) L'espace $H_0^1[a, b]$ est un espace de Hilbert.

2) La forme bilinéaire symétrique (évident) a est continue. En effet soit M le max de α sur $[a,b]$, on a

$$\Big| \int_a^b u'(x)v'(x)dx \, + \, \int_a^b \alpha(x)u(x)v(x)dx \Big| \leq \, max(1,M) \, \|u\| \, \|v\|$$

3) La forme a est elliptique : car ou bien le minimum m de α sur $[a,b]$ est strictement positif et on a

$$\int_a^b v'^2(x)dx \, + \, \int_a^b \alpha(x)v^2(x)dx \, \geq \, min(1,m) \, \|v\|^2$$

et l'ellipticité avec une constante égale à : $min(1,m)$,

ou bien $m = 0$ et on utilise l'inégalité de Poincaré vérifiée dans $H_0^1[a,b]$ et qui assure

$$\|v'\|_{0,2}^2 \geq (1 + \frac{(b-a)^2}{2})^{-1} \, \|v\|_{1,2}^2$$

donc encore l'ellipticité avec une constante égale à : $(1 + \frac{(b-a)^2}{2})^{-1}$

4) La forme l est continue car :

$$\Big| \int_a^b f(x)v(x)dx \Big| \leq \, \|f\|_{0,2} \, \|v\|_{0,2} \, \leq \, \|f\|_{0,2} \, \|v\|_{1,2}$$

En conséquence, les hypothèses du théorème de Lax-Milgram étant vérifiées, le problème (5.1) admet une solution unique dans $H_0^1[a,b]$, qui réalise également le minimum de l'énergie :

$$J(v) = \frac{1}{2} \Big(\int_a^b v'^2(x)dx \, + \, \int_a^b \alpha(x)v^2(x)dx \Big) \, - \int_a^b f(x)v(x)dx \qquad (5.5)$$

5.3.2 Problème de Dirichlet non-homogène

Dans ce cas, les conditions aux limites sont fixées mais non nulles en a et b. soit par exemple

$$u(a) = u_a \quad u(b) = u_b$$

Une solution consiste à se ramener au problème précédent en choisissant une fonction auxiliaire simple u_0 prenant les valeurs fixées

$$u_0(a) = u_a \quad u_0(b) = u_b$$

et à poser $u = \tilde{u} + u_0$. Le problème se ramène alors à un problème de Dirichlet homogène pour la nouvelle inconnue $\tilde{u}$, soit

$$\begin{cases} -\tilde{u}''(x) + \alpha(x)\,\tilde{u}(x) = f(x) + u_0''(x) - \alpha(x)\,u_0(x) & \forall x \in [a,b] \\ \tilde{u}(a) = \tilde{u}(b) = 0 \end{cases}$$

dont la formulation variationnelle s'écrit

$$\begin{cases} \text{Chercher la fonction } \tilde{u} \text{ appartenant à } H_0^1[a,b] \text{ telle que :} \\ \displaystyle\int_a^b \tilde{u}'(x)v'(x)dx + \int_a^b \alpha(x)\tilde{u}(x)v(x)dx = \int_a^b f(x)v(x)dx \\ \displaystyle -\int_a^b u_0'(x)v'(x)dx - \int_a^b \alpha(x)u_0(x)v(x)dx \quad \forall v \in H_0^1[a,b] \end{cases} \tag{5.6}$$

On verra plus loin, lors de l'approximation par éléments finis de ce problème, quel est le bon choix pratique de la fonction u_0.

Remarque 5.3.1 (importante) *L'espace V , espace des fonctions tests de la formulation variationnelle est toujours le même, $H_0^1[a,b]$, dans le cas Dirichlet non homogène comme dans le cas Dirichlet homogène. Ce qui implique le choix de fonctions tests v nulles au bord, c'est le fait que les valeurs de l'inconnue u soient fixées sur la frontière et non qu'elles y soient nulles. C'est ceci que l'approche minimisation d'énergie rend plus évident. On peut alors considérer la solution comme la fonction qui réalise le minimum de l'énergie sur le convexe des fonctions telles que $u(a) = u_a$, $u(b) = u_b$. Les fonctions tests v s'interprètent comme des directions d'accroissement au voisinage de u. Pour que la fonction voisine $u + \lambda v$ reste dans le convexe $\forall \lambda$, il faut imposer $v(a) = v(b) = 0$.*

On démontre sans difficulté, dans le cas Dirichlet non-homogène, l'existence et l'unicité de la solution, par la vérification des hypothèses du théorème de Lax-Milgram.

5.4 Problèmes de Neumann en dimension un

On appelle problème de Neumann, dans le cas d'un problème différentiel du second ordre en dimension un sur un intervalle $[a,b]$, un problème dans lequel les conditions aux limites aux extrémités a et b de l'intervalle sont

$$u'(a) = \mu \quad u'(b) = \nu$$

On parle aussi de conditions aux limites libres ou de conditions aux limites naturelles dans le cas homogène ($\mu = \nu = 0$).

Ces conditions modélisent la donnée d'une force ou d'un flux sur la frontière Γ.

Le problème

$$\begin{cases} -u''(x) + \alpha(x)\, u(x) = f(x) & \forall x \in [a,b] \\ u'(a) = \mu \quad u'(b) = \nu \end{cases} \tag{5.7}$$

est donc un problème de Neumann.

Formulation variationnelle

Les valeurs de u au bord de l'intervalle ne sont plus fixées, ce sont maintenant des inconnues du problème. En conséquence la formulation variationnelle s'écrira dans l'espace de fonctions tests $H^1[a,b]$ tout entier.

$$\begin{cases} \text{Chercher la fonction } u \text{ appartenant à } H^1[a,b] \text{ telle que :} \\ \displaystyle\int_a^b u'(x)v'(x)dx \,+\, \int_a^b \alpha(x)u(x)v(x)dx \,=\, \int_a^b f(x)v(x)dx \\ +\nu v(b) - \mu v(a) \quad \forall v \in H^1[a,b] \end{cases} \tag{5.8}$$

Vérifions, pour ce problème, les hypothèses du théorème de Lax-Milgram.

1) L'espace $H^1[a,b]$ est un espace de Hilbert.

2) La forme bilinéaire symétrique (évident) a est continue. En effet soit M le max de α sur $[a,b]$, on a

$$\Big|\, \int_a^b u'(x)v'(x)dx \,+\, \int_a^b \alpha(x)u(x)v(x)dx \,\Big| \,\leq\, max(1,M)\, \|u\|\, \|v\|$$

3) La forme a est elliptique si le minimum m de α sur $[a,b]$ est strictement positif et on a

$$\int_a^b v'^2(x)dx \,+\, \int_a^b \alpha(x)v^2(x)dx \,\geq\, min(1,m)\, \|v\|^2$$

et l'ellipticité avec une constante égale à : $min(1,m)$.

Remarque 5.4.1 *Dans le cas $\alpha = 0$ on n'a plus ellipticité de a, le problème est mal posé. Il est d'ailleurs clair qu'alors, la fonction inconnue u n'apparaissant dans le problème que par l'intermédiaire des valeurs de sa dérivée, on a une infinité de solutions à une constante additive près. Il y a, de plus, une condition de compatibilité entre le second membre et les conditions aux limites de Neumann à vérifier pour avoir une solution. En effet, on a*

$$-u''(x) = f(x) \Longrightarrow u'(a) - u'(b) = \int_a^b f(x)dx$$

4) La forme l est continue car :

$$\Big| \int_a^b f(x)v(x)dx \Big| \leq \|f\|_{0,2} \|v\|_{0,2} \leq \|f\|_{0,2} \|v\|_{1,2}$$

et gràce à l'inclusion $H^1[a,b] \subset C[a,b]$

$$|\nu v(b) - \mu v(a)| \leq c \|v\|_{1,2}$$

En conséquence, les hypothèses du théorème de Lax-Milgram étant vérifiées, le problème (5.7) admet une solution unique dans $H^1[a,b]$, qui réalise également le minimum de l'énergie :

$$J(v) = \frac{1}{2}\Big(\int_a^b v'^2(x)dx + \int_a^b \alpha(x)v^2(x)dx \Big) - \int_a^b f(x)v(x)dx - \nu v(b) + \mu v(a)$$

dans $H^1[a,b]$

Retour à la formulation différentielle initiale

En intégrant par parties

$$\int_a^b u'(x)v'(x)dx = -\int_a^b u''(x)v(x)dx + u'(b)v(b) - u'(a)v(a)$$

donc :

$$\int_a^b (-u''(x) + \alpha(x)u(x) - f(x))v(x)dx + (u'(b) - \nu)v(b) - (u'(a) - \mu)v(a) = 0$$

pour tout $v \in H^1[a,b]$. On choisit d'abord $v \in H_0^1[a,b]$ ce qui donne $-u'' + \alpha u = f$ au sens de $L^2[a,b]$. Puis on obtient

$$(u'(b) - \nu)v(b) - (u'(a) - \mu)v(a) = 0 \quad \forall v \in H^1[a,b]$$

donc

$$u'(a) = \mu \quad \text{et} \quad u'(b) = \nu$$

5.5 Problèmes mélés monodimensionnels

Les conditions aux limites mélés sont des conditions de Dirichlet sur une partie de la frontière, et de Neumann sur une autre partie. En dimension un, cela correspond à une condition de Dirichlet en un point et à une condition de Neumann sur l'autre extrémité du segment domaine de calcul.

$$u(a) = u_a = 0 \qquad u'(b) = \beta$$

Nous obtenons le problème :

$$\begin{cases} -u''(x) + \alpha(x)\, u(x) = f(x) & \forall x \in [a,b] \\ u(a) = u_a = 0 \qquad u'(b) = \beta \end{cases} \tag{5.9}$$

Formulation variationnelle

La valeur de u au point a est fixée, tandis que sa valeur au point b est une inconnue du problème. On recherchera donc la solution du problème variationnel dans un espace intermédiaire entre $H^1[a,b]$ et $H_0^1[a,b]$, l'espace V des fonctions de $H^1[a,b]$ nulles au point a. V est bien un espace de Hilbert, car c'est un sous-espace fermé de $H^1[a,b]$.

Comme précédemment, pour passer de la formulation différentielle à la formulation variationnelle, on multiplie l'équation

$$-u''(x) + \alpha(x)\, u(x) = f(x)$$

par $v(x)$ et on intègre sur $[a,b]$

$$-\int_a^b u''(x)v(x)dx \; + \int_a^b \alpha(x)u(x)v(x)dx \; = \int_a^b f(x)v(x)dx \quad \forall v \in V$$

puis en intégrant par parties la première intégrale on obtient le problème variationnel

$$\begin{cases} \text{Chercher la fonction } u \text{ appartenant à } V \text{ telle que :} \\ \displaystyle\int_a^b u'(x)v'(x)dx \; + \int_a^b \alpha(x)u(x)v(x)dx \; = \int_a^b f(x)v(x)dx + \beta v(b) \quad \forall v \in V \end{cases}$$

La démonstration de l'existence et de l'unicité se fait comme dans les cas précédents. L'ellipticité de a est immédiate si α est strictement positif, sinon on l'obtient par une inégalité de Poincaré grâce à la condition de Dirichlet au point a. Le cas d'une condition de Dirichlet non-homogène en a se traite comme ci-dessus, sans difficulté.

5.6 Problèmes de Fourier en dimension un

Les conditions aux limites de Fourier, fréquentes en thermique, expriment une relation affine entre les valeurs de la solution u et de sa dérivée en un point frontière. Elles s'écrivent de manière générale selon :

$$u'(a) = k_a\,(u(a) - u_a) + \beta_a$$

$$u'(b) = -k_b\,(u(b) - u_b) + \beta_b$$

avec ici k_a et k_b coefficients réels positifs.

Nous obtenons le problème :

$$\begin{cases} -u''(x) + \alpha(x)\,u(x) = f(x) & \forall x \in [a,b] \\ u'(a) = k_a\,(u(a) - u_a) + \beta_a \\ u'(b) = -k_b\,(u(b) - u_b) + \beta_b \end{cases} \tag{5.10}$$

Formulation variationnelle

Les valeurs de u aux points a et b sont des inconnues du problème. On recherchera donc la solution du problème variationnel dans l'espace $H^1[a,b]$.

Comme précédemment, pour passer de la formulation différentielle à la formulation variationnelle, on multiplie l'équation

$$-u''(x) + \alpha(x)\,u(x) = f(x)$$

par $v(x)$ et on intègre sur $[a,b]$

$$-\int_a^b u''(x)v(x)dx + \int_a^b \alpha(x)u(x)v(x)dx = \int_a^b f(x)v(x)dx \quad \forall v \in H^1[a,b]$$

puis en intégrant par parties la première intégrale on obtient le problème variationnel

$$\begin{cases} \text{Chercher la fonction } u \text{ appartenant à } H^1[a,b] \text{ telle que : } \forall v \in H^1[a,b] \\ \displaystyle\int_a^b u'(x)v'(x)dx + \int_a^b \alpha(x)u(x)v(x)dx + k_a u(a)v(a) + k_b u(b)v(b) \\ \qquad = \displaystyle\int_a^b f(x)v(x)dx + k_a u_a v(a) + k_b u_b v(b) - \beta_a v(a) + \beta_b v(b) \end{cases}$$

La démonstration de l'existence et de l'unicité se fait comme dans les cas précédents. Il convient de remarquer, que dans le cas de conditions aux limites de Fourier, la forme bilinéaire a est modifiée :

$$a(u,v) = \int_a^b u'(x)v'(x)dx \; + \; \int_a^b \alpha(x)u(x)v(x)dx \; + ku(a)v(a) + k_b u(b)v(b)$$

La continuité de a utilise les majorations

$$|u(a)| \leq \|u\|_{0,\infty} \leq \; C \, \|u\|_{1,2}$$

$$|u(b)| \leq \|u\|_{0,\infty} \leq \; C \, \|u\|_{1,2}$$

L'ellipticité de a est immédiate si α ou k_a et k_b sont strictement positifs.

Remarque 5.6.1 *Les conditions de type Fourier fournissent une forme générale comprenant comme cas particuliers :*

1) les conditions de Neumann dans le cas $k = 0$,

2) les conditions de Dirichlet $u(a) = u_a$, $u(b) = u_b$, à la limite quand k devient infiniment grand.

Enfin, signalons l'existence d'autres types de conditions aux limites importantes, les conditions aux limites périodiques qui s'écrivent :

$$u(a) = u(b) \quad u'(a) = u'(b) \tag{5.11}$$

5.7 Problèmes elliptiques en dimensions supérieures

Nous présenterons le cas de la dimension deux d'espace. Mais tout ce qui suit se généralise sans difficulté conceptuelle à la dimension trois. Par contre, il est clair que la discrétisation et l'implémentation pratique sont beaucoup plus complexes pour les problèmes tridimensionnels.

On introduit, comme en dimension un, un produit scalaire de 2 fonctions selon :

$$(v,w) = \iint_\Omega v(x,y)\, w(x,y)\, dxdy$$

et l'espace $L^2[\Omega]$ des fonctions de carré sommable sur Ω, c'est à dire telles que l'intégrale suivante existe :

$$\iint_\Omega v^2(x,y)\, dxdy$$

Soit $H^1[\Omega]$ l'espace des fonctions v, de carré sommable et dont les dérivées partielles sont également de carré sommable, et soit $H_0^1[\Omega]$ l'espace des fonctions v de $H^1[\Omega]$ "nulles" sur la frontière Γ de Ω (Les fonctions de $H^1[\Omega]$ ne sont pas continues pour Ω de dimension deux et trois, les valeurs nulles sont définies au sens des traces, voir annexe B et bibliographie)
On utilisera la formule de Green (équivalente à l'intégration par parties en dimension un) pour $u \in H^2(\Omega)$ et $v \in H^1(\Omega)$ (voir annexe B) :

$$-\iint_\Omega \Delta u \, v dxdy = \iint_\Omega \mathbf{grad} u \; \mathbf{grad} v dxdy - \int_\Gamma \frac{\partial u}{\partial n} \, v \, d\gamma$$

5.7.1 Problème de Dirichlet homogène

Soit Ω un domaine ouvert de $I\!R^2$ de frontière Γ , on considère le problème

$$\begin{cases} -\Delta \; u\,(x,y) = f\,(x,y) & \forall \; x,y \in \Omega \\ u_{\,|\,\Gamma} = 0 \end{cases} \tag{5.12}$$

Formulation variationnelle

On cherche une solution u fixée à zéro sur la frontière. On écrit donc la formulation variationnelle du problème dans l'espace $H_0^1(\Omega)$. Après multiplication par $v(x,y)$ et intégration on obtient :

$$-\iint_\Omega \Delta u \; v dxdy = \iint_\Omega f \, v \, dxdy \quad \forall \, v \in H_0^1(\Omega)$$

On utilise alors la formule de Green et on obtient :

$$\iint_\Omega \mathbf{grad} u \; \mathbf{grad} v dxdy - \int_\Gamma \frac{\partial u}{\partial n} \, v \, d\gamma = \iint_\Omega f \, v \, dxdy \quad \forall \, v \in H_0^1(\Omega)$$

Le terme d'intégrale de bord sur Γ disparaît puisqu'on a pris des fonctions tests v nulles au bord, et on obtient enfin le problème variationnel :

$$\begin{cases} \text{Trouver } \; u \in H_0^1(\Omega) \; \text{ telle que :} \\ \displaystyle\iint_\Omega \mathbf{grad} u \; \mathbf{grad} v \, dxdy = \iint_\Omega f \, v \, dxdy \quad \forall \, v \in H_0^1(\Omega) \end{cases} \tag{5.13}$$

Existence et unicité du problème

Ce problème s'écrit sous la forme générale :

$$\begin{cases} \text{Trouver } u \in V \text{ telle que :} \\ a(u,v) = l(v) \quad \forall\, v \in V \end{cases} \tag{5.14}$$

avec $V = H_0^1(\Omega)$: espace de Hilbert
$a : u, v \longrightarrow a(u,v) = \iint_\Omega \mathbf{grad}u\ \mathbf{grad}v\, dxdy$ forme bilinéaire symétrique sur V
$l : v \longrightarrow l(v) = \iint_\Omega f\, v\, dxdy$ forme linéaire sur V.
La démonstration de l'existence et de l'unicité se fait par l'utilisation du théorème de Lax-Milgram. On vérifie

1) la continuité de a :

$$|a(u,v)| \leq \Big| \iint_\Omega \mathbf{grad}u\ \mathbf{grad}v\, dxdy \Big| \leq \|\mathbf{grad}u\,\|_{0,2}\, \|\mathbf{grad}v\|_{0,2}\ \leq \|u\,\|_{1,2}\, \|v\|_{1,2}$$

2) l'ellipticité de a :

$$a(v,v) = \iint_\Omega (\mathbf{grad}v)^2\, dxdy \geq C(\Omega)\|v\|_{1,2}^2$$

grâce à l'inégalité de Poincaré en dimension deux vérifiée pour $v \in H_0^1(\Omega)$.

3) la continuité de l :

$$|l(v)| = \Big| \iint_\Omega f\, v\, dxdy \Big| \leq \|f\,\|_{0,2}\, \|v\|_{0,2}\ \leq \|f\,\|_{0,2}\, \|v\|_{1,2}$$

Dès lors qu'il vérifie toutes les hypothèses du théorème de Lax-Milgram, le problème variationnel admet une solution unique $u \in H_0^1(\Omega)$. Comme de plus la forme bilinéaire a est symétrique, cette solution réalise le minimum de l'énergie $J(v) = \frac{1}{2}a(v,v) - l(v)$ soit ici :

$$J(v) = \frac{1}{2} \iint_\Omega (\mathbf{grad}v)^2\, dxdy\ -\ \iint_\Omega f\, v\, dxdy \tag{5.15}$$

Retour à la forme différentielle

Pour montrer que la solution u obtenue est solution forte, c'est à dire solution du problème (5.12) dans le cas ou elle est suffisamment régulière (C^2), il suffit comme en dimension un d'intégrer "par parties", c'est à dire d'utiliser la formule

de Green, à l'envers. On obtient alors le résultat suivant : u solution du problème dans $H_0^1(\Omega)$, donc nulle au bord vérifie

$$-\iint_\Omega \Delta u \; v dxdy = \iint_\Omega f\, v\, dxdy \quad \forall\, v \in H_0^1(\Omega)$$

Ceci entraîne par densité :

$$-\iint_\Omega (\Delta u \; + f)\, v dxdy = 0 \quad \forall\, v \in L^2(\Omega)$$

et donc $-\Delta u = f$ au sens de L^2. Si $u \in C^2(\Omega)$ on en déduit l'égalité au sens fort.

5.7.2 Problème de Dirichlet non-homogène

Dans ce cas les valeurs de la solution sont fixées mais non nécessairement nulles sur la frontière. Soit Ω un domaine ouvert de $\mathbb{R}^2$ de frontière Γ , on considère le problème

$$\begin{cases} -\Delta \; u\,(x,y) = f\,(x,y) & \forall\; x,y \in \Omega \\ u_{\,|\,\Gamma} = u_d \end{cases} \tag{5.16}$$

Comme en dimension un, on se ramène au problème homogène en introduisant une fonction auxiliaire simple prenant les valeurs imposées sur la frontière Γ : $u_{0_{|\Gamma}} = u_d$ et à poser $u = \tilde{u} + u_0$. Le problème se ramène alors à un problème de Dirichlet homogène pour la nouvelle inconnue $\tilde{u}$, soit :

$$\begin{cases} -\Delta \; \tilde{u}\,(x,y) = f\,(x,y) + \Delta \; u_0\,(x,y) & \forall\; x,y \in \Omega \\ \tilde{u}_{\,|\,\Gamma} = 0 \end{cases}$$

La formulation variationnelle de ce problème s'écrit :

$$\begin{cases} \text{Trouver } \; \tilde{u} \in H_0^1(\Omega) \; \text{ telle que} \quad \forall\, v \in H_0^1(\Omega) : \\ \displaystyle\iint_\Omega \mathbf{grad}\tilde{u}\, \mathbf{grad} v \, dxdy = \iint_\Omega f\, v\, dxdy - \iint_\Omega \mathbf{grad} u_0\, \mathbf{grad} v \, dxdy \end{cases} \tag{5.17}$$

On verra plus loin, lors de l'approximation par éléments finis de ce problème, quel est le bon choix pratique de la fonction u_0. Nous renvoyons à la remarque 5.3.1 pour le choix de l'espace des fonctions tests, qui est ici, comme dans le cas monodimensionnel, l'espace H_0^1.

On démontre également, dans le cas Dirichlet non-homogène, l'existence et l'unicité de la solution, par la vérification des hypothèses du théorème de Lax-Milgram.

5.8 Problème de Neumann en dimension deux

Il s'agit de problèmes où les conditions aux limites sur Γ portent sur la dérivée normale de la solution. Ces conditions modélisent la donnée physique d'une force ou d'un flux sur Γ. On considère le problème de Neumann suivant :

$$\begin{cases} -\Delta\, u\,(x,y) \,+\, \alpha(x,y)\, u(x,y) = f\,(x,y) \qquad \forall\, x,y \in \Omega \\ \dfrac{\partial u}{\partial n}\Big|_{\Gamma} = g \end{cases} \tag{5.18}$$

Formulation variationnelle

Les valeurs de u sur Γ sont ici des inconnues du problème. La formulation variationnelle s'écrit dans l'espace $H^1[\Omega]$.

$$\begin{cases} \text{Trouver } \; u \in H^1(\Omega) \;\text{ telle que : } \quad \forall\, v \in H^1(\Omega) \\ \displaystyle\iint_{\Omega} \mathbf{grad} u\ \mathbf{grad} v\, dxdy \,+\, \iint_{\Omega} \alpha\, u\ v\, dxdy \,=\, \iint_{\Omega} f\, v\, dxdy + \int_{\Gamma} g\, v\, d\gamma \end{cases}$$

Existence et unicité du problème

On suppose que la fonction $\alpha \in L^{\infty}[\Omega]$ est minorée sur Ω par un nombre m **strictement** positif et majoré par un réel M et que $g \in L^2[\Gamma]$. La démonstration de l'existence et de l'unicité se fait par l'utilisation du théorème de Lax-Milgram. On vérifie dans l'espace de Hilbert $H^1(\Omega)$

1) la continuité de la forme bilinéaire symétrique :

$$\Big|\iint_{\Omega} \mathbf{grad} u\ \mathbf{grad} v\, dxdy \,+\, \iint_{\Omega} \alpha\, u\ v\, dxdy\Big| \leq max(1,M)\, \|u\|_{1,2}, \|v\|_{1,2}$$

2) l'ellipticité de la forme bilinéaire symétrique :

$$\iint_{\Omega} \mathbf{grad} v^2\, dxdy \,+\, \iint_{\Omega} \alpha\ v^2\, dxdy \,\geq min(1,m)\, \|\, v\,\|^2_{1,2}$$

3) la continuité de la forme linéaire :

$$\Big|\iint_{\Omega} f\, v\, dxdy + \int_{\Gamma} g\, v\, d\gamma\Big| \leq (\|f\|_{0,2,\Omega} + \|g\|_{0,2,\Gamma})\|v\|_{1,2,\Omega}$$

Les hypothèses du théorème de Lax-Milgram étant vérifiées, le problème admet une solution unique dans $H^1(\Omega)$ qui réalise également, puisque la forme bilinéaire est symétrique, le minimum de l'énergie :

$$J(v) = \frac{1}{2}\iint_{\Omega} (\mathbf{grad} v)^2\, dxdy \;-\; \iint_{\Omega} f\, v\, dxdy - \int_{\Gamma} g\, v\, d\gamma$$

dans $H^1(\Omega)$.

Remarque 5.8.1 *Dans le cas $\alpha = 0$, comme en dimension un, on n'a plus ellipticité de a, le problème est mal posé. Il est d'ailleurs clair qu'alors, la fonction inconnue u n'apparaissant dans le problème que par l'intermédiaire des valeurs de ses dérivées, on a une infinité de solutions à une constante additive près. Il y a, de plus, une condition de compatibilité entre le second membre et les conditions aux limites de Neumann à vérifier pour avoir une solution. En effet, par la formule de Stokes, on a*

$$\iint_\Omega \Delta u dxdy = \int_\Gamma \frac{\partial u}{\partial n} d\gamma$$

ce qui impose

$$\iint_\Omega f dxdy + \int_\Gamma g d\gamma = 0$$

Retour à la forme différentielle

Pour montrer que la solution u obtenue est solution forte, c'est à dire solution du problème différentiel dans le cas ou elle est suffisamment régulière (C^2), il suffit d'utiliser la formule de Green, à l'envers.

On obtient alors le résultat suivant : u solution du problème dans $H^1(\Omega)$ vérifie

$$\iint_\Omega (-\Delta u + \alpha u)\, v dxdy + \int_\Gamma (\frac{\partial u}{\partial n} - g) v\, d\gamma = \iint_\Omega f\, v\, dxdy \quad \forall\, v \in H^1(\Omega)$$

Ceci entraîne en choisissant v dans $H_0^1(\Omega)$ puis par densité dans $L^2(\Omega)$:

$$\iint_\Omega (-\Delta u + \alpha u - f)\, v dxdy = 0 \quad \forall\, v \in L^2(\Omega)$$

et donc $-\Delta u + \alpha u = f$ au sens de L^2. Puis :

$$\int_\Gamma (\frac{\partial u}{\partial n} - g) v\, d\gamma = 0 \quad \forall\, v \in H^1(\Omega)$$

donc la vérification de la condition aux limites de Neuman, mais au sens faible, c'est à dire au sens de L^2. Si $u \in C^2(\Omega)$ on en déduit l'égalité et la vérification du problème de Neuman non-homogène au sens fort.

5.9 Problème de Fourier en dimension deux

Les conditions aux limites de Fourier expriment une relation affine entre u et sa dérivée normale $\dfrac{\partial u}{\partial n}$ sur la frontière Γ. Elles incluent comme cas particulier les

conditions de Neumann et permettent dans la pratique de modéliser les conditions de Dirichlet par une technique de pénalisation en prenant un coefficient d'échange k très grand. On considère le problème de Fourier suivant :

$$\begin{cases} -\Delta\, u\,(x,y) \,+\, \alpha(x,y)\, u(x,y) = f\,(x,y) \qquad \forall\, x,y \in \Omega \\ \dfrac{\partial u}{\partial n} = -k(u-u_0) + \beta \quad \text{sur} \quad \Gamma \end{cases} \tag{5.19}$$

Formulation variationnelle de ce problème

La formulation de ce problème s'écrit dans $H^1(\Omega)$ tout entier. En multipliant scalairement dans $L^2(\Omega)$ par v et en utilisant la formule de Green on obtient la formulation variationnelle :

$$\begin{cases} \text{Trouver } \; u \in H^1(\Omega) \; \text{ telle que :} \\ \displaystyle\iint_\Omega \mathbf{grad} u \; \mathbf{grad} v \, dxdy \,+\, \iint_\Omega \alpha\, u \; v \, dxdy \,+\, \int_\Gamma ku\, v \, d\gamma \\ \qquad\qquad = \displaystyle\iint_\Omega f\, v \, dxdy + \int_\Gamma (ku_0 + \beta)\, v \, d\gamma \quad \forall\, v \in H^1(\Omega) \end{cases}$$

Existence et unicité du problème

On suppose que la fonction α est minorée sur Ω par un nombre m **strictement** positif et majorée par un réel M et que k est un réel positif. La démonstration de l'existence et de l'unicité se fait sans difficulté par l'utilisation du théorème de Lax-Milgram.

Remarque 5.9.1 *On a également ellipticité de a et donc existence et unicité si $\alpha = 0$ et $k > 0$. Ceci permet en particulier de justifier l'existence et l'unicité d'une solution dans le cas du traitement pratique des conditions de Dirichlet comme des conditions de Fourier avec k très grand.*

Chapitre 6

L'équation de Laplace

Introduction aux problèmes elliptiques.

6.1 Equation de Laplace dans $\mathbb{R}^n$

Définition 6.1.1 *Soit $u(x, y, ...)$ une fonction définie sur un domaine D de $\mathbb{R}^n$, et vérifiant dans ce domaine l'équation de Laplace :*

$$\Delta u = 0 \tag{6.1}$$

Les fonctions qui vérifient cette équation sont dites ***harmoniques*** *dans D.*

L'étude mathématique de l'équation bidimensionelle (c. à d. dans $\mathbb{R}^2$) va permettre de dégager, dans un cas simple, les principales propriétés des problèmes regroupés sous le vocable de **problèmes elliptiques** (cf. ch. 6).

6.2 Equation de Laplace dans un demi-plan

Considérons le problème physique suivant : on veut connaître la température dans un demi plan connaissant la température sur le bord, sachant que cette température tend vers 0 en s'éloignant de ce bord et qu'il n'y a aucun apport de chaleur.
Le modèle mathématique correspondant s'écrit :

$$\begin{cases} \text{Trouver } u\,(x,y) \to u(x,y) \text{ telle que :} \\ \Delta u = \dfrac{\partial^2 u}{\partial x^2}(x,y) + \dfrac{\partial^2 u}{\partial y^2}(x,y) = 0 \qquad \forall x \in \mathbb{R} \text{ et } \forall y \in \mathbb{R}^+ \\ u(x,0) = f(x) \qquad \text{donnée} \\ u(x,+\infty) = 0 \end{cases} \tag{6.2}$$

6.2.1 Résolution par la transformée de Fourier

On introduit la transformée de Fourier de u par rapport à x , notée $U(\nu, y)$ et définie par :

$$U(\nu, y) = \int_{-\infty}^{+\infty} u(x, y) \exp(-2i\pi\nu x) dx \tag{6.3}$$

La transformée de Fourier de l'équation s'écrit :

$$(-2i\pi\nu)^2 U(\nu, y) + \frac{\partial^2}{\partial y^2} U(\nu, y) = 0 \tag{6.4}$$

La résolution de l'équation différentielle du second ordre en y donne :

$$U(\nu, y) = A(\nu) \exp(2\pi\nu y) + B(\nu) \exp(-2\pi\nu y) \tag{6.5}$$

Comme la fonction $u(x, y)$ tend vers 0 lorsque y tend vers $+\infty$ quelque soit x d'après (6.2), la transformée de Fourier $U(\nu, y)$ doit aussi tendre vers 0. Comme $\exp(2\pi\nu y)$ tend vers $+\infty$ lorsque y tend vers $+\infty$ quand $v > 0$, il faut donc que $A(\nu) = 0$ pour $\nu > 0$. De même il faut que $B(\nu) = 0$ pour $\nu < 0$.
La condition pour $y = 0$, implique $U(\nu, 0) = F(\nu)$ où $F(\nu)$ est la transformée de Fourier de la donnée $f(x)$. On a donc $A(\nu) = F(\nu)$ pour $\nu < 0$ et $B(\nu) = F(\nu)$ pour $\nu > 0$.
On obtient donc :

$$U(\nu, y) = F(\nu) \exp(-2\pi|\nu|y) \tag{6.6}$$

Or, d'après les formules de transformées de Fourier,

$$\exp(-2\pi|\nu|y) = F(\frac{y}{\pi(x^2 + y^2)}) \tag{6.7}$$

La transformée de Fourier d'un produit de convolution est le produit des transformées de Fourier.Comme U est le produit de $F(\nu)$ par $\exp(-2\pi|\nu|y)$, la transformée de Fourier inverse de U est le produit de convolution, par rapport à la variable x , de la fonction f avec la fonction $\frac{y}{\pi(x^2+y^2)}$. La solution u s'écrit donc :

$$u(x, y) = \int_{-\infty}^{+\infty} f(s) \frac{y}{\pi((x - s)^2 + y^2)} ds \tag{6.8}$$

6.3 Equation de Laplace dans un cercle ou une sphère

Si l'on veut résoudre l'**équation de Laplace dans un cercle** centré à l'origine de rayon a, il est naturel de passer en coordoonnées polaires.

Si l'on ne cherche que les solutions **indépendantes de l'angle polaire**, c'est à dire les fonctions $v(r) = v(\sqrt{x^2+y^2}) = u(x,y)$, on doit alors résoudre l'équation différentielle suivante :

$$\frac{\partial^2 v}{\partial r^2} + \frac{1}{r}\frac{\partial v}{\partial r} = 0 \quad \text{pour } r \in]0, a] \tag{6.9}$$

Ce qui entraine $\dfrac{\partial v}{\partial r} = \dfrac{C_1}{r}$ et donc

$$v(r) = C_1 \ln r + C_2 \tag{6.10}$$

On vérifie que cette solution n'est pas définie en $r = 0$. De la même façon, on

montre que les fonctions harmoniques qui ne dépendent que de la distance au centre dans une **sphère** de centre O et de rayon a, doivent vérifier l'équation différentielle suivante :

$$\frac{\partial^2 v}{\partial r^2} + \frac{2}{r}\frac{\partial v}{\partial r} = 0 \quad \text{pour } r \in]0, a] \tag{6.11}$$

et sont de la forme :

$$v(r) = C_1 \frac{1}{r} + C_2 \tag{6.12}$$

6.4 Equation de Laplace dans un rectangle

On considère le problème physique suivant : un rectangle sans apport de chaleur à l'intérieur, sans échange de chaleur sur deux des côtés opposés (côtés adiabatiques) et avec une température imposée sur les deux autres côtés. On cherche la température dans le rectangle. Les équations du problème sont alors :

$$\begin{cases} \text{Trouver } u\,(x,y) \to u(x,y) \text{ telle que :} \\ \\ \Delta u = \dfrac{\partial^2 u}{\partial x^2}(x,y) + \dfrac{\partial^2 u}{\partial y^2}(x,y) = 0 \quad \forall x \in]0,a[\text{ et } \forall y \in]0,b[\\ \\ \frac{\partial u}{\partial n}(0,y) = 0 \quad \frac{\partial u}{\partial n}(a,y) = 0 \quad \forall y \in]0,b[\\ \\ u(x,0) = \varphi(x) \quad \text{et} \quad u(x,b) = \psi(x) \;\; \forall x \in]0,a[\end{cases} \tag{6.13}$$

où φ et ψ sont des données.

6.4.1 Méthode de séparation des variables.

Pour résoudre dans un rectangle ce problème on utilise la méthode de séparation des variables appelée aussi méthode de Fourier.

— On néglige temporairement la condition non-homogène qui dans cet exemple est une condition de type Dirichlet.

— On cherche les solutions du problème (6.13) avec conditions de Neumann homogènes , qui soient du type $u(x,y) = X(x)Y(y)$ et non identiquement nulles.
L'équation (6.13) conduit à :

$$X"(x)Y(y) + Y"(y)X(x) = 0 \tag{6.14}$$

Comme une fonction ne dépendant que de x ne peut être égale à une fonction de y que si ces deux fonctions sont constantes, on a :

$$\frac{X"(x)}{X(x)} = -\frac{Y"(y)}{Y(y)} = \lambda \tag{6.15}$$

La condition de Neumann implique :

$$X'(0)Y(y) = 0 \quad \text{et} \quad X'(a)Y(y) = 0 \tag{6.16}$$

Puisque l'on cherche les solutions $X(x)Y(y)$ non identiquement nulles, il faut que $Y(y) \neq 0$ et donc :

$$X'(0) = 0 \quad \text{et} \quad X'(a) = 0 \tag{6.17}$$

— On commence donc par résoudre le problème en X. On cherche donc pour quelles valeurs de la constante λ le problème suivant a des solutions $X(x)$ non identiquement nulles :

$$\begin{cases} X"(x) = \lambda X(x) \quad \forall x \in]0,a[\\ X'(0) = 0 \\ X'(a) = 0 \end{cases} \tag{6.18}$$

Ce problème consiste en fait à chercher les valeurs propres et les vecteurs propres de l'opérateur "dérivée seconde" dans le sous espace vectoriel de $L^2(0,a)$ formé par les fonctions deux fois continûment dérivables sur $]0,a[$ et de dérivées nulles en $x = 0$ et $x = a$.

— En multipliant (6.18) par $X(x)$ et en intégrant par parties sur $]0,a[$, on obtient :

$$[X'(x)X(x)]_0^a - \int_0^a (X'(x))^2 dx = \lambda \int_0^a (X(x))^2 dx \tag{6.19}$$

Les conditions de Neumann homogènes impliquent que le premier terme est nul et comme les intégrales sont positives ou nulles, les valeurs propres ne peuvent être que négatives ou nulles.

— Pour $\lambda = 0$, on a $X(x) = A + Bx$, avec $B = 0$ à cause de les conditions de Neumann homogènes.
Le problème a donc pour valeur propre $\lambda_0 = 0$ et les fonctions propres associées sont les fonctions contantes $X_0(x) = A_0$.

— Pour $\lambda < 0$, on pose pour simplifier l'écriture $\lambda = -\omega^2$, on a alors $X(x) = A\cos(\omega x) + B\sin(\omega x)$, d'où $X'(x) = \omega(-A\sin(\omega x) + B\cos(\omega x))$ et d'après les conditions de Neumann homogènes et puisque $\omega \neq 0$, on a :

$$B = 0 \tag{6.20}$$

$$-A\sin(\omega a) + B\cos(\omega a) = 0 \tag{6.21}$$

Si on veut avoir $X(x)$ non identiquement nulle il faut que $A \neq 0$ et donc que $\sin(\omega a) = 0$, ce qui n'est possible que si $\omega a = n\pi$ où n est un entier relatif.
Les valeurs propres sont donc de la forme :

$$\begin{aligned} &\lambda_1 = -(\frac{\pi}{a})^2, \qquad \lambda_2 = -(\frac{2\pi}{a})^2, \qquad \lambda_3 = -(\frac{3\pi}{a})^2, \\ &\lambda_4 = -(\frac{4\pi}{a})^2, \qquad ..., \qquad ..., \lambda_n = -(\frac{n\pi}{a})^2, \end{aligned} \tag{6.22}$$

Les fonctions propres associées à chaque λ_n sont les fonctions :

$$X_n(x) = \cos(\frac{n\pi}{a}x) \tag{6.23}$$

— Pour chacunes des valeurs propres λ_n trouvées, on résout le problème en Y_n correspondant :

$$Y_n"(y) = (\frac{n\pi}{a})^2 Y_n(y) \quad \forall y \in]0, b[\tag{6.24}$$

D'où, pour $\lambda_0 = 0$:

$$Y_0(y) = C_0 y + D_0 \tag{6.25}$$

et pour $\lambda_n = -(\frac{n\pi}{a})^2$:

$$Y_n(y) = C_n \exp(\frac{n\pi y}{a}) + D_n \exp(-\frac{n\pi y}{a}) \tag{6.26}$$

— La fonction $U_n(x,y) = X_n(x)Y_n(y)$ est donc solution du problème linéaire (6.13)avec conditions de Neumann homogènes).
D'après le principe de superposition, la somme de toutes les fonctions U_n est aussi solution de ce problème. Cette solution s'écrit :

$$U(x,y) = C_0 y + D_0 + \sum_{n=1}^{\infty} \cos(\frac{n\pi}{a}x)\left(C_n \exp(\frac{n\pi y}{a}) + D_n \exp(-\frac{n\pi y}{a})\right) \tag{6.27}$$

— Il faut donc maintenant s'assurer que la condition de Dirichlet (que l'on avait négligée jusqu'alors) peut être vérifiée par la fonction U si on choisit de façon convenable les coefficients C_n et D_n. Il faut donc que :

$$D_0 + \sum_{n=1}^{\infty} \cos(\frac{n\pi}{a}x)\,(C_n + D_n) = \varphi(x)$$

$$C_0 b + D_0 + \sum_{n=1}^{\infty} \cos(\frac{n\pi}{a}x)\left(C_n \exp(\frac{n\pi b}{a}) + D_n \exp(-\frac{n\pi b}{a})\right) = \psi(x)$$

Or la propriété fondamentale est que les fonctions propres $X_n(x)$ trouvées forment une base des fonctions de $L^2(0,a)$. On peut donc décomposer de façon unique φ et ψ sur cette base :

$$\varphi(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos(\frac{n\pi}{a}x) \tag{6.28}$$

$$\psi(x) = \alpha_0 + \sum_{n=1}^{\infty} \alpha_n \cos(\frac{n\pi}{a}x) \tag{6.29}$$

En utilisant l'unicité de la décomposition, on obtient :

$$D_0 = a_0 \tag{6.30}$$

$$C_0 b + D_0 = \alpha_0 \tag{6.31}$$

$$C_n + D_n = a_n \tag{6.32}$$

$$C_n \exp(\frac{n\pi b}{a}) + D_n \exp(-\frac{n\pi b}{a}) = \alpha_n \tag{6.33}$$

Ce qui permet de déterminer de façon unique les constantes C_n et D_n.

— Exemple : Cas $\varphi(x) = 0$ et $\psi(x) = \cos^2(\frac{\pi}{a}x)$.
Dans ce cas $a_n = 0$ pour tout $n \geq 0$ et $\alpha_0 = 1/2$, $\alpha_2 = 1/2$ et $\alpha_n = 0$ pour

tout n , $n \neq 0$ et $n \neq 2$.
On a alors :

$$D_0 = 0 \tag{6.34}$$

$$C_0 = \frac{1}{2b} \tag{6.35}$$

$$C_n + D_n = 0 \quad \forall n > 0 \tag{6.36}$$

$$C_2 = \frac{1}{4sh(\frac{2\pi b}{a})} \tag{6.37}$$

$$C_n = 0 \quad \forall n,\ n \neq 0 \ \text{ et } \ n \neq 2 \tag{6.38}$$

La solution explicite du problème est alors :

$$u(x,y) = \frac{1}{2b}y + \cos(\frac{2\pi}{a}x)\frac{sh(\frac{2\pi y}{a})}{2sh(\frac{2\pi b}{a})} \tag{6.39}$$

Remarque : Pour pouvoir utiliser la méthode de séparation des variables, il est essentiel qu'il y aît des conditions homogènes sur deux cotés opposés afin de pouvoir chercher les valeurs propres et les fonctions propres d'un problème linéaire et homogène.
Si ce n'est pas le cas, il est toujours possible d'après la linéarité du problème de superposer des problèmes ayant cette propriété.
Par exemple, soit u solution du problème (P) défini par :

$$\begin{cases} \text{Trouver } u\,(x,y) \to u(x,y) \text{ telle que :} \\ \dfrac{\partial^2 u}{\partial x^2}(x,y) + \dfrac{\partial^2 u}{\partial y^2}(x,y) = 0 \quad \forall x \in]0,a[\text{ et } \forall y \in]0,b[\\ \dfrac{\partial u}{\partial n}(0,y) = \chi(y) \quad \dfrac{\partial u}{\partial n}(a,y) = \eta(y) \quad \forall y \in]0,b[\\ u(x,0) = \varphi(x) \quad \text{et} \quad u(x,b) = \psi(x) \ \ \forall x \in]0,a[\end{cases} \tag{6.40}$$

alors u est la somme de u_1 et u_2 , solutions respectives des problèmes (P1) et (P2) suivants :

$$\begin{cases} \dfrac{\partial^2 u_1}{\partial x^2}(x,y) + \dfrac{\partial^2 u_1}{\partial y^2}(x,y) = 0 \quad \forall x \in]0,a[\text{ et } \forall y \in]0,b[\\ \dfrac{\partial u_1}{\partial n}(0,y) = 0 \text{ et } \dfrac{\partial u_1}{\partial n}(a,y) = 0 \qquad \forall y \in]0,b[\\ u_1(x,0) = \varphi(x) \text{ et } u_1(x,b) = \psi(x) \quad \forall x \in]0,a[\end{cases} \tag{6.41}$$

et :

$$\begin{cases} \dfrac{\partial^2 u_2}{\partial x^2}(x,y) + \dfrac{\partial^2 u_2}{\partial y^2}(x,y) = 0 \quad \forall x \in]0,a[\text{ et } \forall y \in]0,b[\\ \dfrac{\partial u_2}{\partial n}(0,y) = \chi(y) \text{ et } \dfrac{\partial u_2}{\partial n}(a,y) = \eta(y) \quad \forall y \in]0,b[\\ u_2(x,0) = 0 \text{ et } u_2(x,b) = 0 \; quad \forall x \in]0,a[\end{cases} \tag{6.42}$$

6.5 Equation de Laplace dans un domaine borné Ω quelconque

Si le domaine Ω est de forme quelconque (ni rectangle, ni cercle, ni couronne circulaire), il n'est pas possible de donner une solution explicite comme dans les deux cas précédents. On peut seulement chercher une solution numérique à l'aide de la méthode des éléments finis et après avoir écrit le problème sous la forme faible. On obtient cette forme faible en multipliant l'équation (6.1) par une fonction quelconque $\varphi(x,y)$ et en intégrant sur le domaine Ω . En utilisant la deuxième formule de Green on a :

$$\int_\Omega \varphi \Delta u = \int_{\partial\Omega} \varphi \frac{\partial u}{\partial n} - \int_\Omega \mathbf{grad}\ \varphi . \mathbf{grad}\ u \tag{6.43}$$

d'où :

$$0 = \int_{\partial\Omega} \varphi \frac{\partial u}{\partial n} - \int_\Omega \mathbf{grad}\ \varphi . \mathbf{grad}\ u \tag{6.44}$$

Il reste ensuite à tenir compte des conditions aux bords.
Ceci sera étudié en détail en GM3.

6.6 Propriétés fondamentales des fonctions harmoniques

Une propriété remarquable des fonctions harmoniques, c'est à dire telles que $\Delta u = 0$, est le principe suivant :

6.6.1 Principe du maximum

Théorème 6.6.1 *Si u est une fonction deux fois continûment dérivable qui vérifie $\Delta u = 0$ dans l'ouvert Ω , et si u n'est pas constante, alors u atteint son maximum sur le bord de Ω .*

6.6.2 Unicité de la solution

Nous avons utilisé des méthodes très diverses pour trouver des solutions de l'équation de Laplace avec des conditions de Dirichlet ou de Neumann sur le bord du domaine. La question qui se pose est de savoir si la solution est unique.

Le principe du maximum entraine l'unicité de la solution du problème suivant.

Théorème 6.6.2 *Le problème suivant a une solution u et cette solution est unique si Γ_1 est un ouvert non vide du bord $\partial\Omega$ et si $\Gamma_1 \cup \Gamma_2 = \partial\Omega$*

$$\Delta u = 0 \text{ dans } \Omega \tag{6.45}$$

$$u = \varphi \text{ sur } \Gamma_1 \tag{6.46}$$

$$\frac{\partial u}{\partial n} = \eta \text{ sur } \Gamma_2 \tag{6.47}$$

Si $\Gamma_2 = \partial\Omega$ et donc si Γ_1 est vide, il est clair que, si il y a une solution u , alors il en a une infinité puisque pour toute constante c , la fonction $u + c$ est aussi solution. Alors, il n'y a pas unicité.

6.6.3 Propriété de la moyenne

Théorème de la moyenne
Si u est une fonction deux fois continûment dérivable qui vérifie $\Delta u = 0$ dans le disque D , alors la valeur de u au centre est égale à la moyenne des valeurs de u sur le bord de D.

La démonstration est basée sur la troisième formule de Green :

$$\int_\Omega \varphi\Delta\psi - \psi\Delta\varphi = \int_{\partial\Omega} \varphi\frac{\partial\psi}{\partial n} - \psi\frac{\partial\varphi}{\partial n} \tag{6.48}$$

On pose $\psi = u$, on choisit pour fonction φ une fonction harmonique dans le cercle D privé de son centre et on prend pour domaine Ω, l'anneau formé par le disque D privé du disque de rayon ϵ petit.
Dans $\mathbb{R}^2$ on peut prendre pour fonction φ la fonction harmonique $\ln r$ où r est la distance du point considéré au centre du disque.
On obtient alors le résultat, en faisant tendre ϵ vers 0 . Un théorème similaire

se démontre dans une sphère de $\mathbb{R}^3$ (on choisit alors la fonction harmonique $\varphi(r) = \frac{1}{r}$).

Chapitre 7

Éléments finis monodimensionnels

7.1 Principes généraux de l'approximation

7.1.1 Une famille de problèmes variationnels linéaires

Dans le chapitre 5, nous avons obtenu les formulations variationnelles de problèmes aux limites sous la forme générale :

$$(P)\begin{cases} \text{Trouver la fonction } u \text{ appartenant à l'Hilbert } V \text{ telle que :} \\ a(u,v) = l(v) \quad \forall v \in V \end{cases} \tag{7.1}$$

Sous les hypothèses : a forme bilinéaire continue sur V et elliptique, l une forme linéaire continue sur V, le théorème de Lax-Milgram conduit aux conclusions suivantes :

1) Le problème P admet une solution unique u dans V

2) Si la forme bilinéaire a est symétrique le problème variationnel P est équivalent au problème de minimisation suivant :

$$\begin{cases} \text{Trouver la fonction } u \in V \text{ qui minimise la forme quadratique} \\ J(v) = \frac{1}{2}\, a(v,v) \;-\; l(v) \end{cases} \tag{7.2}$$

7.1.2 Approximation interne du problème

On suppose maintenant que l'on connaît un sous-espace $V_h \subset V$ de dimension finie, paramétré par h et tel que pour tout $v \in V$, il existe un élément $r_h v \in V_h$ vérifiant :

$$\lim_{h \to 0} \|r_h v - v\| = 0$$

On parle d'approximation interne car $V_h \subset V$.

Considérons alors le problème P_h

$$(P_h) \begin{cases} \text{Trouver la fonction } u_h \text{ appartenant à } V_h \text{ telle que :} \\ a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h \end{cases} \tag{7.3}$$

Ce problème admet également une solution unique car $V_h \subset V$, et donc les hypothèses du théorème de Lax-Milgram sont également vérifiées dans V_h.

De même, si la forme bilinéaire a est symétrique le problème variationnel P_h est équivalent au problème de minimisation suivant :

$$\begin{cases} \text{Trouver la fonction } u_h \in V_h \text{ qui minimise la forme quadratique} \\ J(v_h) = \frac{1}{2}\, a(v_h, v_h) \;-\; l(v_h) \end{cases} \tag{7.4}$$

et on a donc évidemment

$$J(u_h) \;\geq\; J(u)$$

7.1.3 Un résultat général de majoration d'erreur

Théorème 7.1.1 *Soit M la constante intervenant dans l'hypothèse de continuité de a : $a(u,v) \leq M\,\|u\|\,\|v\|$ et m la constante intervenant dans l'hypothèse d'ellipticité : $a(v,v) \geq m\|v\|^2$, on a la majoration d'erreur suivante :*

$$\|u - u_h\| \leq \frac{M}{m} \inf_{v_h \in V_h} \|u - v_h\|$$

Démonstration. De

$$a(u,v) = l(v) \quad \forall v \in V$$

et

$$a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h$$

on obtient

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_h$$

et

$$a(u - u_h, u_h - u + u - v_h) = 0 \quad \forall v_h \in V_h$$

soit

$$m\|u - u_h\|^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \leq M\|u - u_h\|\|u - v_h\|$$

et le résultat.

Remarque 7.1.1 *Si a est symétrique :*

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_h$$

signifie que u_h est la projection de u dans V_h au sens du produit scalaire a. Dans ce cas on a

$$\|u - u_h\| \leq \sqrt{\frac{M}{m}} \inf_{v_h \in V_h} \|u - v_h\|$$

7.1.4 Un premier exemple d'approximation interne : la méthode de Galerkin

On suppose l'espace de Hilbert V séparable. Il existe donc une base $\{w_j\}_{j=1}^{+\infty}$ dénombrable engendrant un sous-espace dense dans V. On considère alors un sous-ensemble fini $B_m = \{w_j\}_{j=1}^{m}$ et l'espace V_m engendré par B_m.

Soit Π_m l'opérateur de projection orthogonale de V dans V_m on a :

$$\lim_{m\to\infty} \|\Pi_m v - v\| = 0 \quad \forall v \in V$$

Le problème P_m suivant :

$$P_m \begin{cases} \text{Trouver la fonction } u_m \text{ appartenant à } V_m \text{ telle que :} \\ a(u_m, v_m) = l(v_m) \quad \forall v_m \in V_m \end{cases} \tag{7.5}$$

qui admet une solution unique dans V_m peut s'écrire sous forme d'un système linéaire :

$$\sum_{j=1,m} A_{i,j}\, u_j \;=\; L_i \quad \forall\, i = 1, m$$

avec

$$A_{i,j} = a(w_j, w_i)$$

et

$$L_i = L(w_i)$$

7.2 Éléments finis P1 pour le problème de Dirichlet

7.2.1 Problème de Dirichlet homogène

Reprenons le problème de Dirichlet homogène dont la formulation variationnelle s'écrit :

$$\begin{cases} \text{Trouver la fonction } u \text{ appartenant à } H_0^1[a,b] \text{ telle que :} \\ \displaystyle\int_a^b u'(x)v'(x)dx \;+\; \int_a^b \alpha(x)u(x)v(x)dx \;=\; \int_a^b f(x)v(x)dx \quad \forall v \in H_0^1[a,b] \end{cases}$$

et introduisons une discrétisation de l'intervalle [a,b] en N sous-intervalles ou éléments $T_i = [x_{i-1}, x_i]$. Les éléments T_i n'ont pas forcément même longueur. $V_{0,h}$ est alors l'espace des fonctions continues affines par morceaux (affines sur les segments T_i) et nulles aux extrémités a et b.

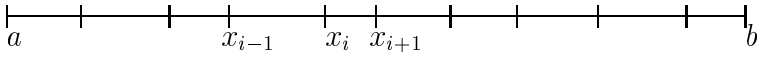


FIGURE 7.1 – Discrétisation (maillage) du segment [a,b] en éléments finis.

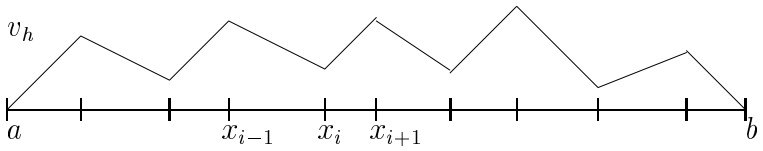


FIGURE 7.2 – Une fonction de $V_{0,h}$

Rappelons que chaque fonction $v_h \in V_{0,h}$ est déterminée de manière unique par la donnée de ses valeurs aux points x_i pour $i = 1, N-1$. L'espace $V_{0,h}$ est de dimension $N-1$ et est engendré par la base de Lagrange de $V_{0,h}$, formée des N-1 fonctions $w_i \in V_{0,h}$ définies par les N-1 conditions suivantes :

$$w_i(x_j) = \delta_{ij} \quad \forall i = 1, N-1 \quad \text{et} \quad \forall j = 1, N-1 \tag{7.6}$$

Une fonction v_h quelconque s'écrit dans cette base :

$$v_h(x) = \sum_{i=1}^{i=N-1} v_i \, w_i(x)$$

avec $v_i = v_h(x_i)$. Les coefficients v_i sont donc les valeurs de v_h aux points (x_i)

7.2.2 Écriture du problème approché

Ecrivons le problème approché dans $V_{0,h}$

$$\begin{cases} \text{Trouver la fonction } u_h \text{ appartenant à } V_{0,h} \text{ telle que} \quad \forall v_h \in V_{0,h} \\ \displaystyle\int_a^b u_h'(x) v_h'(x) dx \; + \; \int_a^b \alpha(x) u_h(x) v_h(x) dx \; = \int_a^b f(x) v_h(x) dx \end{cases} \tag{7.7}$$

Le problème étant linéaire, l'égalité est vraie pour tout v_h si elle est vraie pour une base de l'espace vectoriel $V_{0,h}$

$$\forall v_h \in V_h \Longleftrightarrow \forall w_i \quad \text{pour} \quad i = 1, N-1$$

D'autre part, écrivons u_h, solution du problème approché dans $V_{0,h}$, dans la base des w_i

$$u_h(x) = \sum_{j=1}^{j=N-1} u_j \, w_j(x)$$

avec $u_j = u_h(x_j)$ valeur approchée de la solution exacte au point (x_j)

On obtient l'écriture suivante du problème approché :

$$\left\{ \begin{array}{l} \text{Trouver } u_1,\ u_2, ...u_{N-1} \text{ tels que} \quad \forall i = 1, N-1 \\ \displaystyle\sum_{j=1}^{j=N-1} \Big(\int_a^b w_j'(x)\, w_i'(x)\, dx + \int_a^b \alpha(x)\, w_j(x)\, w_i(x)\, dx \Big)\, u_j = \int_a^b f(x)\, w_i(x)\, dx \end{array} \right.$$

Soit en posant

$$\int_a^b f(x)\, w_i(x)\, dx = F_i$$

et

$$\int_a^b w_j'(x)\, w_i'(x)\, dx + \int_a^b \alpha(x)\, w_j(x)\, w_i(x)\, dx \ = A_{ij}$$

$$\sum_{j=1}^{j=N-1} A_{ij}\ u_j = F_i \qquad \forall i = 1, N-1$$

On a ainsi obtenu un système linéaire de $N-1$ équations à $N-1$ inconnues, qui peut s'écrire sous la forme matricielle suivante :

$$A\ U \ = \ F$$

7.2.3 Calcul des coefficients de la matrice

Le calcul des coefficients de la matrice A et du second membre se fait, comme dans le chapitre 1, par assemblage des contributions des éléments $T_i = [x_{i-1}, x_i]$ pour $i = 1, ..N$. La matrice A apparaît comme la somme de deux matrices K et M.

K constituée des coefficients

$$K_{ij} = \int_a^b w_j'(x)\, w_i'(x)\, dx$$

s'appelle la matrice de raideur.

M constituée, dans le cas $\alpha = 1$ des coefficients

$$M_{ij} = \int_a^b w_j(x)\, w_i(x)\, dx$$

s'appelle la matrice de masse. On obtient sans difficulté les contributions de chaque élément T_i aux matrices de raideur et de masse, dites matrices élémentaires de raideur et matrices élémentaires de masse.

Matrice élémentaire de raideur

On calcule les coefficients K_{ij} en sommant les contributions des différents éléments selon :

$$K_{ij} = \int_a^b w_j'(x)\, w_i'(x)\, dx = \sum_{k=1}^{k=N} \int_{x_{k-1}}^{x_k} w_j'(x)\, w_i'(x)\, dx$$

Considérons par exemple l'élément $T_i = [x_{i-1}, x_i]$. Sur cet élément, il n'y a que 2 fonctions de base non nulles : w_{i-1} et w_i

$$w_{i-1|_{T_i}} = \frac{x_i - x}{x_i - x_{i-1}} \qquad w_{i|_{T_i}} = \frac{x - x_{i-1}}{x_i - x_{i-1}}$$
$$w'_{i-1|_{T_i}} = \frac{-1}{x_i - x_{i-1}} \qquad w'_{i|_{T_i}} = \frac{1}{x_i - x_{i-1}}$$

L'élément T_i produira donc effectivement une contribution non nulle aux 4 coefficients $K_{i-1,i-1}$, $K_{i-1,i}$, $K_{i,i}$ et $K_{i,i-1}$ de la matrice globale K.

Calculons les contributions élémentaires de T_i et disposons les sous la forme d'une matrice élémentaire 2×2

$$ElemK_i = \begin{pmatrix} e^i_{1,1} & e^i_{1,2} \\ e^i_{2,1} & e^i_{2,2} \end{pmatrix}$$

avec

$$e^i_{1,1} = \int_{x_{i-1}}^{x_i} w'_{i-1}(x)^2\, dx = \int_{x_{i-1}}^{x_i} \frac{1}{(x_i - x_{i-1})^2}\, dx = \frac{1}{x_i - x_{i-1}}$$

$$e^i_{1,2} = e^i_{2,1} = \int_{x_{i-1}}^{x_i} w'_{i-1}(x)\, w'_i(x)\, dx = \int_{x_{i-1}}^{x_i} -\frac{1}{(x_i - x_{i-1})^2}\, dx = -\frac{1}{x_i - x_{i-1}}$$

$$e^i_{2,2} = \int_{x_{i-1}}^{x_i} w'_i(x)^2\, dx = \int_{x_{i-1}}^{x_i} \frac{1}{(x_i - x_{i-1})^2}\, dx = \frac{1}{x_i - x_{i-1}}$$

d'où

$$ElemK_i = \begin{pmatrix} \dfrac{1}{x_i - x_{i-1}} & \dfrac{-1}{x_i - x_{i-1}} \\ \dfrac{-1}{x_i - x_{i-1}} & \dfrac{1}{x_i - x_{i-1}} \end{pmatrix} = \frac{1}{x_i - x_{i-1}} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Matrice élémentaire de masse

On obtient de même la matrice de masse élémentaire

$$ElemM_i = \frac{x_i - x_{i-1}}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

7.2.4 Calcul des composantes du second membre

Chaque composante F_i du vecteur second-membre global

$$F_i = \int_a^b f(x)\, w_i(x)\, dx$$

est calculée également par assemblage de contributions élémentaires.

$$F_i = \sum_{k=1}^{k=N} \int_{x_{k-1}}^{x_k} f(x)\, w_i(x)\, dx$$

Tout dépend alors de la donnée de f. Si f est donnée analytiquement, on peut parfois calculer les intégrales exactement à la main. Mais en général, f n'est connue que par ses valeurs aux points x_i pour $i = 0, N$. On écrit donc f dans la base des w_i selon :

$$f(x) = \sum_{j=0}^{j=N} f_j\, w_j(x)\, dx$$

Le problème se ramène alors au calcul des intégrales :

$$\int_{x_{k-1}}^{x_k} w_j(x)\, w_i(x)\, dx$$

On utilise des formules d'intégration numérique. Par exemple la formule des trapèzes

$$\int_{x_{k-1}}^{x_k} F(x)\, dx = \frac{x_k - x_{k-1}}{2}[F(x_{k-1}) + F(x_k)]$$

exacte pour des polynômes de degré 1.

ou la formule de Simpson

$$\int_{x_{k-1}}^{x_k} F(x)\, dx = \frac{x_k - x_{k-1}}{6}[F(x_{k-1}) + 4\, F(x_{k-\frac{1}{2}}) + F(x_k)]$$

exacte pour des polynômes de degré inférieur ou égal à 3.

Avec des fonctions tests $w_i \in P1$, la méthode des trapèzes conduit à une valeur approchée de l'intégrale, qui dans le cas de points équidistribués de pas h redonne le résultat

$$F_i = h\, f_i$$

obtenu en différences finies. La méthode de Simpson permet un calcul exact et donne dans le même cas

$$F_i = \frac{h}{6}[f_{i-1} + 4\, f_i + f_{i+1}]$$

7.2.5 Problème de Dirichlet non-homogène

Le problème de Dirichlet non-homogène s'écrit

$$\begin{cases} -u''(x) + \alpha(x)\, u(x) = f(x) & \forall x \in [a,b] \\ u(a) = u_a \quad u(b) = u_b \end{cases} \tag{7.8}$$

On pose $u = \tilde{u} + u_0$ en choisissant une fonction auxiliaire simple u_0 prenant les valeurs fixées

$$u_0(a) = u_a \quad u_0(b) = u_b$$

Le problème se ramène alors à un problème de Dirichlet homogène pour la nouvelle inconnue $\tilde{u}$ dont la formulation variationnelle s'écrit :

$$\begin{cases} \text{Trouver la fonction } \tilde{u} \text{ appartenant à } H_0^1[a,b] \text{ telle que :} \\ \displaystyle\int_a^b \tilde{u}'(x)v'(x)dx \,+\, \int_a^b \alpha(x)\tilde{u}(x)v(x)dx \,=\, \int_a^b f(x)v(x)dx \\ \displaystyle -\int_a^b u_0'(x)v'(x)dx \,-\, \int_a^b \alpha(x)u_0(x)v(x)dx \quad \forall v \in H_0^1[a,b] \end{cases}$$

On obtient donc dans ce cas le problème approché suivant dans $V_{0,h}$

$$\begin{cases} \text{Trouver la fonction } \tilde{u}_h \text{ appartenant à } V_{0,h} \text{ telle que :} \\ \displaystyle\int_a^b \tilde{u}_h'(x)v_h'(x)dx \,+\, \int_a^b \alpha(x)\tilde{u}_h(x)v_h(x)dx \,=\, \int_a^b f(x)v_h(x)dx \\ \displaystyle -\int_a^b u_0'(x)v_h'(x)dx \,-\, \int_a^b \alpha(x)u_0(x)v_h(x)dx \quad \forall v_h \in V_{0,h} \end{cases}$$

Il reste simplement à préciser le choix pratique de u_0. On prend habituellement pour u_0, la fonction suivante de l'espace V_h des fonctions continues affines par éléments :

$$u_0 = u_a\, w_0 \,+\, u_b\, w_N$$

L'existence et l'unicité des solutions des problèmes continus et discrets se démontrent aisément par application du Théorème de Lax-Milgram.

D'un point de vue pratique, la seule modification à apporter au système par rapport au cas Dirichlet homogène concerne les seules composantes 1 et N-1 du second membre.

On verra plus loin une autre technique de prise en compte de conditions aux limites de Dirichlet déduite du cas général de conditions aux limites de Fourier.

7.3 Approximation du problème de Neumann

La formulation variationnelle du problème de Neumann s'écrit dans l'espace $H^1[a,b]$ tout entier.

$$\left\{\begin{array}{l} \text{Chercher la fonction } u \text{ appartenant à } H^1[a,b] \text{ telle que :} \\ \displaystyle\int_a^b u'(x)v'(x)dx + \int_a^b \alpha(x)u(x)v(x)dx = \int_a^b f(x)v(x)dx \\ \qquad\qquad +\nu v(b) - \mu v(a) \quad \forall v \in H^1[a,b] \end{array}\right.$$

Le problème approché s'écrira dans l'espace V_h des fonctions continues affines par éléments T_i. Sans restriction sur les valeurs aux bords, l'espace V_h est de dimension $N+1$. Il est engendré par les $N+1$ fonctions affines par morceaux w_i définies pour $i = 0,..N$ par :

$$w_i(x_j) = \delta_{ij} \quad \forall i = 0,N \quad \text{et} \quad \forall j = 0,N$$

avec

$$x_0 = a \quad \text{et} \quad x_N = b$$

On obtient cette fois le problème approché :

$$\left\{\begin{array}{l} \text{Trouver } u_0,\ u_1,\ u_2,....u_N \text{ tels que} \quad \forall i = 0,..N \\ \displaystyle\sum_{j=0}^{j=N}\Big(\int_a^b w_j'(x)\,w_i'(x)\,dx + \int_a^b \alpha(x)\,w_j(x)\,w_i(x)\,dx\Big)\,u_j = \int_a^b f(x)\,w_i(x)\,dx \\ \qquad\qquad +\nu w_i(b) - \mu w_i(a) \end{array}\right.$$

On voit que $w_i(b)$ est non nul et égal à 1 pour le seul indice $i = N$. De même $w_i(a)$ est non nul et égal à 1 pour le seul indice $i = 0$.

En posant

$$\int_a^b f(x)\,w_i(x)\,dx = F_i \quad \text{pour} \quad i = 1,..N-1$$

$$\int_a^b f(x)\,w_0(x)\,dx - \mu = F_0$$

$$\int_a^b f(x)\,w_N(x)\,dx + \nu = F_N$$

et

$$\int_a^b w_j'(x)\,w_i'(x)\,dx + \int_a^b \alpha(x)\,w_j(x)\,w_i(x)\,dx = A_{ij} \quad \text{pour} \quad i,j = 0...N$$

le problème approché s'écrit sous la forme du système linéaire de $N+1$ équations à $N+1$ inconnues :

$$\sum_{j=0}^{j=N} A_{ij}\ u_j = F_i \qquad \forall i = 0, N$$

Remarquons que les conditions aux limites sont intervenues uniquement dans les composantes extrêmes du vecteur second-membre.

7.4 Approximation du problème de Fourier

La formulation variationnelle du problème de Fourier

$$\begin{cases} -u''(x) + \alpha(x)\, u(x) = f(x) \qquad \forall x \in [a,b] \\ u'(a) = k_a\,(u(a) - u_a) + \beta_a \\ u'(b) = -k_b\,(u(b) - u_b) + \beta_b \end{cases} \tag{7.9}$$

s'écrit dans l'espace des fonctions de $H^1[a,b]$.

$$\begin{cases} \text{Chercher la fonction } u \text{ appartenant à } H^1[a,b] \text{ telle que : } \forall v \in H^1[a,b] \\ \displaystyle\int_a^b u'(x)v'(x)dx \ + \ \int_a^b \alpha(x)u(x)v(x)dx \ + k_a u(a)v(a) + k_b u(b)v(b) \\ \displaystyle = \int_a^b f(x)v(x)dx + k_a u_a v(a) + k_b u_b v(b) - \beta_a v(a) + \beta_b v(b) \end{cases}$$

Le problème approché s'écrira dans l'espace V_h des fonctions continues affines par éléments T_i. L'espace V_h est de dimension $N+1$. Il est engendré par les N fonctions affines par morceaux $w_i \quad$ pour $\quad i = 0, ..N$

On obtient le problème approché :

$$\begin{cases} \text{Trouver } u_0,\ u_1,\ u_2, ...u_N \text{ tels que} \quad \forall i = 0, ..N \\ \displaystyle\sum_{j=0}^{j=N} \Big(\int_a^b w_j'(x) w_i'(x)\, dx + \int_a^b \alpha(x) w_j(x) w_i(x)\, dx \ + k_a w_j(a) w_i(a) \\ \displaystyle + k_b\, w_j(b)\, w_i(b) \Big)\, u_j = \int_a^b f(x) w_i(x)\, dx + k_a u_a w_i(a) + k_b u_b w_i(b) - \beta_a w_i(a) + \beta_b w_i(b) \end{cases}$$

Par rapport aux cas étudiés précédemment, il convient de noter que les conditions aux limites de Fourier introduisent une modification de la matrice du système

linéaire. L'ajout du terme $k_a\, w_j(a)\, w_i(a)$ entraîne l'addition de la valeur k_a au coefficient $A_{0,0}$ de la matrice et de même pour k_b au coefficient $A_{N,N}$. En effet le produit $w_j(a)\, w_i(a)$ est non nul et égal à 1 dans le seul cas $i = j = 0$ (idem pour $w_j(b)\, w_i(b)$).

Comme on l'a signalé plus haut, le choix d'une valeur très grande de k permet de modéliser une condition aux limites de Dirichlet à partir de la formulation d'une condition de Fourier.

7.5 Assemblage

L'assemblage des matrices et second membres élémentaires, afin de constituer le système global, s'effectue sans difficulté dans une boucle générale de calcul passant en revue les éléments T_i et sommant leurs contributions en les affectant aux coefficients adéquats du système global.
Ici, chaque coefficient A_{ij} est obtenu en sommant les contributions des éléments T_i. Remarquons que seuls 2 éléments produisent une contribution non nulle à chaque coefficient. Ainsi on vérifie que sur chaque ligne i de la matrice A il n'y a que 3 coefficients non nuls $A_{i,i-1}$, $A_{i,i}$, $A_{i,i+1}$.

Supposons un maillage en N éléments. Notons A la matrice globale à assembler (matrice de raideur ou de masse globale, second membre) et a_k les matrices élémentaires correspondantes relatives à chaque élément T_k.
L'algorithme d'assemblage est très simple, dès lors que l'on dispose d'un tableau associant les points d'un élément T_k et les noeuds du maillage global.

Dans ce cas très simple d'éléments de degré un en dimension un, chaque élément T_k comprend 2 noeuds x_{k-1}, x_k.

D'où l'algorithme :

POUR K = 1, N FAIRE ! boucle sur les éléments

POUR i = 1, 2 FAIRE ! boucle sur les numéros locaux
POUR j = 1, 2 FAIRE ! boucle sur les numéros locaux

I = K+i-2 ! numéros globaux
J = K+i-2 ! numéros globaux

A(I,J) = A(I,J) + a(i,j) ! A : matrice globale, a matrice élémentaire

FIN DES 3 BOUCLES

Remarque 7.5.1 *Les indices des coefficients de la matrice correspondent à la numérotation globale adoptée. Ils vont donc de* 0 *à* N.

7.6 Éléments finis de Lagrange de degré deux ou éléments P2

Considérons une discrétisation de l'intervalle $[a, b]$ en N sous-intervalles ou éléments T_i. Les éléments T_i n'ont pas forcément même longueur.
Choisissons comme espace V_h d'approximation, l'espace des fonctions continues sur $[a, b]$, et polynomiales de degré deux sur chaque sous-intervalle. Un polynôme de degré deux est fixé par ses valeurs en trois points. On prend les extrémités et le milieu de chaque élément T_i. On est ainsi amené à considérer une discrétisation de $[a, b]$ en N sous-intervalles comportant eux-mêmes trois points, ce qui nous conduit globalement à une discrétisation par $2N + 1$ points ou noeuds x_i pour $i = 0, ...2N$ selon la figure suivante :

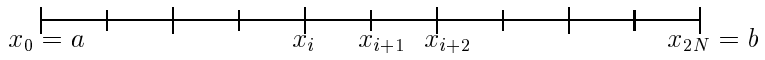


FIGURE 7.3 – Maillage P2 en dimension un

D'où l'allure générale d'une fonction $P2$ appartenant à V_h.

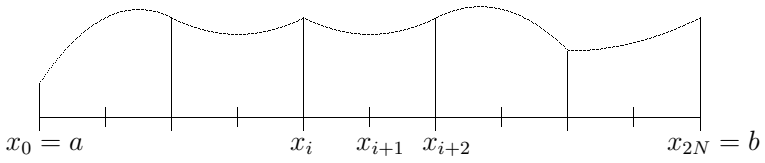


FIGURE 7.4 – Une fonction $P2$ en dimension un

On prend comme base de V_h l'ensemble des $2N + 1$ fonctions $w_i \in V_h$ définies classiquement par les $2N + 1$ conditions

$$w_i(x_j) = \delta_{ij} \quad \forall i = 0, 2N \quad \text{et} \quad \forall j = 0, 2N \tag{7.10}$$

Sur chaque sous-intervalle $[x_i, x_{i+2}]$ on réalise donc une interpolation de Lagrange de degré deux basée sur les trois points équidistants x_i, x_{i+1}, x_{i+2}. Selon la technique classique de Lagrange on construit les trois polynômes :

$$L_i(x) = \frac{(x - x_{i+1})(x - x_{i+2})}{(x_i - x_{i+1})(x_i - x_{i+2})}$$

$$L_{i+1}(x) = \frac{(x - x_i)(x - x_{i+2})}{(x_{i+1} - x_i)(x_{i+1} - x_{i+2})}$$
$$L_{i+2}(x) = \frac{(x - x_i)(x - x_{i+1})}{(x_{i+2} - x_i)(x_{i+2} - x_{i+1})}$$

On aura donc deux types de fonctions de base $P2$. Les fonctions w_i correspondant à un point x_i extrémité d'un élément (les points d'indice pair dans notre numérotation) dont le graphe a l'allure suivante :

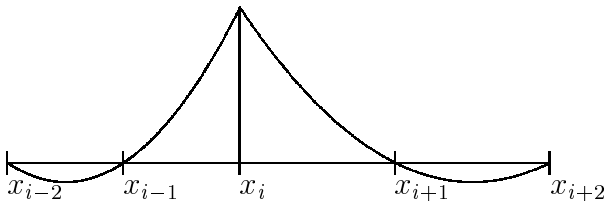


FIGURE 7.5 – Fonction de base $P2$ associée à une extrémité

et qui sont définies par

$$w_i\big|_{[x_{i-2},x_i]}(x) = \frac{(x - x_{i-2})(x - x_{i-1})}{(x_i - x_{i-2})(x_i - x_{i-1})}$$
$$w_i\big|_{[x_i,x_{i+2}]}(x) = \frac{(x - x_{i+1})(x - x_{i+2})}{(x_i - x_{i+1})(x_i - x_{i+2})}$$
$$w_i = 0 \text{ à l'extérieur de } [x_{i-2}, x_{i+2}]$$

Les fonctions de base correspondant à un point milieu d'un élément (points de numéros impair dans le cas de la figure) dont le graphe est le suivant :

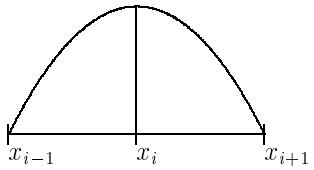


FIGURE 7.6 – Fonction de base $P2$ associée à un milieu

et qui sont définies par

$$w_i\big|_{[x_{i-1},x_{i+1}]}(x) = \frac{(x - x_{i-1})(x - x_{i+1})}{(x_i - x_{i-1})(x_i - x_{i+1})}$$
$$w_i = 0 \text{ à l'extérieur de } [x_{i-1}, x_{i+1}]$$

7.6.1 Approximation du problème de Neumann

Reprenons le problème modèle de Neumann homogène :

$$\begin{cases} -u''(x) + \alpha(x)\, u(x) = f(x) & \forall x \in [a,b] \\ u'(a) = 0 \quad u'(b) = 0 \end{cases} \tag{7.11}$$

avec α strictement positive sur $[a,b]$ telle que : $\alpha(x) \geq \alpha > 0 \quad \forall x \in [a,b]$.

La formulation variationnelle du problème de Neumann homogène s'écrit :

$$\begin{cases} \text{Chercher la fonction } u \text{ appartenant à } H^1[a,b] \text{ telle que :} \\ \displaystyle\int_a^b u'(x)v'(x)dx + \int_a^b \alpha(x)u(x)v(x)dx = \int_a^b f(x)v(x)dx \quad \forall v \in H^1[a,b] \end{cases}$$

Et le problème approché dans l'espace V_h engendré par les fonctions de base w_i :

$$\begin{cases} \text{Trouver } u_0,\ u_1,\ u_2, ...u_{2N} \text{ tels que} \quad \forall i = 0, ..2N \\ \displaystyle\sum_{j=0}^{j=2N} \Big(\int_a^b w_j'(x)\, w_i'(x)\, dx + \int_a^b \alpha(x)\, w_j(x)\, w_i(x)\, dx \Big)\, u_j = \int_a^b f(x)\, w_i(x)\, dx \end{cases}$$

Le calcul des coefficients de la matrice de masse et du second membre fait intervenir la détermination des intégrales :

$$\int_a^b w_j(x)\, w_i(x)\, dx$$

Le calcul de la matrice de raideur nécessite l'évaluation des intégrales :

$$\int_a^b w_j'(x)\, w_i'(x)\, dx$$

Ces calculs se font, comme précédemment, par assemblage des contributions de chaque élément $T_i = [x_i, x_{i+2}]$ pour i pair. On va maintenant présenter une technique commode de calcul des matrices et second membre élémentaires dans le cas d'éléments d'ordre élevé. Cette technique consiste à ramener par un changement de variable les calculs sur un élément quelconque à un calcul simplifié sur un élément de référence bien choisi.

7.6.2 Technique de l'élément de référence

Par le changement de variable

$$x = x_{i+1} + \frac{x_{i+2} - x_i}{2}\, t$$

on passe de $t \in [-1,1]$ à $x \in [x_i, x_{i+2}]$. Les fonctions de base dans $[x_i, x_{i+2}]$ se ramènent alors aux trois fonctions simples suivantes dans $[-1,1]$:

$$\lambda_{-1}(t) = \frac{t(t-1)}{2} \quad \lambda_0(t) = -(t-1)(t+1) \quad \lambda_1(t) = \frac{t(t+1)}{2}$$

dont les dérivées sont respectivement égales à :

$$\frac{d\lambda_{-1}}{dt} = t - \frac{1}{2} \quad \frac{d\lambda_0}{dt} = -2t \quad \frac{d\lambda_1}{dt} = t + \frac{1}{2}$$

7.6.3 Calcul de la matrice de masse élémentaire

Le calcul des coefficients de la matrice de masse se ramène, dans le cas α constante, à l'évaluation des intégrales

$$\int_{-1}^{1} \lambda_i(t)\, \lambda_j(t)\, dt \quad \text{pour} \quad i,j = -1,0,1$$

On obtient ainsi la matrice de masse élémentaire suivante pour l'élément $[x_i, x_{i+2}]$

$$M_i = \frac{x_{i+2} - x_i}{2} \begin{pmatrix} \frac{4}{15} & \frac{2}{15} & \frac{-1}{15} \\ \frac{2}{15} & \frac{16}{15} & \frac{2}{15} \\ \frac{-1}{15} & \frac{2}{15} & \frac{4}{15} \end{pmatrix}$$

7.6.4 Calcul de la matrice de raideur élémentaire

Le calcul de la matrice de raideur fait intervenir les dérivées des fonctions w_i par rapport à x. Après changement de variable, elles s'expriment comme produit des dérivées des λ par rapport à t par la dérivée de t par rapport à x.

$$\frac{dw_i}{dx} = \frac{d\lambda_k}{dt}\, \frac{dt}{dx}$$

si λ_k est dans $[-1,1]$ la fonction correspondant à la restriction de w_i dans $[x_i, x_{i+2}]$. On obtient ainsi :

$$K_i = \frac{2}{x_{i+2} - x_i} \begin{pmatrix} \frac{7}{6} & \frac{-4}{3} & \frac{1}{6} \\ \frac{-4}{3} & \frac{8}{3} & \frac{-4}{3} \\ \frac{1}{6} & \frac{-4}{3} & \frac{7}{6} \end{pmatrix}$$

7.6.5 Calcul du second membre élémentaire

Chaque composante F_i du vecteur second-membre global

$$F_i = \int_a^b f(x)\, w_i(x)\, dx$$

est calculée également par assemblage de contributions élémentaires $F_i^{(k)}$ selon :

$$F_i = \sum_{k=1}^{k=N} F_i^{(k)} = \sum_{k=1}^{k=N} \int_{x_{2k-2}}^{x_{2k}} f(x)\, w_i(x)\, dx$$

où les $F_i^{(k)}$ désignent les contributions des éléments k.

Tout dépend alors de la donnée de f. Si f est donnée analytiquement, on peut parfois calculer les intégrales exactement à la main. Mais en général, f n'est connue que par ses valeurs aux points x_i pour $i = 0, 2N$. On écrit donc f dans la base des w_i selon :

$$f(x) = \sum_{j=0}^{j=2N} f_j\, w_j(x)\, dx$$

On obtient :

$$F_i = \sum_{j=0}^{j=2N} f_j \Big(\sum_{k=1}^{k=N} \int_{x_{2k-2}}^{x_{2k}} w_j(x)\, w_i(x)\, dx \Big)$$

Le problème se ramène alors au calcul des intégrales :

$$\int_{x_{2k-2}}^{x_{2k}} w_j(x)\, w_i(x)\, dx$$

On retrouve des expressions calculées pour la matrice de masse. L'élément d'indice $k = i/2 + 1$: $[x_i, x_{i+2}]$ (pour i pair) produira ainsi une contribution non nulle aux trois composantes d'indices i, $i+1$, $i+2$ selon

$$\begin{pmatrix} F_i^{(k)} \\ F_{i+1}^{(k)} \\ F_{i+2}^{(k)} \end{pmatrix} = \frac{x_{i+2} - x_i}{2} \begin{pmatrix} \frac{4}{15} & \frac{2}{15} & \frac{-1}{15} \\ \frac{2}{15} & \frac{16}{15} & \frac{2}{15} \\ \frac{-1}{15} & \frac{2}{15} & \frac{4}{15} \end{pmatrix} \begin{pmatrix} f_i \\ f_{i+1} \\ f_{i+2} \end{pmatrix}$$

7.6.6 Intégration approchée. Condensation de masse

On verra plus loin qu'un calcul de l'erreur d'interpolation avec éléments finis $P2$ conduit à une majoration en h^2 selon

$$\|u - P_h u\|_{1,2} \leq C\, h^2 \max_{x\in[a,b]} |u'''(x)|$$

Il est par conséquent inutile de faire exactement tous les calculs d'intégrales. Ce qui est nécessaire, c'est

1) d'assurer l'ellipticité de la forme bilinéaire approchée après intégration numérique, donc l'existence et l'unicité de la solution approchée.

2) de conserver le même ordre global pour l'erreur. Donc de choisir une formule d'intégration numérique telle que l'erreur de quadrature soit du même ordre que l'erreur d'interpolation.

Dans le cas présent d'éléments $P2$, on choisira d'utiliser la formule de Simpson

$$\int_{x_i}^{x_{i+2}} \Phi(x)\, dx \approx \frac{x_{i+2} - x_i}{6}[\Phi(x_i) + 4\,\Phi(x_{i+1}) + \Phi(x_{i+2})]$$

qui est exacte pour les polynômes de degré inférieur ou égal à 3. Ceci donne un calcul exact pour les matrices de raideur élémentaires.

En ce qui concerne la matrice de masse élémentaire on obtient la forme approchée diagonale suivante :

$$M_i = \frac{x_{i+2} - x_i}{6} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Dans de nombreuses applications, il sera commode d'utiliser une matrice de masse diagonale (ou condensée).

On obtient de même une écriture plus simple du second membre pour l'élément d'indice $k = i/2 + 1$: $[x_i, x_{i+2}]$ (pour i pair) :

$$\begin{pmatrix} F_i^{(k)} \\ \\ F_{i+1}^{(k)} \\ \\ F_{i+2}^{(k)} \end{pmatrix} = \frac{x_{i+2} - x_i}{6} \begin{pmatrix} 1 & 0 & 0 \\ \\ 0 & 4 & 0 \\ \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} f_i \\ \\ f_{i+1} \\ \\ f_{i+2} \end{pmatrix}$$

7.6.7 Technique d'assemblage

Supposons un maillage en N éléments T_k pour $k = 1, 2..N$.

Notons A la matrice globale à assembler (matrice de raideur ou de masse globale, second membre) et a_k les matrices élémentaires correspondantes relatives à chaque élément T_k.

L'algorithme d'assemblage est très simple, dès lors que l'on dispose d'un tableau associant les points d'un élément T_k et les noeuds du maillage global.

Dans ce cas très simple d'éléments de degré deux en dimension un, chaque élément T_k comprend trois noeuds x_{2k-2}, x_{2k-1}, x_{2k}.

D'où l'algorithme :

POUR K = 1, N FAIRE ! boucle sur les éléments

POUR i = 1, 3 FAIRE ! boucle sur les numéros locaux
POUR j= 1, 3 FAIRE ! boucle sur les numéros locaux

I = 2*K + i - 3 ! numéros globaux
J = 2*K + j - 3 ! numéros globaux

A(I,J) = A(I,J) + a(i,j) ! A : matrice globale, a matrice élémentaire

FIN DES 3 BOUCLES

Remarque 7.6.1 *Les indices des coefficients de la matrice correspondent à la numérotation globale des inconnues adoptée. Ils vont donc de* 0 *à* $2N$.

7.7 Éléments finis de Lagrange de degré k ou éléments Pk

Comme dans les cas $P1$ et $P2$, considérons une discrétisation de l'intervalle $[a, b]$ en N sous-intervalles ou éléments T_i.

Choisissons comme espace V_h d'approximation, l'espace des fonctions continues sur $[a, b]$, et polynomiales de degré k sur chaque sous-intervalle. Un polynôme de degré k est fixé par ses valeurs en $k + 1$ points. On prend les extrémités et $k - 1$ points internes à chaque élément T_i. On est ainsi amené à considérer une discrétisation globale en $kN + 1$ points ou noeuds x_i pour $i = 0, ...kN$. L'erreur d'interpolation en norme H^1 est cette fois un infiniment petit en $O(h^k)$, si h est la mesure du plus grand des sous-intervalles.

Les calculs des matrices de raideur et de masse ainsi que le calcul du second membre se font comme précédemment et ne posent que des difficultés techniques.

Remarque 7.7.1 (importante) *Quel que soit le degré du polynôme d'interpolation utilisé dans chaque élément, les fonctions de V_h ne se raccordent, dans le cas d'éléments de Lagrange, que par leurs valeurs aux extrémités des éléments. On a donc uniquement la continuité des valeurs de la fonction inconnue et non celles de ses dérivées. On parle d'éléments de régularité C^0*

7.8 Éléments finis de Hermite cubiques ou éléments poutres

7.8.1 Problème de la poutre encastrée

Dans certains problèmes, la continuité C^0 ne suffit plus et on demande des solutions approchées plus régulières. Considérons le problème d'une poutre encastrée en ses extrémités et soumise à un chargement d'intensité f.

$$\left\{ \begin{array}{l} \dfrac{d^4}{dx^4}u(x) = f(x) \qquad \forall x \in [a,b] \\ u(a) = 0 \quad u'(a) = 0 \quad u(b) = 0 \quad u'(b) = 0 \end{array} \right. \tag{7.12}$$

La formulation variationnelle de ce problème s'écrit dans $H_0^2[a,b]$, espace des fonctions de carré sommable, de dérivées première et seconde de carré sommables dans $[a,b]$, nulles et à dérivées premières nulles aux points a et b.

On vérifiera qu'après 2 intégrations par parties successives, on obtient la formulation variationnelle suivante :

$$\left\{ \begin{array}{l} \text{Trouver la fonction } u \text{ appartenant à } H_0^2[a,b] \text{ telle que :} \\ \displaystyle\int_a^b u''(x)v''(x)dx = \int_a^b f(x)v(x)dx \quad \forall v \in H_0^2[a,b] \end{array} \right. \tag{7.13}$$

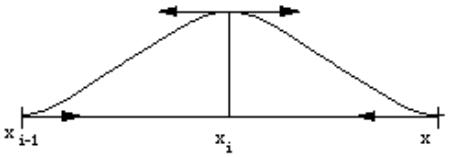
Les fonctions de $H^2[a,b]$ sont de classe C^1, on va donc construire un espace V_h d'approximation interne de fonctions C^1.

Discrétisons l'intervalle $[a,b]$ en N sous-intervalles ou éléments. On choisit pour V_h l'espace des fonctions continues à dérivées premières continues et polynomiales de degré trois par éléments.

Un polynôme de degré 3 est déterminé de manière unique par la donnée de ses valeurs et des valeurs de sa dérivée première aux deux points extrémités du sous-intervalle. Comme on a choisi les valeurs aux extrémités, le recollement global assurera la continuité C^1 des fonctions de V_h.

Chaque fonction de V_h est alors déterminée de manière unique par la donnée de ses valeurs et de celles de sa dérivée aux $N+1$ points x_i de la discrétisation. V_h est de dimension $2N+2$.

Si l'on considère maintenant le sous espace $V_{0,h} \subset H_0^2[a,b]$ des fonctions de V_h nulles et à dérivées nulles aux points a et b, il est de dimension $2N-2$, et est engendré par la double famille de fonctions de base w_i, θ_i définies par les $2N-2$ conditions suivantes :



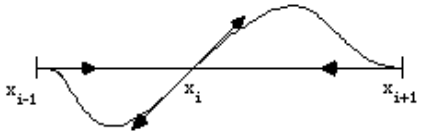


FIGURE 7.7 – Graphe d'une fonction w_i Graphe d'une fonction θ_i

$$w_i(x_j) = \delta_{ij} \quad w_i'(x_j) = 0 \quad \forall i = 1, N-1 \quad \text{et} \quad \forall j = 1, N-1$$

$$\theta_i(x_j) = 0 \quad \theta_i'(x_j) = \delta_{ij} \quad \forall i = 1, N-1 \quad \text{et} \quad \forall j = 1, N-1$$

Une fonction v_h quelconque de $V_{0,h}$ s'écrit dans cette base :

$$v_h(x) = \sum_{i=1}^{i=N-1} v_i\, w_i(x) \; + v_i'\, \theta_i(x)$$

avec $v_i = v_h(x_i)$ et $v_i' = v_h'(x_i)$.

7.8.2 Écriture du problème approché

Ecrivons le problème approché dans $V_{0,h}$

$$\begin{cases} \text{Trouver la fonction } u_h \text{ appartenant à } V_{0,h} \text{ telle que} \\ \displaystyle\int_a^b u_h''(x) v_h''(x) dx \; = \int_a^b f(x) v_h(x) dx \quad \forall v_h \in V_{0,h} \end{cases}$$

Le problème étant linéaire, l'égalité est vraie pour tout v_h si elle est vraie pour une base de l'espace vectoriel $V_{0,h}$

$$\forall v_h \in V_{0,h} \Longleftrightarrow \forall w_i \quad \text{et} \quad \forall \theta_i \quad \text{pour} \quad i = 1, N-1$$

D'autre part, écrivons u_h, solution du problème approché dans $V_{0,h}$, dans la base des w_i, θ_i

$$u_h(x) = \sum_{j=1}^{j=N-1} \left(u_j \, w_j(x) \, + \, u'_j \, \theta_j(x) \right)$$

avec $u_j = u_h(x_j)$ et $u'_j = u'_h(x_j)$ valeurs de la solution approchée et de sa dérivée au point (x_j)

On obtient l'écriture suivante du problème approché :

$$\begin{cases} \text{Trouver } u_1, \; u'_1, \; u_2, \; u'_2, ...u_{N-1}, \; u'_{N-1} \text{ tels que} \quad \forall i = 1, N-1 \\ \displaystyle\sum_{j=1}^{j=N-1} \left(\int_a^b w''_j(x) w''_i(x) dx \right) u_j + \int_a^b \theta''_j(x) w''_i(x) dx \, u'_j = \int_a^b f(x) w_i(x) dx \\ \displaystyle\sum_{j=1}^{j=N-1} \left(\int_a^b w''_j(x) \theta''_i(x) dx \right) u_j + \left(\int_a^b \theta''_j(x) \theta''_i(x) dx \right) u'_j = \int_a^b f(x) \theta_i(x) dx \end{cases}$$

En posant pour i et $j = 1, N-1$

$$\int_a^b f(x) \, w_i(x) \, dx = F_{2i-1} \qquad \int_a^b f(x) \, \theta_i(x) \, dx = F_{2i}$$

et

$$\int_a^b w''_j(x) \, w''_i(x) \, dx \; = A_{2i-1,2j-1} \qquad \int_a^b \theta''_j(x) \, \theta''_i(x) \, dx \; = A_{2i,2j}$$

$$\int_a^b w''_j(x) \, \theta''_i(x) \, dx \; = A_{2i,2j-1} \qquad \int_a^b \theta''_j(x) \, w''_i(x) \, dx \; = A_{2i-1,2j}$$

ceci donne

$$\sum_{J=1}^{J=2N-2} A_{I,J} \, U_J = F_I \qquad \forall I = 1, 2N-2$$

On a ainsi obtenu un système linéaire de $2N-2$ équations à $2N-2$ inconnues, qui peut s'écrire sous la forme matricielle suivante :

$$A \; U \; = \; F$$

7.8.3 Calcul des matrices et second membre élémentaires pour l'élément de Hermite cubique

En adoptant la technique de l'élément de référence, on se ramène sur l'intervalle $[0, 1]$. Par le changement de variable

$$x = x_i + (x_{i+1} - x_i) \, t$$

on passe de $t \in [0,1]$ à $x \in [x_i, x_{i+1}]$. Les restrictions des fonctions de base w_j, θ_j dans un élément quelconque $[x_i, x_{i+1}]$ tel que $x_{i+1} - x_i = h_i$ se ramènent alors sur $[0,1]$ aux 4 fonctions simples λ_0, λ_1, μ_0, μ_1 suivantes :

$$\lambda_0(t) = 2t^3 - 3t^2 + 1 \quad \lambda_1(t) = 3t^2 - 2t^3$$

et après avoir remarqué que

$$\frac{d\theta_i}{dx} = \frac{d\mu_k}{dt}\,\frac{1}{h_i}$$

si μ_k est dans $[0,1]$ la fonction correspondant à la restriction de θ_i dans $[x_i, x_{i+1}]$

$$\mu_0(t) = h_i\, t(t-1)^2 \quad \mu_1(t) = h_i\,(t^3 - t^2)$$

Leurs dérivées secondes sont respectivement

$$\lambda_0''(t) = 12t - 6 \quad \lambda_1''(t) = 6 - 12t$$

$$\mu_0''(t) = h_i\,(6t - 4) \quad \mu_1''(t) = h_i\,(6t - 2)$$

D'où,

$$\frac{d^2 w_i}{dx^2} = \frac{d^2\lambda_k}{dt^2}\,\frac{1}{h_i^2}$$

si λ_k est dans $[0,1]$ la fonction correspondant à la restriction de w_i dans $[x_i, x_{i+1}]$, et de même

$$\frac{d^2\theta_i}{dx^2} = \frac{d^2\mu_k}{dt^2}\,\frac{1}{h_i^2}$$

si μ_k est dans $[0,1]$ la fonction correspondant à la restriction de θ_i dans $[x_i, x_{i+1}]$. On obtient en définitive l'expression suivante de la matrice élémentaire relative à un élément $[x_i, x_{i+1}]$ tel que $x_{i+1} - x_i = h_i$

$$K_i = \frac{1}{h_i^3}\begin{pmatrix} 12 & 6h_i & -12 & 6h_i \\ 6h_i & 4h_i^2 & -6h_i & 2h_i^2 \\ -12 & -6h_i & 12 & -6h_i \\ 6h_i & 2h_i^2 & -6h_i & 4h_i^2 \end{pmatrix}$$

Chapitre 8

Éléments finis bidimensionnels

8.1 Rappel de la formulation générale abstraite

Tous les problèmes bidimensionnels du chapitre 5, comme les problèmes monodimensionnels résolus par éléments finis lors du chapitre précédent, peuvent s'écrire sous la forme générale du problème P suivant :

$$(P)\begin{cases} \text{Trouver la fonction } u \text{ appartenant à l'Hilbert } V \text{ telle que :} \\ a(u,v) = l(v) \quad \forall v \in V \end{cases} \tag{8.1}$$

avec V un espace de Hilbert réel muni du produit scalaire $(\ .\ ,\ .\ )$ et de la norme associée $\|\ .\ \|$, a une forme bilinéaire continue et elliptique sur V et l une forme linéaire continue sur V. Le théorème de Lax-Milgram conduit alors aux conclusions suivantes :

1) Le problème P admet une solution unique u dans V
2) Si la forme bilinéaire a est symétrique, le problème variationnel (P) est équivalent au problème de minimisation suivant :

$$\begin{cases} \text{Trouver } u \in V \text{ qui réalise le minimum de la forme quadratique} \\ J(v) = \frac{1}{2}\, a(v,v) - l(v) \end{cases} \tag{8.2}$$

8.2 Approximation interne du problème

On doit maintenant construire un sous-espace $V_h \subset V$ de dimension finie dans lequel s'écrira le problème approché P_h :

$$(P_h)\begin{cases} \text{Trouver la fonction } u_h \text{ appartenant à l'espace } V_h \text{ telle que :} \\ a(u_h,v_h) = l(v_h) \quad \forall v_h \in V_h \end{cases} \tag{8.3}$$

Ce problème admet également une solution unique car $V_h \subset V$, et donc les hypothèses du théorème de Lax-Milgram sont également vérifiées dans V_h.

Si la forme bilinéaire a est symétrique le problème variationnel P_h est équivalent au problème de minimisation suivant :

$$\left\{ \begin{array}{l} \text{Trouver } u_h \in V_h \text{ qui réalise le minimum de la forme quadratique} \\ \\ \qquad\qquad J(v_h) = \frac{1}{2}\, a(v_h, v_h) \;-\; l(v_h) \end{array} \right. \tag{8.4}$$

et on a donc évidemment

$$J(u_h) \;\geq\; J(u)$$

8.3 Maillage

Dans la méthode des éléments finis, la construction du sous-espace V_h nécessite la discrétisation préalable du domaine Ω en éléments géométriques simples.

En dimension un, la discrétisation préalable du domaine, un intervalle de $\mathbb{R}$, en éléments, ne posait pas de difficultés. En dimension deux et plus encore en dimension trois, la discrétisation du domaine Ω est un problème technique difficile. La qualité du maillage est cruciale pour la qualité de l'approximation. Le problème de la réalisation du maillage se pose à la fois en amont de la résolution numérique, qui s'appuie sur une discrétisation a priori, et en aval dans les techniques de maillages adaptatifs par lesquelles on s'efforce d'améliorer la qualité de la discrétisation en fonction des résultats obtenus (voir chapitre 18).

L'exposé des techniques mises en oeuvre pour construire un maillage en éléments finis dépasse le cadre de ce cours et nous renvoyons à la littérature (notamment P.L.George : Génération automatique de maillages.).

Disons simplement que l'on peut distinguer deux types de maillages :

- Les maillages structurés qui fournissent une discrétisation "régulière" obtenue par transformation d'une grille régulière sur un domaine rectangulaire.
- Les maillages non-structurés, principalement construits par la méthode de Delaunay-Voronoi, qui peuvent s'appliquer aux géométries les plus générales.

En dimension deux, les éléments sont des triangles ou des quadrangles de côtés droits ou curvilignes. En dimension trois, ce sont des tétraèdres, pentaèdres ou hexaèdres.

Un maillage en éléments finis doit satisfaire les critères suivants :

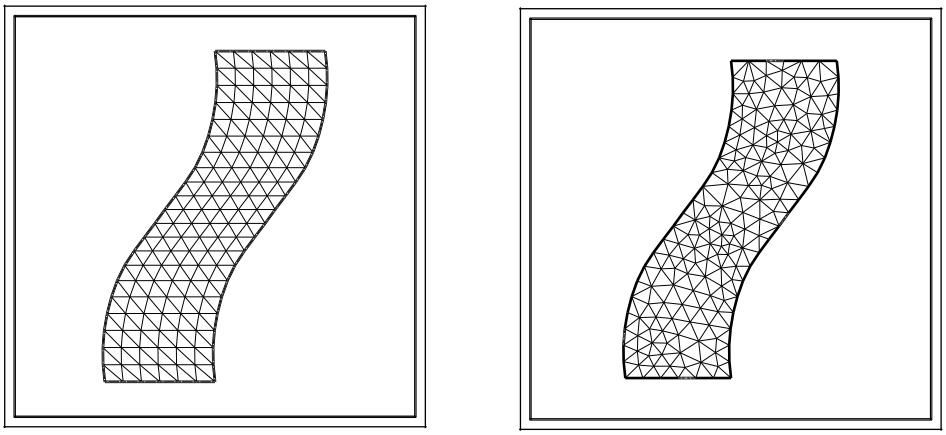


FIGURE 8.1 – Maillage structuré et maillage non-structuré

1) Les éléments K_i du maillage doivent recouvrir le domaine Ω

$$\bigcup_i K_i = \bar{\Omega}$$

Ceci implique que, par exemple en dimension deux, ce domaine soit polygonal ou approché par un polygone si l'on utilise des éléments droits.

2) L'intersection de deux éléments distincts ne peut être que
- l'ensemble vide
- un sommet
- un côté
- une face (en dimension trois)

Ceci a pour but d'assurer la continuité des fonctions de V_h, on parle alors d'éléments **conformes**. En particulier la disposition présentée dans la figure 8.2 est interdite.

8.4 Éléments finis de Lagrange triangulaires de degré un : les éléments finis P1

Dans ce cas l'espace d'approximation V_h est un espace de fonctions continues affines par éléments triangulaires. Dans chaque triangle, la restriction des fonc-

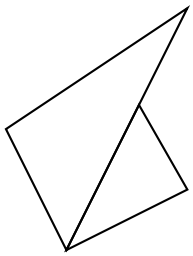


FIGURE 8.2 – Configuration interdite

tions de V_h est donc un polynôme de degré un de la forme $a_0 + a_1\,x + a_2\,y$ qui est donc déterminé de façon unique par ses valeurs en trois points distincts. On choisit les trois sommets du triangle. Ceci assure la continuité globale des fonctions de V_h sur le domaine polygonal Ω. En effet sur chaque arête commune à deux triangles adjacents, les restrictions des fonctions de V_h sont des fonctions affines fixées de manière unique par la donnée de leurs valeurs aux deux sommets sur l'arête. Globalement les fonctions de V_h seront uniquement déterminées par

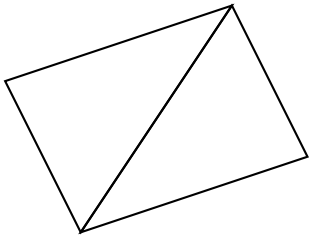


FIGURE 8.3 – Deux éléments adjacents

leurs valeurs aux sommets ou noeuds de la triangulation. **La dimension totale de V_h est donc égale au nombre de noeuds du maillage.** Dans le cas de conditions aux limites de Dirichlet sur une partie de la frontière, la dimension de V_h est évidemment réduite au nombre des noeuds associés à une valeur inconnue de la solution. Elle est donc égale au nombre de noeuds total moins le nombre de noeuds où la solution est fixée par une condition de Dirichlet.

8.4.1 Les fonctions de base P1

La construction d'une base de V_h se fait selon la technique classique de Lagrange. On prend comme fonctions de base les N fonctions w_I de V_h définies par les N conditions suivantes aux N noeuds (x_I, y_I) du maillage :

$$w_I(x_J, y_J) = \delta_{IJ}$$

On remarquera que ses fonctions ont un support réduit à l'union des triangles dont le point (x_I, y_I) est un sommet. Dans cette base une fonction de V_h s'écrit :

$$v_h(x,y) = \sum_I v_I \, w_I(x,y)$$

FIGURE 8.4 – Une fonction de base P1

8.4.2 Les fonctions de forme P1

On appellera fonctions de forme les restrictions des fonctions de base dans un élément. Dans chaque triangle T de sommets A_1, A_2, A_3, il n'y a que 3 fonctions de base non-nulles. Les restrictions de ces fonctions de base $w_{I_1}, w_{I_2}, w_{I_3}$ sont les trois fonctions polynomiales de degré un prenant la valeur 1 en un des sommets et nulles aux deux autres sommets. Notons les respectivement $\lambda_1, \lambda_2, \lambda_3$.

λ_1 est le polynôme de degré un prenant la valeur 1 en A_1 et nul en A_2 et A_3.

$$\lambda_1(x,y) = a_0 + a_1 x + a_2 y$$

λ_1 est donc déterminé par le système linéaire suivant :

$$\begin{cases} \lambda_1(x_1,y_1) = a_0 + a_1 x_1 + a_2 y_1 = 1 \\ \lambda_1(x_2,y_2) = a_0 + a_1 x_2 + a_2 y_2 = 0 \\ \lambda_1(x_3,y_3) = a_0 + a_1 x_3 + a_2 y_3 = 0 \end{cases}$$

Le déterminant de ce système

$$\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} = 2.Aire(T)$$

est différent de zéro si les points A_1, A_2, A_3 ne sont pas alignés.

En résolvant le système 3×3 ci-dessus et les systèmes analogues pour λ_2 et λ_3 on obtient les formules suivantes :

$$\lambda_1(x,y) = \frac{x_2y_3 - x_3y_2 + x(y_2 - y_3) + y(x_3 - x_2)}{2\,Aire\,(T)} \tag{8.5}$$

$$\lambda_2(x,y) = \frac{x_3y_1 - x_1y_3 + x(y_3 - y_1) + y(x_1 - x_3)}{2\,Aire\,(T)} \tag{8.6}$$

$$\lambda_3(x,y) = \frac{x_1y_2 - x_2y_1 + x(y_1 - y_2) + y(x_2 - x_1)}{2\,Aire\,(T)} \tag{8.7}$$

On en déduit les expressions suivantes des gradients :

$$\frac{\partial\lambda_1}{\partial x} = \frac{y_2 - y_3}{2\,Aire\,(T)} \qquad \frac{\partial\lambda_1}{\partial y} = \frac{x_3 - x_2}{2\,Aire\,(T)} \tag{8.8}$$

$$\frac{\partial\lambda_2}{\partial x} = \frac{y_3 - y_1}{2\,Aire\,(T)} \qquad \frac{\partial\lambda_2}{\partial y} = \frac{x_1 - x_3}{2\,Aire\,(T)} \tag{8.9}$$

$$\frac{\partial\lambda_3}{\partial x} = \frac{y_1 - y_2}{2\,Aire\,(T)} \qquad \frac{\partial\lambda_3}{\partial y} = \frac{x_2 - x_1}{2\,Aire\,(T)} \tag{8.10}$$

Les trois fonctions $\lambda_1, \lambda_2, \lambda_3$ s'appellent les coordonnées barycentriques du triangle T. On les désigne sous le nom "area coordinates" en anglais car elles représentent en chaque point M de coordonnées x, y le rapport des aires algébriques des triangles MA_iA_j et T . Par exemple

$$\lambda_1(M) = \frac{Aire(A_2A_3M)}{Aire(T)}$$

FIGURE 8.5 – Deux représentations de la fonction λ_1

Une fonction quelconque de V_h prenant les valeurs v_1, v_2, v_3 aux sommets A_1, A_2, A_3 du triangle T s'écrit dans T sous la forme

$$v_h(x,y) = v_1\lambda_1(x,y) + v_2\lambda_2(x,y) + v_3\lambda_3(x,y)$$

8.5 Application à un problème elliptique modèle

Soit Ω un domaine borné polygonal de $\mathbb{R}^2$ de frontière Γ. Supposons Γ constituée de deux parties Γ_0 et Γ_1, $\Gamma = \Gamma_0 \cup \Gamma_1$.

Sur Γ_0 sont imposées des conditions de Dirichlet.
Sur Γ_1 sont imposées des conditions de Neumann.

Nous obtenons le problème mixte suivant :

$$\begin{cases} -\Delta\, u\,(x,y) = f\,(x,y) \qquad \forall\ x,y \in \Omega \\ u_{\,|\,\Gamma_0} = u_d \\ \dfrac{\partial u}{\partial n}_{\,|\,\Gamma_1} = g \end{cases} \tag{8.11}$$

8.5.1 Formulation variationnelle de ce problème

La formulation du problème s'écrit dans l'espace V des fonctions de $H^1(\Omega)$ nulles sur la partie Γ_0 de la frontière. On se ramène aux conditions de Dirichlet homogènes sur Γ_0 en introduisant une fonction auxiliaire simple prenant les valeurs imposées sur la frontière Γ_0 : $u_{0_{|\Gamma_0}} = u_d$ et en posant $u = \tilde{u} + u_0$. On obtient le problème variationnel :

$$\begin{cases} \text{Trouver } \tilde{u} \in V \text{ telle que : } \quad \forall\, v \in V \\ \displaystyle\iint_\Omega \mathbf{grad}\tilde{u}\ \mathbf{grad}v\, dxdy = \iint_\Omega f v\, dxdy - \iint_\Omega \mathbf{grad}u_0\, \mathbf{grad}v\, dxdy + \int_{\Gamma_1} g\, v\, d\gamma \end{cases}$$

Ce problème s'écrit sous la forme générale standard :

$$\begin{cases} \text{Trouver } \tilde{u} \in V \text{ telle que :} \\ a(\tilde{u}, v) = l(v) \quad \forall\, v \in V \end{cases}$$

avec $a : u, v \longrightarrow a(u,v) = \displaystyle\iint_\Omega \mathbf{grad}u\ \mathbf{grad}v\, dxdy$: forme bilinéaire symétrique sur V
$l : v \longrightarrow l(v) = \displaystyle\iint_\Omega f\, v\, dxdy - \iint_\Omega \mathbf{grad}u_0\, \mathbf{grad}v\, dxdy + \int_{\Gamma_1} g\, v\, d\gamma$: forme linéaire sur V.

8.5.2 Écriture du problème approché en éléments finis P1

Le problème approché s'écrit dans l'espace V_h des fonctions continues affines par triangles et nulles sur la partie Γ_0 de la frontière. Notons $\mathbf{I}$ l'ensemble des indices des noeuds du maillage correspondants à une valeur inconnue de la solution u. C'est à dire ici l'ensemble des noeuds n'appartenant pas à Γ_0. Notons $\mathbf{J}$ l'ensemble des indices des sommets du maillage appartenant à Γ_0.

La solution $\tilde{u}_h$ s'écrira dans la base des w_J pour $J \in \mathbf{I}$ selon :

$$\tilde{u}_h(x,y) = \sum_{J \in \mathbf{I}} u_J \, w_J(x,y)$$

La fonction auxiliaire u_0 sera approchée par une fonction $u_{0,h}$ continue et affine par morceaux prenant les valeurs imposées sur Γ_0 et nulle sur tous les noeuds d'indices $J \in \mathbf{I}$

$$u_{0,h}(x,y) = \sum_{J \in \mathbf{J}} u_d(x_J, y_J) \; w_J(x,y)$$

Ce choix de $u_{0,h}$ présente deux avantages :

1) La solution cherchée u_h et la solution calculée $\tilde{u}_h$ prennent les mêmes valeurs aux points où la solution u est inconnue. Il n'y a donc pas de transformation a posteriori à effectuer sur les résultats.

2) Les conditions aux limites ne produisent qu'une modification limitée du système linéaire qui n'intervient que sur quelques composantes du second-membre.

En intégrant dans la formulation variationnelle tous ces éléments on obtient en définitive le problème approché dans V_h

$$\left\{ \begin{array}{l} \text{Trouver les valeurs } u_J \text{ pour } J \in \mathbf{I} \text{ telles que :} \\ \sum_{J \in \mathbf{I}} \Big(\iint_\Omega \mathbf{grad} w_J \; \mathbf{grad} w_I \, dxdy \Big) \, u_J = \iint_\Omega f \, w_I \, dxdy \; + \int_{\Gamma_1} g \, w_I \, d\gamma \\ - \sum_{J \in \mathbf{J}} \Big(\iint_\Omega \mathbf{grad} w_J \, \mathbf{grad} w_I \, dxdy \Big) \, u_d(x_J, y_J) \quad \forall \; I \in \mathbf{I} \end{array} \right.$$

On obtient un système de $N_{\mathbf{I}}$ équations à $N_{\mathbf{I}}$ inconnues où $N_{\mathbf{I}}$ désigne le nombre de points du maillage d'indices $I \in \mathbf{I}$ donc le nombre de noeuds correspondant à des valeurs inconnues de la solution. Ce système s'écrit sous la forme matricielle

$$K \; U \; = \; F$$

où K est la matrice de raideur de coefficients

$$K_{I,J} = \iint_\Omega \mathbf{grad} w_J \; \mathbf{grad} w_I \, dxdy$$

et F le vecteur second-membre de composantes :

$$F_I = \iint_\Omega f\, w_I\, dxdy + \int_{\Gamma_1} g\, w_I\, d\gamma - \sum_{J \in \mathbf{J}} \Big(\iint_\Omega \mathbf{grad} w_J\ \mathbf{grad} w_I\, dxdy \Big)\ u_d(x_J, y_J)$$

où l'on reconnaît :

- un premier terme représentant les efforts surfaciques (correspondant au second-membre du problème différentiel).
- un deuxième terme, intégrale curviligne, provenant des conditions aux limites de Neumann.
- un dernier terme, expression des conditions de Dirichlet non-homogènes.

Le calcul des coefficients $K_{I,J}$ de la matrice de raideur et des composantes F_I du second-membre se fait par une procédure d'assemblage des contributions apportées par chacun des éléments T_k de la triangulation.

Par exemple pour la matrice de raideur K :

$$K_{I,J} = \iint_\Omega \mathbf{grad} w_J\ \mathbf{grad} w_I\, dxdy = \sum_k \iint_{T_k} \mathbf{grad} w_J\ \mathbf{grad} w_I\, dxdy \qquad (8.12)$$

On observe que la matrice K est très "creuse", un grand nombre de ses coefficients sont nuls, en raison du choix de fonctions w_I de support limité.

8.5.3 Calcul de la matrice de raideur élémentaire P1

Un des outils de base essentiels à la programmation de la méthode des éléments finis est un tableau de correspondances entre les noeuds X_I du maillage global et les points d'un élément particulier : ici les sommets A_1, A_2, A_3 des éléments triangulaires.

Dans chaque élément triangulaire T_k de sommets A_1, A_2, A_3 correspondants aux noeuds X_I, X_J, X_K, les seules fonctions de base non nulles sont les fonctions w_I, w_J, w_K. Leurs restrictions dans le triangle sont respectivement les trois coordonnées barycentriques $\lambda_1, \lambda_2, \lambda_3$ calculées précédemment. La matrice élémentaire relative au triangle T_k est donc une matrice 3×3 de coefficients :

$$elemK_{i,j} = \iint_{T_k} \mathbf{grad}\lambda_j\ \mathbf{grad}\lambda_i\, dxdy \quad \forall\, i, j = 1, 2, 3$$

Comme, dans ce cas particulier simple d'éléments $P1$, les gradients sont constants par triangles, les intégrales à calculer sont des intégrales de fonctions constantes. Il suffit d'en multiplier la valeur par l'aire de l'élément.

On obtient ainsi la matrice élémentaire $P1$ de coefficients :

$$\begin{pmatrix} \frac{(y_2-y_3)^2+(x_2-x_3)^2}{4\,Aire(T)} & \frac{(y_2-y_3)(y_3-y_1)+(x_2-x_3)(x_3-x_1)}{4\,Aire(T)} & \frac{(y_2-y_3)(y_1-y_2)+(x_2-x_3)(x_1-x_2)}{4\,Aire(T)} \\ \frac{(y_2-y_3)(y_3-y_1)+(x_2-x_3)(x_3-x_1)}{4\,Aire(T)} & \frac{(y_3-y_1)^2+(x_3-x_1)^2}{4\,Aire(T)} & \frac{(y_3-y_1)(y_1-y_2)+(x_3-x_1)(x_1-x_2)}{4\,Aire(T)} \\ \frac{(y_2-y_3)(y_1-y_2)+(x_2-x_3)(x_1-x_2)}{4\,Aire(T)} & \frac{(y_3-y_1)(y_1-y_2)+(x_3-x_1)(x_1-x_2)}{4\,Aire(T)} & \frac{(y_1-y_2)^2+(x_1-x_2)^2}{4\,Aire(T)} \end{pmatrix}$$

En particulier, dans le cas du triangle rectangle isocèle de sommets

$$A_1 = (0,0), A_2 = (1,0), A_3 = (0,1)$$

souvent utilisé comme triangle de référence (voir plus loin), on a la matrice élémentaire de raideur suivante :

$$elemK = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

8.5.4 Calcul des seconds membres élémentaires

Le second membre se compose de 3 termes :

1) un premier terme surfacique

$$Fs_I = \iint_\Omega f\, w_I\, dxdy \qquad \forall I$$

son calcul s'effectue en sommant les contributions de chaque élément triangulaire.

$$Fs_I = \sum_k Fs_{I,T_k} = \sum_k \iint_{T_k} f\, w_I\, dxdy$$

Si f est connue analytiquement et suffisamment simple on peut calculer exactement, à la main, les intégrales. Mais en général, le second membre f est connu par ses valeurs aux noeuds. On le représente dans la base des w_I et on est ramené aux calculs suivants

$$Fs_I = \iint_\Omega \sum_J f_J\, w_J\, w_I\, dxdy \qquad \forall I$$

Ce qui ramène aux calculs de

$$\iint_{\Omega} w_J \, w_I \, dxdy \qquad \forall I, J$$

Dans chaque élément on doit donc calculer la matrice de masse élémentaire.

Matrices élémentaires de masse

On obtient la matrice élémentaire de masse, en utilisant les formules d'intégration exactes suivantes :

$$\iint_T \lambda_i \, dxdy = \frac{Aire(T)}{3}$$

$$\iint_T \lambda_i^2 \, dxdy = \frac{Aire(T)}{6}$$

$$\iint_T \lambda_j \, \lambda_i \, dxdy = \frac{Aire(T)}{12} \qquad \text{si } i \text{ différent de } j$$

Ce qui donne le résultat :

$$elemM_{T_k} = Aire(T_k) \begin{pmatrix} \frac{1}{6} & \frac{1}{12} & \frac{1}{12} \\ ... & \frac{1}{6} & \frac{1}{12} \\ ... & ... & \frac{1}{6} \end{pmatrix}$$

Matrice de masse condensée (mass lumping)

Il peut être avantageux de calculer la matrice de masse élémentaire de manière approchée en utilisant la formule suivante (exacte sur P1) :

$$\iint_T \phi \, dxdy = \frac{Aire(T)}{3} \sum_{i=1}^{3} \phi(A_i)$$

On obtient dans ce cas une matrice de masse diagonale égale à :

$$elemM_{T_k} = \frac{Aire(T_k)}{3} \, I$$

On en déduit le second membre élémentaire surfacique :

$$\begin{pmatrix} Fs_{I,T_k} \\ Fs_{J,T_k} \\ Fs_{K,T_k} \end{pmatrix} = elemM_{T_k} \begin{pmatrix} f_I \\ f_J \\ f_K \end{pmatrix}$$

2) un terme de bord provenant des conditions de Neumann

$$Fn_I = \int_{\Gamma_1} g\, w_I\, d\gamma$$

Son calcul s'effectue sur les éléments ayant un côté sur Γ_1. Il se ramène à une intégrale simple sur un côté A d'un triangle. Si la fonction g est donnée par ses valeurs aux noeuds du maillage et si A a pour extrémités X_I et X_J, on a sur A : $g = g_I w_I + g_J w_J$ et on doit donc calculer :

$$g_J \int_A w_J w_I\, d\gamma$$

$$g_J \int_A w_J^2\, d\gamma$$

et

$$g_I \int_A w_I^2\, d\gamma$$

Ces calculs s'effectuent exactement par la formule de Simpson :

$$g_J \int_A w_J\, w_I\, d\gamma = \frac{Longueur(A)}{6}\, g_J$$

$$g_J \int_A w_J^2\, d\gamma = \frac{Longueur(A)}{3}\, g_J$$

etc..

FIGURE 8.6 – Triangle frontière

Dans le cas de la figure (8.6), on obtient :

$$\begin{pmatrix} Fn_{T,I} \\ Fn_{T,J} \\ Fn_{T,K} \end{pmatrix} = \frac{Longueur(A)}{6} \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} g_I \\ g_J \\ g_K \end{pmatrix}$$

On peut également calculer ce terme de façon approchée par la formule des trapèzes ; Ce qui donnerait :

$$\begin{pmatrix} Fn_{T,I} \\ Fn_{T,J} \\ Fn_{T,K} \end{pmatrix} = \frac{Longueur(A)}{2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} g_I \\ g_J \\ g_K \end{pmatrix}$$

Dans tous les cas le terme provenant des conditions de Neuman n'induit de contributions non nulles que pour les composantes du second membre relatives à des noeuds du maillage situés sur Γ_1.

3) un troisième terme provenant des conditions de Dirichlet non-homogènes

$$Fd_I = -\sum_{J \in \mathbf{J}} \Big(\iint_\Omega \mathbf{grad} w_J \ \mathbf{grad} w_I \, dxdy \Big) \ u_d(x_J, y_J)$$

Son calcul se ramène encore à une somme de contributions des triangles T.
Pour chaque triangle dont des points d'indices $J \in \mathbf{J}$ et le point I sont sommets, on obtient une contribution à la I^{eme} composante du second membre égale à :

$$\sum_{J \in \mathbf{J}} \Big(\iint_T \mathbf{grad} \lambda_j \ \mathbf{grad} \lambda_i \, dxdy \Big) \ u_d(x_J, y_J)$$

On retrouve des coefficients déjà calculés pour la matrice de raideur. Pour un triangle I, J, K comme celui de la figure (8.6) on obtient une contribution unique à la composante K :

$$Fd_{T,K} = (a_{i,k} \quad , a_{j,k} \quad , a_{k,k}) \begin{pmatrix} ud_I \\ ud_J \\ 0 \end{pmatrix}$$

où les i, j, k sont les numéros internes au triangle T correspondants aux indices globaux I, J, K.

8.5.5 Algorithme d'assemblage

Supposons un maillage en N éléments T_k pour $k = 1, 2..N$. Notons A la matrice globale à assembler (matrice de raideur ou de masse globale, second membre) et a_k les matrices élémentaires correspondantes relatives à chaque élément T_k. L'algorithme d'assemblage est très simple, dès lors que l'on dispose d'un tableau *numero* associant les sommets d'un élément T_k et les noeuds du maillage global. Dans ce cas simple d'éléments triangulaires P1, chaque élément T_k comprend trois

noeuds X_I, X_J, X_K correspondants aux sommets A_1, A_2, A_3 du triangle T_k. D'où l'algorithme :

POUR k = 1, N FAIRE ! boucle sur les éléments

POUR i = 1, 3 FAIRE ! boucle sur les numéros locaux
I = numero(k,i) ! numéros globaux
POUR j = 1, 3 FAIRE
J = numero(k,j)
A(I,J) = A(I,J) + a_k(i,j) ! A : matrice globale, a matrice élémentaire

FIN DES 3 BOUCLES

8.6 Éléments triangulaires généraux

Considérons un maillage du domaine Ω de $I\!R^2$ en triangles T_l pour $l = 1, ..Nbe$, satisfaisant aux critères énoncés précédemment. Le choix d'une approximation par éléments finis Pk correspond au choix d'un espace V_h d'approximation constitué de fonctions continues sur $\bar{\Omega}$, polynomiales de degré k par éléments triangulaires. Notons que ce choix, associé à une triangulation correcte assure l'inclusion $V_h \subset H^1(\Omega)$.

Un polynôme de degré k à deux variables a $\dfrac{(k+1)(k+2)}{2}$ coefficients. On devra donc imposer $\dfrac{(k+1)(k+2)}{2}$ conditions dans chaque élément triangulaire pour y fixer les restrictions des fonctions de V_h. Soit $\dfrac{(k+1)(k+2)}{2}$ paramètres ou degré de liberté locaux. Dans le cas d'éléments de Lagrange, ces degrés de liberté sont exclusivement les valeurs de la fonction en certains points du triangle.

Pour assurer la continuité globale sur $\bar{\Omega}$ des fonctions de V_h, il faut (et cela suffit) assurer le raccordement aux frontières entre éléments triangulaires. La restriction des polynômes Pk à deux variables sur les côtés des triangles est un polynôme de degré k à une seule variable. Il faut $k+1$ conditions pour le fixer de manière unique. Donc pour assurer le raccordement des fonctions de V_h entre éléments adjacents, il faut disposer, dans le cas d'éléments de Lagrange, $k+1$ points sur leur arête commune.

En définitive les éléments triangulaires de Lagrange de degré k comprennent $\dfrac{(k+1)(k+2)}{2}$ points par triangles, dont $k+1$ points sur chaque côté.

8.6.1 Éléments P1

L'élément triangle P1 comporte 3 degrés de liberté, donc 3 points nodaux, et 2 points par côtés. Il y a une seule possibilité : le choix des 3 sommets du triangle.

8.6.2 Éléments P2

L'élément triangle P2 comporte 6 degrés de liberté, donc 6 points nodaux, 3 points par côtés. On choisit les 3 sommets et les 3 milieux des côtés du triangle.

8.6.3 Éléments P3

L'élément triangle P3 comporte 10 degrés de liberté, donc 10 points nodaux, 4 points par côtés. On choisit les 3 sommets du triangle, plus 2 points au un tiers, deux tiers de chaque côté et le barycentre.

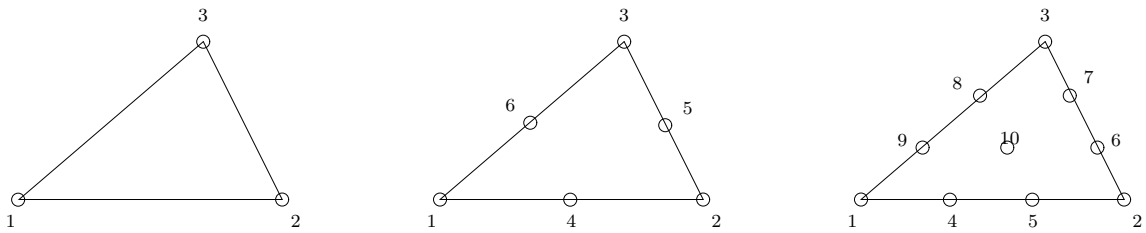


FIGURE 8.7 – Eléments triangulaires : P1, P2, P3.

8.7 Fonctions de base Pk

Globalement sur $\bar{\Omega}$ les fonctions de V_h sont définies par leurs valeurs aux noeuds de la triangulation, c'est à dire sur l'ensemble des points constituant l'union des ensembles de points nodaux élémentaires.

Par exemple, dans le cas d'éléments P2, les noeuds X_I du maillage seront les sommets des triangles et les milieux des côtés des arêtes. La base globale de Lagrange sera la base des fonctions w_I définies par les conditions :

$$w_J(x_I, y_I) = \delta_{I,J}$$

pour tout indice I et tout indice J correspondants à un noeud de la triangulation.

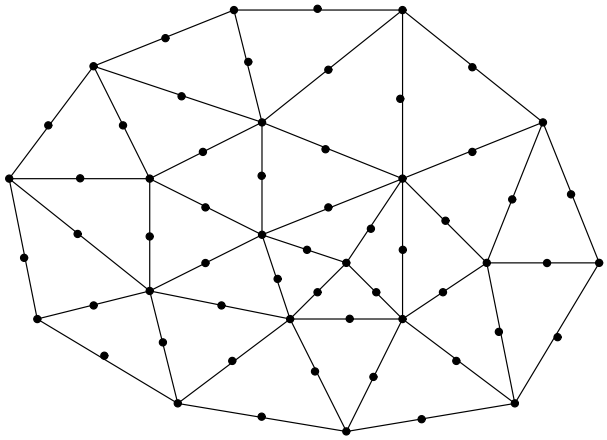


FIGURE 8.8 – Maillage en triangles P2 droits

8.8 Fonctions de forme Pk

Ce sont les restrictions des fonctions de base w_I dans un élément triangulaire. Notons les N_i pour $i = 1, ..n_k$ avec $n_k = \dfrac{(k+1)(k+2)}{2}$.

On les définit localement selon la technique de Lagrange :

$$N_i(A_j) = \delta_{i,j} \quad \forall i, j = 1, ...n_k$$

où les A_j dénotent dans un triangle T_l les noeuds X_I du maillage.

Les fonctions de forme N_i s'expriment simplement en fonctions des coordonnées barycentriques $\lambda_1, \lambda_2, \lambda_3$ du triangle.

8.8.1 Fonctions de forme P1

En exprimant les conditions de Lagrange on retrouve que les N_i coincident avec les λ_i. On a donné leur expression analytique ainsi que celles de leurs gradients dans le chapitre précédent.

8.8.2 Fonctions de forme P2

On obtient pour les noeuds aux sommets correspondants aux indices locaux $1, 2, 3$:

$$N_i = \lambda_i \left(2\lambda_i - 1\right) \quad \text{pour } i = 1, 2, 3 \tag{8.13}$$

et

$$\mathbf{grad} N_i = (4\lambda_i - 1)\mathbf{grad}\lambda_i \quad \text{pour } i = 1, 2, 3 \tag{8.14}$$

FIGURE 8.9 – Fonctions de forme P2

Pour les noeuds aux milieux des côtés correspondants aux indices locaux $4, 5, 6$:

$$N_4 = 4\lambda_1\,\lambda_2 \tag{8.15}$$

$$N_5 = 4\lambda_2\,\lambda_3 \tag{8.16}$$

$$N_6 = 4\lambda_3\,\lambda_1 \tag{8.17}$$

et

$$\mathbf{grad}N_4 = 4\left(\lambda_1\,\mathbf{grad}\lambda_2 + \lambda_2\,\mathbf{grad}\lambda_1\right) \tag{8.18}$$

$$\mathbf{grad}N_5 = 4\left(\lambda_2\,\mathbf{grad}\lambda_3 + \lambda_3\,\mathbf{grad}\lambda_2\right) \tag{8.19}$$

$$\mathbf{grad}N_6 = 4\left(\lambda_3\,\mathbf{grad}\lambda_1 + \lambda_1\,\mathbf{grad}\lambda_3\right) \tag{8.20}$$

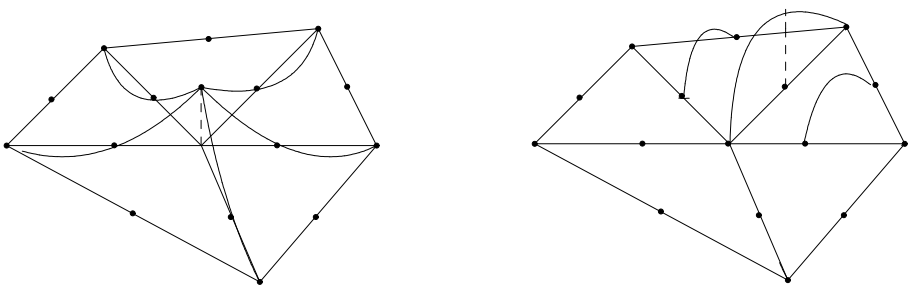


FIGURE 8.10 – Fonctions de base P2 et leur support.

8.8.3 Fonctions de forme P3

On obtient pour les noeuds aux sommets correspondants aux indices locaux $1, 2, 3$:

$$N_i = \frac{\lambda_i\left(3\lambda_i - 1\right)\left(3\lambda_i - 2\right)}{2} \quad \text{pour } i = 1, 2, 3 \tag{8.21}$$

Pour les noeuds, aux un tiers et deux tiers des côtés, correspondants aux indices locaux 4 à 9 :

$$N_4 = \frac{9\lambda_1\,(3\lambda_1 - 1)\lambda_2}{2} \quad \text{etc .. pour } i = 5, 6, 7, 8, 9 \tag{8.22}$$

Enfin, pour le noeud 10 au barycentre, on obtient la fonction "bulle", nulle sur les côtés du triangle et souvent utilisée dans les formulations mixtes.

$$N_{10} = 27\,\lambda_1\,\lambda_2\,\lambda_3 \tag{8.23}$$

8.9 Application aux problèmes elliptiques

8.9.1 Calcul des matrices et second-membres élémentaires

Ces problèmes font intervenir des intégrales de la forme :

$$\iint_\Omega \mathbf{grad}u\ \mathbf{grad}v\,dxdy$$

$$\iint_\Omega u\,v\,dxdy$$

$$\iint_\Omega f\,v\,dxdy$$

et éventuellement

$$\int_\Gamma g\,v\,d\gamma$$

Leur discrétisation, utilisant les fonctions de base w_I, nécessite le calcul des intégrales suivantes :

$$\iint_\Omega \mathbf{grad}w_J\ \mathbf{grad}w_I\,dxdy$$

$$\iint_\Omega w_J\,w_I\,dxdy$$

et

$$\int_\Gamma w_J\,w_I\,d\gamma$$

Ces calculs se ramènent à des intégrales sur les éléments triangulaires et leurs arêtes de fonctions composées de puissances et de produits des coordonnées barycentriques λ_i. Par conséquent la formule exacte suivante est souvent très utile.

$$\iint_T \lambda_1^n\,\lambda_2^p\,\lambda_3^q\,dxdy = 2\,Aire(T)\,\frac{n!p!q!}{(n+p+q+2)!}$$

Cas particuliers importants :

$$\iint_T \lambda_i \, dxdy = \frac{Aire(T)}{3} \quad \forall i = 1, 2, 3$$

$$\iint_T \lambda_i^2 \, dxdy = \frac{Aire(T)}{6} \quad \forall i = 1, 2, 3$$

$$\iint_T \lambda_i \, \lambda_j \, dxdy = \frac{Aire(T)}{12} \quad \forall i, j = 1, 2, 3$$

Par exemple le calcul de la matrice de raideur correspondant au laplacien nécessite l'évaluation des intégrales :

$$\iint_\Omega \mathbf{grad} w_J(x, y) \, \mathbf{grad} w_I(x, y) \, dxdy$$

En éléments P2 droits, les matrices élémentaires correspondantes sont constituées des coefficients :

$$a_{i,j} = \iint_T \mathbf{grad} N_j(x, y) \, \mathbf{grad} N_i(x, y) \, dxdy \quad \forall i, j = 1, 6$$

L'utilisation des formules des gradients données plus haut conduit à la matrice élémentaire qui s'exprime en fonction des constantes $b_{i,j} = \mathbf{grad}\lambda_i \, \mathbf{grad}\lambda_j$ déjà calculées dans le chapitre précédent. On obtient

$$b_{1,1} = \frac{(y_2 - y_3)^2 + (x_2 - x_3)^2}{4 \, (Aire(T))^2}$$

$$b_{1,2} = \frac{(y_2 - y_3)(y_3 - y_1) + (x_2 - x_3)(x_3 - x_1)}{4 \, (Aire(T))^2}$$

etc.. en tournant sur les indices $1, 2, 3$. Le calcul complet de la matrice élémentaire P2 est laissé au lecteur.

Les intégrations sur des morceaux de frontière se font en utilisant des formules de quadrature approchée à une dimension de type trapèzes ou Simpson.

8.9.2 Technique de l'élément de référence

Les calculs des fonctions de forme et des matrices et second-membres élémentaires peuvent s'effectuer en se ramenant à un élément de référence simple.
Dans le cas d'éléments triangulaires on choisit le triangle rectangle isocèle unité de sommets :

$$a_1 = (0, 0), \; a_2 = (1, 0), \; a_3 = (0, 1)$$

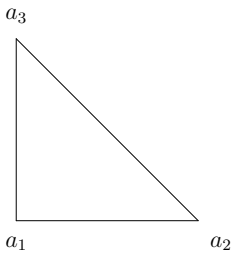


FIGURE 8.11 – Élément de référence

Dans ce triangle de référence les coordonnées barycentriques admettent les expressions suivantes en fonction des coordonnées cartésiennes notées $\hat{x}$ et $\hat{y}$:

$$\begin{cases} \hat{\lambda}_1 = 1 - \hat{x} - \hat{y} \\ \hat{\lambda}_2 = \hat{x} \\ \hat{\lambda}_3 = \hat{y} \end{cases}$$

Par transformation affine inversible on fait correspondre à ce triangle de référence un élément triangulaire droit quelconque T de sommets

$$A_1(x_1, y_1), A_2(x_2, y_2), A_3(x_3, y_3)$$

Pour cela il suffit de construire une transformation affine F qui associe respectivement les sommets a_1, a_2, a_3 du triangle de référence aux sommets A_1, A_2, A_3 du triangle quelconque.

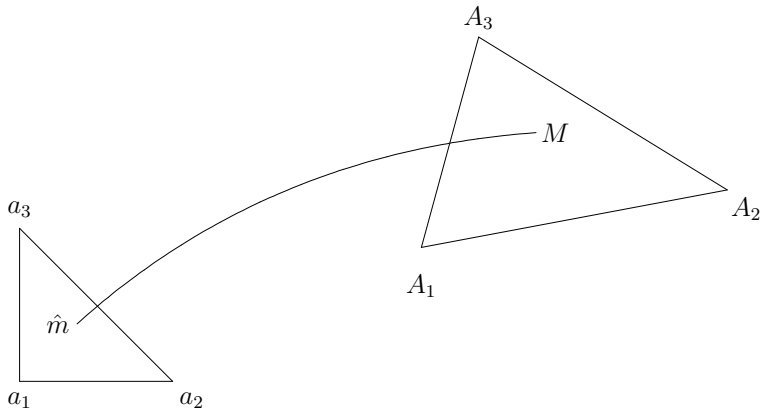


FIGURE 8.12 – Transformation affine entre l'élément de référence et un triangle quelconque

Il est très commode pour cela d'utiliser les fonctions coordonnées barycentriques. On obtient les coordonnées x, y du point M de T correspondant au point $\hat{m}(\hat{x}, \hat{y})$

du triangle de référence selon :

$$x = x_1\hat{\lambda}_1(\hat{x},\hat{y}) + x_2\hat{\lambda}_2(\hat{x},\hat{y}) + x_3\hat{\lambda}_3(\hat{x},\hat{y})$$

$$y = y_1\hat{\lambda}_1(\hat{x},\hat{y}) + y_2\hat{\lambda}_2(\hat{x},\hat{y}) + y_3\hat{\lambda}_3(\hat{x},\hat{y})$$

Le jacobien de la transformation est égal à deux fois l'aire algébrique du triangle T :

$$det(J) = \begin{vmatrix} \dfrac{\partial x}{\partial \hat{x}} & \dfrac{\partial x}{\partial \hat{y}} \\ \dfrac{\partial y}{\partial \hat{x}} & \dfrac{\partial y}{\partial \hat{y}} \end{vmatrix} = \begin{vmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{vmatrix} = 2.Aire(T)$$

car c'est le rapport de l'aire du triangle T et de celle du triangle de référence (qui vaut $\frac{1}{2}$).

Cette transformation affine qui associe les sommets a_1, a_2, a_3 du triangle de référence aux sommets A_1, A_2, A_3, est donc inversible si les points A_1, A_2, A_3 ne sont pas alignés. Pour tout point M correspondant au point $\hat{m}$ on a :

$$\hat{\lambda}_i(\hat{x},\hat{y}) = \lambda_i(x,y) \quad \forall i = 1,2,3$$

autrement dit, les coordonnées barycentriques sont conservées dans la transformation affine ainsi définie.

On ramène alors simplement le calcul des intégrales sur un T quelconque au calcul homologue sur le triangle de référence $\hat{T}$ selon :

$$\iint_T \Phi(x,y)\,dxdy = 2.Aire(T) \iint_{\hat{T}} \Phi(x(\hat{x},\hat{y}), y(\hat{x},\hat{y}))\,d\hat{x}d\hat{y}$$

8.9.3 Calcul des gradients

Le calcul des matrices de raideur fait intervenir le calcul des gradients des fonctions de forme. Dans le cas d'éléments Pk cela se ramène au calcul des gradients des coordonnées barycentriques λ_i. Il nous faut donc calculer ces gradients en fonction de ceux des coordonnées barycentriques de l'élément de référence. On écrit :

$$\frac{\partial \lambda_i}{\partial x} = \frac{\partial \hat{\lambda}_i}{\partial \hat{x}}\frac{\partial \hat{x}}{\partial x} + \frac{\partial \hat{\lambda}_i}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial x}$$

et de même

$$\frac{\partial \lambda_i}{\partial y} = \frac{\partial \hat{\lambda}_i}{\partial \hat{x}}\frac{\partial \hat{x}}{\partial y} + \frac{\partial \hat{\lambda}_i}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial y}$$

Ce qui s'exprime sous forme matricielle selon :

$$\mathbf{grad}\lambda_i = \frac{1}{2Aire(T)} \begin{pmatrix} y_3 - y_1 & y_1 - y_2 \\ x_1 - x_3 & x_2 - x_1 \end{pmatrix} \mathbf{grad}\hat{\lambda}_i$$

et donne à nouveau les résultats obtenus dans le chapitre précédent.

8.10 Éléments finis isoparamétriques triangulaires et quadrangulaires

Les éléments finis isoparamétriques sont des éléments dont les fonctions de forme sont définies à partir de fonctions de forme mères construites sur un élément de référence. Les fonctions de forme de l'élément réel sont les transformées des fonctions de forme mères dans la transformation géométrique faisant passer de l'élément de référence à l'élément réel. **Si cette transformation géométrique s'exprime dans la base des fonctions de formes mères construites sur l'élément de référence, on parle d'élément isoparamétrique.** Car, dans ce cas, les mêmes fonctions servent à définir la base de l'espace d'approximation et la géométrie des éléments. A l'inverse, dans le cas des éléments $P2$ droits, par exemple, les fonctions de formes mères sont des polynômes de degré deux alors que la transformation qui fait passer du triangle de référence au triangle réel est affine. On parle alors d'élément sous-paramétrique.

8.10.1 Les éléments quadrilatéraux bilinéaires de Lagrange : les éléments Q1

Considérons un maillage du domaine Ω de $\mathbb{R}^2$ en quadrangles C_l pour $l = 1, ..Nbe$, satisfaisant aux critères caractérisant un maillage par éléments finis. Essayons de construire une approximation par éléments finis correspondant au choix d'un espace V_h d'approximation constitué de fonctions continues sur $\bar{\Omega}$, régulières et d'expression simple sur chaque élément quadrangulaire. La première idée qui vient à l'esprit consisterait à choisir des polynômes dans chaque quadrangle.
Prenons par exemple un quadrangle droit et choisissons comme degrés de liberté dans ce quadrangle les valeurs de la fonction inconnue aux 4 sommets du quadrangle. Ces 4 valeurs définissent de façon unique un polynôme à 4 coefficients de la forme

$$a_0 + a_1 x + a_2 y + a_3 xy$$

Mais ce choix d'espace d'approximation V_h constitué des fonctions polynomiales de la forme ci-dessus par quadrangles ne fournit pas de méthode d'éléments finis

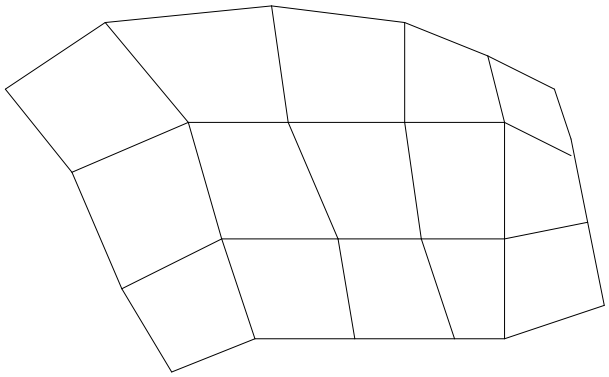


FIGURE 8.13 – Maillage en quadrangles

conformes acceptable. En effet, sur chaque côté commun à deux quadrangles adjacents (d'équation $y = mx + p$), la restriction des polynômes de la forme :

$$a_0 + a_1 x + a_2 y + a_3 xy$$

est un polynôme de degré 2 en x. Et 2 points ne suffisent pas à fixer de manière unique un polynôme de degré 2 à une variable donc à assurer la continuité inter-éléments et la continuité globale des fonctions de cet espace d'approximation.

Remarque 8.10.1 *Par contre, dans le cas de maillages par des éléments rectangulaires (ce qui conduit nécessairement à des maillages structurés et ne s'applique qu'à des domaines simples décomposables en rectangles), en prenant les côtés des rectangles parallèlement aux axes de coordonnées, le choix d'un espace d'approximation V_h de fonctions continues polynomiales de la forme $a_0 + a_1 x + a_2 y + a_3 xy$ par élément rectangulaire est cohérent. Dans ce cas les côtés des éléments ont des équations de la forme $x = cste$ ou $y = cste$ et les restrictions des fonctions de base sur un côté sont des fonctions affines à une variable complètement déterminées par leurs valeurs aux deux extrémités de ce côté. On laisse au lecteur le développement complet de cette méthode d'éléments finis rectangulaires. Signalons que l'on retrouve ainsi des méthodes classiques de discrétisation par différences finies. En particulier dans le cas du laplacien on retrouve, avec ces éléments rectangulaires bilinéaires et la formule d'intégration approchée des trapèzes, le schéma à 5 points usuel.*

Voyons maintenant la solution au problème évoqué plus haut : comment construire un espace d'approximation acceptable avec une discrétisation géométrique du domaine de calcul en quadrangles ?

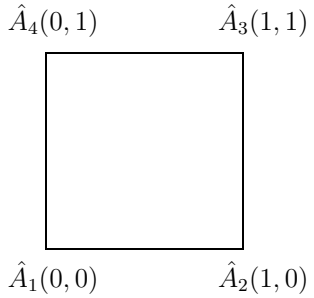


FIGURE 8.14 – Élément de référence Q1 : le carré unité

On choisit comme élément de référence le carré unité et comme points nodaux les 4 sommets. On note $\hat{x}$ et $\hat{y}$ les coordonnées d'un point du carré unité. On choisit comme espace de fonctions de forme dans l'élément de référence l'espace des polynômes bilinéaires de la forme :

$$a_0 + a_1\hat{x} + a_2\hat{y} + a_3\hat{x}\hat{y}$$

Sur ce carré unité, les fonctions de forme mères sont définies classiquement par les conditions de Lagrange

$$\hat{N}_i(\hat{A}_j) = \delta_{ij} \quad \forall\, i,j = 1,..4$$

Ce qui donne les formules suivantes

$$\begin{cases} \hat{N}_1 = (1-\hat{x})(1-\hat{y}) \\ \hat{N}_2 = \hat{x}(1-\hat{y}) \\ \hat{N}_3 = \hat{x}\hat{y} \\ \hat{N}_4 = (1-\hat{x})\hat{y} \end{cases}$$

Ces quatre fonctions forment une base de l'espace des fonctions bilinéaires et toute fonction bilinéaire est déterminée uniquement par la donnée de ses valeurs aux 4 sommets du carré unité. Considérons alors l'application F

$$(\hat{x},\hat{y}) \xrightarrow{F} (x,y)$$

du carré unité dans un quadrilatère quelconque de sommets

$$A_1(x_1,y_1), A_2(x_2,y_2), A_3(x_3,y_3), A_4(x_4,y_4)$$

définie par les formules suivantes :

$$x = x_1\hat{N}_1 + x_2\hat{N}_2 + x_3\hat{N}_3 + x_4\hat{N}_4$$

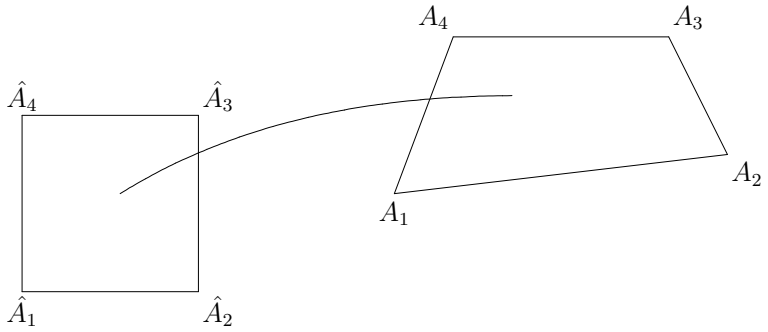


FIGURE 8.15 – Transformation du carré unité en quadrangle quelconque

$$y = y_1\hat{N}_1 + y_2\hat{N}_2 + y_3\hat{N}_3 + y_4\hat{N}_4$$

Cette application fait correspondre aux quatre sommets du carré unité numérotés dans le sens direct les quatre sommets A_1, A_2, A_3, A_4 du quadrilatère également pris dans le sens direct . On va montrer que c'est une transformation inversible si le quadrilatère est convexe. Pour cela on calcule le déterminant jacobien de cette transformation.

$$det(J) = \frac{\partial x}{\partial \hat{x}}\frac{\partial y}{\partial \hat{y}} - \frac{\partial x}{\partial \hat{y}}\frac{\partial y}{\partial \hat{x}}$$

On trouve :

$$\begin{aligned} det(J) &= [(x_2 - x_1)(y_4 - y_1) - (x_4 - x_1)(y_2 - y_1)](1 - \hat{x})(1 - \hat{y}) \\ &+[(x_3 - x_2)(y_1 - y_2) - (x_1 - x_2)(y_3 - y_2)]\hat{x}(1 - \hat{y}) \\ &+[(x_4 - x_3)(y_2 - y_3) - (x_2 - x_3)(y_4 - y_3)]\hat{x}\hat{y} \\ &+[(x_1 - x_4)(y_3 - y_4) - (x_3 - x_4)(y_1 - y_4)](1 - \hat{x})\hat{y} \end{aligned}$$

Les coefficients de $det(J)$ dans la base des $\hat{N}_i$ sont les aires algébriques des 4 parallélogrammes construits sur deux côtés adjacents du quadrilatère. Ces aires sont toutes positives si le quadrilatère est convexe. En effet en notant L_{ij} la longueur du côté A_iA_j et θ_i l'angle interne au sommet A_i on a :

$$\begin{aligned} det(J) &= L_{12}L_{14}sin(\theta_1)\hat{N}_1 + L_{12}L_{23}sin(\theta_2)\hat{N}_2 \\ &+L_{23}L_{34}sin(\theta_3)\hat{N}_3 + L_{14}L_{34}sin(\theta_4)\hat{N}_4 \end{aligned}$$

Si le quadrilatère est convexe, les angles θ_i sont strictement inférieurs à π et leur sinus strictement positifs. Les coefficients de $det(J)$ dans la base des $\hat{N}_i$ sont tous quatre strictement positifs et par conséquent $det(J)$ est toujours strictement positif dans le carré unité. D'ailleurs on peut également remarquer que les termes en $\hat{x}\hat{y}$ s'annulent dans le développement de $det(J)$. Donc le jacobien est une fonction affine (P1). Comme il prend des valeurs strictement positives en quatre points distincts et non alignés, il est toujours strictement positif.

8.10.2 Fonctions de forme Q1

Les fonctions de forme choisies dans un quadrilatère quelconque du maillage sont les transformées des fonctions de forme mères $\hat{N}_i$ selon :

$$N_i(x,y) = \hat{N}_i(\hat{x}(x,y), \hat{y}(x,y))$$

On considère les fonctions q obtenues par la transformation inversible précédente F à partir des fonctions bilinéaires $\hat{q}$ sur le carré de référence.

$$(x,y) \longrightarrow q(x,y) = \hat{q}(\hat{x},\hat{y}) \quad \text{avec} \quad (\hat{x},\hat{y}) \xrightarrow{F} (x,y)$$

Une fonction $q \in Q1$ prenant les valeurs q_i aux sommets A_i du quadrilatère s'écrira donc

$$q(x,y) = \sum_{i=1,..4} q_i \, N_i(x,y) = \sum_{i=1,..4} q_i \, \hat{N}_i(\hat{x},\hat{y})$$

Notons, en effet, que les fonctions de formes mères sont des fonctions de base (locales) des polynômes de type $Q1$ sur l'élément de référence.

8.10.3 Fonctions de base Q1

Ce sont les fonctions w_I, prenant la valeur 1 en un noeud (d'indice I) du maillage et la valeur zéro en tous les autres noeuds, dont les restrictions dans chaque quadrilatère sont des fonctions de forme Q1.

Toute fonction de V_h s'écrira comme combinaison linéaire des w_I. Remarquons que V_h est bien cette fois un espace de fonctions continues. Toute fonction de V_h est uniquement déterminée par ses valeurs aux noeuds du maillage, sommets des quadrilatères. Deux fonctions de V_h égales en deux sommets adjacents A_i et A_j sont égales sur l'arête A_iA_j joignant ces sommets. Car, par transformation continue inverse, les fonctions correspondantes sont égales sur 2 sommets adjacents du carré de référence $\hat{A}_i$ et $\hat{A}_j$ et par construction des fonctions de forme mères, elles sont égales sur tout le côté $\hat{A}_i\hat{A}_j$ qui s'applique lui-même sur l'arête A_iA_j.

8.10.4 Calcul des gradients des fonctions de base Q1

Les restrictions des fonctions de base dans les quadrangles élémentaires C_l sont les fonctions de forme N_i. Rappelons que les N_i sont définies à partir de fonctions de forme $\hat{N}_i$ dans le carré de référence $\hat{C}$ par la transformation F

$$N_i(x,y) = \hat{N}_i(\hat{x}(x,y), \hat{y}(x,y))$$

Le calcul de leur gradients se ramène donc à ceux des fonctions de forme mères sur l'élément de référence.

$$\frac{\partial N_i}{\partial x} = \frac{\partial \hat{N}_i}{\partial \hat{x}}\frac{\partial \hat{x}}{\partial x} + \frac{\partial N_i}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial x}$$

$$\frac{\partial N_i}{\partial y} = \frac{\partial \hat{N}_i}{\partial \hat{x}}\frac{\partial \hat{x}}{\partial y} + \frac{\partial N_i}{\partial \hat{y}}\frac{\partial \hat{y}}{\partial y}$$

Ce qui peut s'écrire sous la forme

$$\mathbf{grad} N_i = \begin{pmatrix} \frac{\partial \hat{x}}{\partial x} & \frac{\partial \hat{y}}{\partial x} \\ \frac{\partial \hat{x}}{\partial y} & \frac{\partial \hat{y}}{\partial y} \end{pmatrix} \mathbf{grad} \hat{N}_i$$

Les expressions données des $\hat{N}_i$ permettent le calcul aisé de leur gradient. Il nous reste donc à calculer la matrice des dérivées partielles ci-dessus. Or on obtient facilement la matrice jacobienne inverse J donnant :

$$\mathbf{grad} \hat{N}_i = \begin{pmatrix} \frac{\partial x}{\partial \hat{x}} & \frac{\partial y}{\partial \hat{x}} \\ \frac{\partial x}{\partial \hat{y}} & \frac{\partial y}{\partial \hat{y}} \end{pmatrix} \mathbf{grad} N_i$$

par :

$$\frac{\partial x}{\partial \hat{x}} = \sum_{i=1,..4} x_i \frac{\partial \hat{N}_i}{\partial \hat{x}}$$

etc..

On en déduit :

$$\mathbf{grad} N_i = J^{-1} \mathbf{grad} \hat{N}_i = \frac{1}{det(J)} \begin{pmatrix} \frac{\partial y}{\partial \hat{y}} & -\frac{\partial y}{\partial \hat{x}} \\ -\frac{\partial x}{\partial \hat{y}} & \frac{\partial x}{\partial \hat{x}} \end{pmatrix} \mathbf{grad} \hat{N}_i$$

Notons que les gradients des fonctions de forme N_i sont des fractions rationnelles (le numérateur est un polynôme de degré deux et le dénominateur $det(J)$ un polynôme $P1$). On aura donc besoin de formules de quadrature numérique.

Remarque 8.10.2 *Le calcul des gradients qui précède est général. Nous le présentons à propos des fonctions $Q1$ mais on peut l'appliquer à tous les cas de fonctions de base éléments finis. Dans le cas d'éléments Pk nous avons trouvé plus simple de passer par les coordonnées barycentriques.*

8.10.5 Application aux problèmes elliptiques. Calcul des matrices et second-membres élémentaires

Ces problèmes font intervenir des intégrales de la forme :

$$\iint_\Omega \mathbf{grad}u\ \mathbf{grad}v\, dxdy$$

$$\iint_\Omega u\, v\, dxdy$$

$$\iint_\Omega f\, v\, dxdy$$

et éventuellement

$$\int_\Gamma g\, v\, d\gamma$$

Leur discrétisation, utilisant les fonctions de base w_I, nécessitent le calcul des intégrales suivantes :

$$\iint_\Omega \mathbf{grad}w_J\ \mathbf{grad}w_I\, dxdy$$

$$\iint_\Omega w_J\, w_I\, dxdy$$

et

$$\int_\Gamma w_J\, w_I\, d\gamma$$

Ces calculs se font par des sommes d'intégrales sur les éléments quadrangulaires qui se ramènent par transformation de variables à des intégrales sur le carré unité.

Par exemple le calcul de la matrice de masse élémentaire revient au calcul de ces coefficients pour tout i et $j = 1, 4$

$$m_{i,j} = \iint_C N_j(x,y)\, N_i(x,y)\, dxdy = \iint_{\hat{C}} \hat{N}_j(\hat{x},\hat{y})\, \hat{N}_i(\hat{x},\hat{y})\, |det(J)|\, d\hat{x}d\hat{y}$$

Le calcul de la matrice de raideur élémentaire correspondant au laplacien nécessite l'évaluation des intégrales :

$$\begin{aligned} a_{i,j} &= \iint_C \mathbf{grad}N_j(x,y)\, \mathbf{grad}N_i(x,y)\, dxdy \\ &= \iint_{\hat{C}} J^{-1}\mathbf{grad}\hat{N}_j(\hat{x},\hat{y})\, J^{-1}\mathbf{grad}\hat{N}_i(\hat{x},\hat{y})\, |det(J)|\, d\hat{x}d\hat{y} \quad \forall i,j = 1,4 \end{aligned}$$

8.10.6 Éléments isoparamétriques Q2

Les développements précédents peuvent être repris dans le cas d'éléments biquadratiques obtenus par transformation isoparamétrique à partir du carré unité à 9 noeuds sur lequel on définit des fonctions de forme mères de la forme :

$$a_0 + a_1 x + a_2 y + a_3 xy + a_4 x^2 + a_5 y^2 + a_6 x^2 y + a_7 xy^2 + a_8 x^2 y^2$$

La transformation permet alors d'obtenir des quadrangles courbes

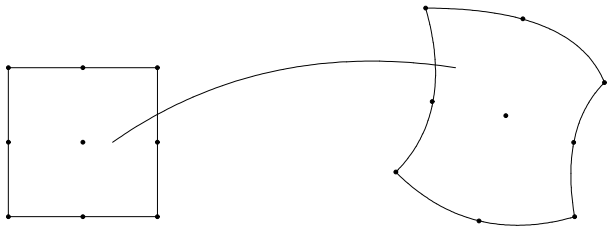


FIGURE 8.16 – Transformation du carré unité Q2 en quadrangle courbe

8.10.7 Éléments isoparamétriques P2

On considère cette fois des triangles curvilignes obtenus par transformation isoparamétrique P2 à partir du triangle de référence isocèle rectangle unité à 6 noeuds. Les fonctions de forme utilisées dans l'élément de référence sont les fonctions de forme P2 définies dans le cours précédent. La transformation s'exprime dans la base des fonctions de forme P2, polynômes de degré 2 dans les variables $\hat{x}, \hat{y}$ selon :

$$x = \sum_{i=1,..6} x_i \hat{N}_i$$

$$y = \sum_{i=1,..6} y_i \hat{N}_i$$

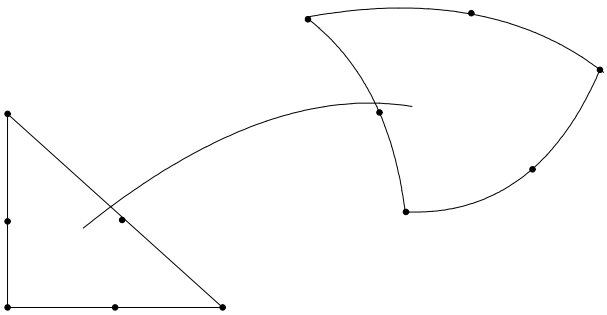


FIGURE 8.17 – Transformation du triangle unité P2 en triangle curviligne

Chapitre 9

Exemple de discrétisation de systèmes : l'élasticité linéaire

9.1 Le modèle en contraintes planes

Les problèmes de calcul de structures dans le cas de l'élasticité linéaire ont été à l'origine de la méthode des éléments finis. Considérons le problème de l'équilibre d'un domaine plan Ω, fixé sur une partie Γ_0 de sa frontière Γ, soumis à des efforts surfaciques de densité $\mathbf{f}$ et à un chargement $\mathbf{g}$ sur la partie Γ_1 de Γ. Ces forces s'exercent dans le plan de Ω. Faisons l'hypothèse de petits déplacements pour

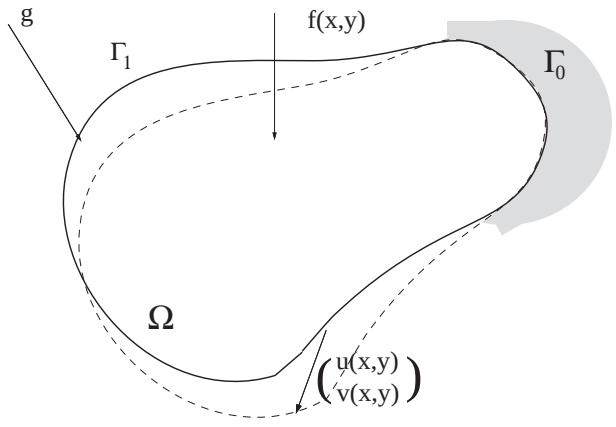


FIGURE 9.1 – Problème en contraintes planes

lesquels les équations de l'élasticité linéaire sont valables et adoptons le modèle de

contraintes planes adapté aux plaques minces sollicitées par des forces s'exerçant dans leur plan. L'inconnue du problème est le vecteur déplacement

$$\mathbf{u} = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix}$$

Le vecteur déformation est donné, dans le cas d'un problème en contraintes planes, par ses composantes fonction des dérivées premières des déplacements selon :

$$\epsilon = \begin{pmatrix} \epsilon_{xx} \\ \epsilon_{yy} \\ 2\epsilon_{xy} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial u}{\partial x} \\ \dfrac{\partial v}{\partial y} \\ \dfrac{\partial u}{\partial y} + \dfrac{\partial v}{\partial x} \end{pmatrix}$$

Ce que l'on peut écrire en utilisant la matrice d'opérateurs différentiels

$$B = \begin{pmatrix} \dfrac{\partial}{\partial x} & 0 \\ 0 & \dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial x} \end{pmatrix}$$

sous la forme

$$\epsilon = B\mathbf{u}$$

Les équations de l'élasticité s'expriment par une relation linéaire entre contraintes et déformation qui dans un problème en contraintes planes s'écrit :

$$\sigma = \mathbf{D}\epsilon$$

où σ est le vecteur des contraintes :

$$\sigma = \begin{pmatrix} \sigma_{xx} \\ \sigma_{yy} \\ \sigma_{xy} \end{pmatrix}$$

et $\mathbf{D}$ est la matrice de Hooke dont l'expression, fonction du module de Young E et du coefficient de Poisson ν, s'écrit :

$$\mathbf{D} = \frac{E}{1-\nu^2} \begin{pmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & (1-\nu)/2 \end{pmatrix}$$

Enfin les équations d'équilibre entre efforts et contraintes s'écrivent :

$$\begin{cases} \dfrac{\partial}{\partial x}\sigma_{xx} + \dfrac{\partial}{\partial y}\sigma_{xy} + f_x = 0 \\ \\ \dfrac{\partial}{\partial x}\sigma_{xy} + \dfrac{\partial}{\partial y}\sigma_{yy} + f_y = 0 \end{cases}$$

où f_x et f_y sont les 2 composantes des forces de surface.

En utilisant la matrice d'opérateurs différentiels B on peut écrire les équations d'équilibre sous la forme :

$$B^T\sigma + \mathbf{f} = 0$$

soit avec

$$\sigma = \mathbf{D}\epsilon \quad \text{et} \quad \epsilon = B\mathbf{u}$$

On obtient ainsi un système d'équations aux dérivées partielles du second ordre

$$B^T\mathbf{D}B\mathbf{u} + \mathbf{f} = 0$$

dont l'inconnue est le vecteur déplacement.

Conditions aux limites

Les conditions aux limites sont habituellement de 2 types.

Conditions de type Dirichlet, dites essentielles lorsque les déplacements sont fixés, par exemple à zéro sur la frontière.

$$u = v = 0 \quad \text{sur} \quad \Gamma_0$$

Conditions de type Neuman, dites libres ou naturelles dans le cas homogène, qui s'écrivent sous la forme générale :

$$\begin{cases} \sigma_{xx}\, n_x + \sigma_{xy}\, n_y = g_x \\ \\ \sigma_{xy}\, n_x + \sigma_{yy}\, n_y = g_y \end{cases}$$

9.2 Formulation variationnelle d'un problème d'élasticité linéaire. Principe des travaux virtuels

La formulation variationnelle s'obtient comme dans le cas du laplacien par multiplication par une fonction test $\mathbf{w}$, intégration et utilisation de la formule de

Green. La difficulté supplémentaire dans ce cas provient du fait que l'inconnue est un vecteur. On devra donc multiplier scalairement par une fonction test vectorielle $\mathbf{w}$. Notons w_x et w_y ses deux composantes. On obtient tout d'abord

$$\iint_\Omega \left[\left(\frac{\partial}{\partial x}\sigma_{xx} + \frac{\partial}{\partial y}\sigma_{xy}\right) w_x \; + \left(\frac{\partial}{\partial x}\sigma_{xy} + \frac{\partial}{\partial y}\sigma_{yy}\right) w_y \right] dxdy \; + \iint_\Omega \mathbf{f}\,\mathbf{w} dxdy = 0$$

On utilise la formule de Green, ce qui donne

$$\iint_\Omega \left[\sigma_{xx}\frac{\partial}{\partial x}w_x + \sigma_{xy}\frac{\partial}{\partial y}w_x \; + \sigma_{xy}\frac{\partial}{\partial x}w_y + \sigma_{yy}\frac{\partial}{\partial y}w_y \right] dxdy$$

$$- \int_\Gamma \left[\sigma_{xx}w_x n_x + \sigma_{xy}w_x n_y \; + \sigma_{xy}w_y n_x + \sigma_{yy}w_y n_y\right] d\gamma = \iint_\Omega \mathbf{f}\,\mathbf{w} dxdy$$

Les conditions aux limites s'intègrent dans le terme de bord. Donc avec des conditions de Dirichlet homogènes sur Γ_0 et un effort $\mathbf{g}$ s'exerçant sur Γ_1, on obtient :

$$\iint_\Omega \left[\sigma_{xx}\frac{\partial}{\partial x}w_x + \sigma_{xy}\left(\frac{\partial}{\partial y}w_x \; + \frac{\partial}{\partial x}w_y\right) + \sigma_{yy}\frac{\partial}{\partial y}w_y \right] dxdy$$
$$= \int_{\Gamma_1} \mathbf{g}\,\mathbf{w}\, d\gamma \; + \iint_\Omega \mathbf{f}\,\mathbf{w}\, dxdy$$

pour toute fonction vectorielle $\mathbf{w}$ de composantes dans l'espace de Hilbert V des fonctions de $H^1(\Omega)$, nulles sur Γ_0.

On obtient ainsi la formulation variationnelle de ce problème en contrainte plane

$$\begin{cases} \text{Trouver la fonction vectorielle } \mathbf{u} \text{ appartenant à } V^2 \text{ telle que :} \\ \displaystyle\iint_\Omega \sigma(\mathbf{u})\,\epsilon(\mathbf{w})\, dxdy = \int_{\Gamma_1} \mathbf{g}\,\mathbf{w}\, d\gamma \; + \iint_\Omega \mathbf{f}\,\mathbf{w}\, dxdy \quad \forall\, \mathbf{w} \in V^2 \end{cases}$$

Ceci s'interprète également comme l'expression du principe des travaux virtuels et peut s'écrire, en utilisant les matrices introduites plus haut sous la forme :

$$\begin{cases} \text{Trouver la fonction vectorielle } \mathbf{u} \text{ appartenant à } V^2 \text{ telle que :} \\ \displaystyle\iint_\Omega \mathbf{D}B\mathbf{u}\,.\,B\mathbf{w}\, dxdy = \int_{\Gamma_1} \mathbf{g}\,\mathbf{w}\, d\gamma \; + \iint_\Omega \mathbf{f}\,\mathbf{w}\, dxdy \quad \forall\, \mathbf{w} \in V^2 \end{cases}$$

L'existence et l'unicité de la solution de ce problème dans l'espace de Hilbert V^2 s'obtient classiquement par le théorème de Lax-Milgram. L'ellipticité de la forme bilinéaire

$$\mathbf{u}, \mathbf{w} \longrightarrow a(\mathbf{u}, \mathbf{w}) = \iint_\Omega \sigma(\mathbf{u})\,\epsilon(\mathbf{w})\, dxdy$$

dans V^2 résulte de l'inégalité de Korn qui nécessite que des conditions de type Dirichlet soient imposées sur une partie Γ_0 de mesure non-nulle. En l'absence de fixation sur une partie de la frontière, il existe une infinité de solutions à un déplacement rigide près.

La solution du problème variationnel réalise le minimum dans l'espace V^2 de l'énergie potentielle :

$$J(\mathbf{v}) = \frac{1}{2}\iint_\Omega \sigma(\mathbf{v})\,\epsilon(\mathbf{v})\,dxdy - \int_{\Gamma_1} \mathbf{g}\,\mathbf{v}\,d\gamma \; - \; \iint_\Omega \mathbf{f}\,\mathbf{v}\,dxdy$$

9.3 Approximation par éléments finis P1

On suppose tout d'abord le domaine Ω maillé en éléments finis triangulaires droits. On écrit le problème approché dans l'espace V_h^2 des vecteurs à composantes continues, affines par éléments triangulaires et nulles sur la partie Γ_0 de la frontière. On est ainsi assuré de l'inclusion $V_h^2 \subset V^2$, donc la méthode d'éléments finis est dite conforme. Soit $\mathbf{I}$ l'ensemble des indices des noeuds du maillage correspondant à une valeur inconnue de la solution $\mathbf{u}$, et $N_\mathbf{I}$ le cardinal de $\mathbf{I}$, nombre de noeuds appartenant à $\mathbf{I}$, l'espace V_h^2 sera de dimension $2N_\mathbf{I}$. On choisit comme base de V_h^2, selon la technique de Lagrange, l'ensemble des vecteurs :

$$\begin{pmatrix} w_i \\ 0 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} 0 \\ w_i \end{pmatrix} \quad \text{pour } i = 1, ...N_\mathbf{I}$$

où les w_i sont les fonctions scalaires de base de Lagrange de l'espace des fonctions continues affines par triangles déjà rencontrées dans le cas de la discrétisation du laplacien.

Les composantes de la solution approchée s'écriront comme combinaisons linéaires des fonctions de base w_i :

$$u_h(x,y) = \sum_{j\in\mathbf{I}} u_j\, w_j(x,y)$$

$$v_h(x,y) = \sum_{j\in\mathbf{I}} v_j\, w_j(x,y)$$

En utilisant l'écriture matricielle de la formulation variationnelle, on obtient alors

le système linéaire suivant :

$$\left\{\begin{array}{l} \text{Trouver les composantes } u_j \text{ et } v_j \text{ pour } j \in \mathbf{I} \text{ telles que :} \\ \displaystyle\iint_\Omega \sum_{j\in\mathbf{I}} \mathbf{D}B \begin{pmatrix} w_j & 0 \\ 0 & w_j \end{pmatrix} \begin{pmatrix} u_j \\ v_j \end{pmatrix} . B \begin{pmatrix} w_i \\ 0 \end{pmatrix} dxdy = \\ \qquad \displaystyle\int_{\Gamma_1} \begin{pmatrix} g_x \\ g_y \end{pmatrix} \begin{pmatrix} w_i \\ 0 \end{pmatrix} d\gamma + \iint_\Omega \begin{pmatrix} f_x \\ f_y \end{pmatrix} \begin{pmatrix} w_i \\ 0 \end{pmatrix} dxdy \quad \forall\, i \in \mathbf{I} \\ \displaystyle\iint_\Omega \sum_{j\in\mathbf{I}} \mathbf{D}B \begin{pmatrix} w_j & 0 \\ 0 & w_j \end{pmatrix} \begin{pmatrix} u_j \\ v_j \end{pmatrix} . B \begin{pmatrix} 0 \\ w_i \end{pmatrix} dxdy = \\ \qquad \displaystyle\int_{\Gamma_1} \begin{pmatrix} g_x \\ g_y \end{pmatrix} \begin{pmatrix} 0 \\ w_i \end{pmatrix} d\gamma + \iint_\Omega \begin{pmatrix} f_x \\ f_y \end{pmatrix} \begin{pmatrix} 0 \\ w_i \end{pmatrix} dxdy \quad \forall\, i \in \mathbf{I} \end{array}\right.$$

Comme il y a 2 inconnues u_i et v_i par noeud du maillage, le système ci-dessus est un système linéaire de $2N_{\mathbf{I}}$ équations à $2N_{\mathbf{I}}$ inconnues.

Numérotons les inconnues dans l'ordre $u_1, v_1, u_2, v_2, ...u_i, v_i, ...u_{N_{\mathbf{I}}}, v_{N_{\mathbf{I}}}$. Les coefficients de la matrice de raideur et du second membre se calculent alors selon :

$$\begin{pmatrix} a_{2i-1,2j-1} & a_{2i-1,2j} \\ a_{2i,2j-1} & a_{2i,2j} \end{pmatrix} = \iint_\Omega \mathbf{D}B \begin{pmatrix} w_j & 0 \\ 0 & w_j \end{pmatrix} . B \begin{pmatrix} w_i & 0 \\ 0 & w_i \end{pmatrix} dxdy$$

et

$$\begin{pmatrix} b_{2i-1} \\ b_{2i} \end{pmatrix} = \int_{\Gamma_1} \begin{pmatrix} w_i & 0 \\ 0 & w_i \end{pmatrix} \begin{pmatrix} g_x \\ g_y \end{pmatrix} d\gamma + \iint_\Omega \begin{pmatrix} w_i & 0 \\ 0 & w_i \end{pmatrix} \begin{pmatrix} f_x \\ f_y \end{pmatrix} dxdy$$

9.4 Calculs des matrices et second-membre élémentaires P1

Un élément fini P1 comporte 3 noeuds qui sont les sommets du triangle T. En chaque noeud, nous avons 2 composantes de l'inconnue, où 2 fonctions scalaires de base non nulles. Il y a donc 6 degrés de liberté par triangles. La matrice de raideur élémentaire est donc une matrice 6×6 et le second membre élémentaire un vecteur à 6 composantes. Les restrictions des fonctions de base vectorielles non

nulles dans l'élément T s'expriment en fonction des 3 coordonnées barycentriques selon

$$\begin{pmatrix} \lambda_i \\ 0 \end{pmatrix} \text{ et } \begin{pmatrix} 0 \\ \lambda_i \end{pmatrix} \quad \text{pour } i = 1, 2, 3$$

Notons $\tilde{B}$ la matrice

$$B \begin{pmatrix} \lambda_1 & 0 & \lambda_2 & 0 & \lambda_3 & 0 \\ 0 & \lambda_1 & 0 & \lambda_2 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} \dfrac{\partial}{\partial x} & 0 \\ 0 & \dfrac{\partial}{\partial y} \\ \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial x} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & \lambda_2 & 0 & \lambda_3 & 0 \\ 0 & \lambda_1 & 0 & \lambda_2 & 0 & \lambda_3 \end{pmatrix}$$

Les λ_i étant des fonctions affines, la matrice $\tilde{B}$ est une matrice à coefficients constants que l'on détermine facilement à l'aide des gradients déjà calculés des coordonnées barycentriques. On obtient :

$$\tilde{B} = \frac{1}{2\,aire(T)} \begin{pmatrix} y_2 - y_3 & 0 & y_3 - y_1 & 0 & y_1 - y_2 & 0 \\ 0 & x_3 - x_2 & 0 & x_1 - x_3 & 0 & x_2 - x_1 \\ x_3 - x_2 & y_2 - y_3 & x_1 - x_3 & y_3 - y_1 & x_2 - x_1 & y_1 - y_2 \end{pmatrix}$$

Le calcul complet de la matrice élémentaire se poursuit sans difficulté. Les intégrations portant sur des fonctions constantes, il suffit de multiplier l'intégrande par l'aire du triangle. D'où le résultat : la matrice élémentaire de raideur K_T vaut

$$K_T = aire(T)\, \tilde{B}^T\, \mathbf{D}\, \tilde{B}$$

soit

$$\frac{1}{4\,aire(T)} \begin{pmatrix} y_2-y_3 & 0 & x_3-x_2 \\ 0 & x_3-x_2 & y_2-y_3 \\ y_3-y_1 & 0 & x_1-x_3 \\ 0 & x_1-x_3 & y_3-y_1 \\ y_1-y_2 & 0 & x_2-x_1 \\ 0 & x_2-x_1 & y_1-y_2 \end{pmatrix} \mathbf{D} \begin{pmatrix} y_2-y_3 & 0 & y_3-y_1 & 0 & y_1-y_2 & 0 \\ 0 & x_3-x_2 & 0 & x_1-x_3 & 0 & x_2-x_1 \\ x_3-x_2 & y_2-y3 & x_1-x_3 & y_3-y_1 & x_2-x_1 & y_1-y_2 \end{pmatrix}$$

Le second membre élémentaire s'obtient selon le même principe. Notons S le pourtour du triangle T, on calcule ses 6 composantes par :

$$\int_{\Gamma_1\cap S}\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_1\\ \lambda_2 & 0\\ 0 & \lambda_2\\ \lambda_3 & 0\\ 0 & \lambda_3\end{pmatrix}\begin{pmatrix} g_x\\ g_y\end{pmatrix} d\gamma + \iint_T \begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_1\\ \lambda_2 & 0\\ 0 & \lambda_2\\ \lambda_3 & 0\\ 0 & \lambda_3\end{pmatrix}\begin{pmatrix} f_x\\ f_y\end{pmatrix} dxdy$$

Si les fonctions **g** et **f** sont données par leurs valeurs aux noeuds du maillage, on trouve :

$$\int_{\Gamma_1\cap S}\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_1\\ \lambda_2 & 0\\ 0 & \lambda_2\\ \lambda_3 & 0\\ 0 & \lambda_3\end{pmatrix}\begin{pmatrix}\lambda_1 & 0 & \lambda_2 & 0 & \lambda_3 & 0\\ 0 & \lambda_1 & 0 & \lambda_2 & 0 & \lambda_3\end{pmatrix}\begin{pmatrix} g_{x1}\\ g_{y1}\\ g_{x2}\\ g_{y2}\\ g_{x3}\\ g_{y3}\end{pmatrix} d\gamma +$$

$$\iint_T \begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_1\\ \lambda_2 & 0\\ 0 & \lambda_2\\ \lambda_3 & 0\\ 0 & \lambda_3\end{pmatrix}\begin{pmatrix}\lambda_1 & 0 & \lambda_2 & 0 & \lambda_3 & 0\\ 0 & \lambda_1 & 0 & \lambda_2 & 0 & \lambda_3\end{pmatrix}\begin{pmatrix} f_{x1}\\ f_{y1}\\ f_{x2}\\ f_{y2}\\ f_{x3}\\ f_{y3}\end{pmatrix} dxdy$$

Soit

$$\int_{\Gamma_1\cap S}\begin{pmatrix}\lambda_1^2 & 0 & \lambda_2\lambda_1 & 0 & \lambda_3\lambda_1 & 0\\ 0 & \lambda_1^2 & 0 & \lambda_2\lambda_1 & 0 & \lambda_3\lambda_1\\ \lambda_1\lambda_2 & 0 & \lambda_2^2 & 0 & \lambda_3\lambda_2 & 0\\ 0 & \lambda_1\lambda_2 & 0 & \lambda_2^2 & 0 & \lambda_3\lambda_2\\ \lambda_1\lambda_3 & 0 & \lambda_2\lambda_3 & 0 & \lambda_3^2 & 0\\ 0 & \lambda_1\lambda_3 & 0 & \lambda_2\lambda_3 & 0 & \lambda_3^2\end{pmatrix}\begin{pmatrix} g_{x1}\\ g_{y1}\\ g_{x2}\\ g_{y2}\\ g_{x3}\\ g_{y3}\end{pmatrix} d\gamma +$$

$$\iint_T \begin{pmatrix}\lambda_1^2 & 0 & \lambda_2\lambda_1 & 0 & \lambda_3\lambda_1 & 0\\ 0 & \lambda_1^2 & 0 & \lambda_2\lambda_1 & 0 & \lambda_3\lambda_1\\ \lambda_1\lambda_2 & 0 & \lambda_2^2 & 0 & \lambda_3\lambda_2 & 0\\ 0 & \lambda_1\lambda_2 & 0 & \lambda_2^2 & 0 & \lambda_3\lambda_2\\ \lambda_1\lambda_3 & 0 & \lambda_2\lambda_3 & 0 & \lambda_3^2 & 0\\ 0 & \lambda_1\lambda_3 & 0 & \lambda_2\lambda_3 & 0 & \lambda_3^2\end{pmatrix}\begin{pmatrix} f_{x1}\\ f_{y1}\\ f_{x2}\\ f_{y2}\\ f_{x3}\\ f_{y3}\end{pmatrix} dxdy$$

Il ne reste plus pour obtenir le résultat qu'à intégrer. Si $\Gamma_1 \cap S$ est le côté A_1A_2

de T, et en calculant exactement les intégrales, le premier terme donne :

$$\frac{Longueur(A_1A_2)}{6}\begin{pmatrix} 2 & 0 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} g_{x1} \\ g_{y1} \\ g_{x2} \\ g_{y2} \\ g_{x3} \\ g_{y3} \end{pmatrix}$$

Si l'on fait un calcul approché par la formule des trapèzes sur l'arête A_1A_2, on trouve :

$$\frac{Longueur(A_1A_2)}{2}\begin{pmatrix} g_{x1} \\ g_{y1} \\ g_{x2} \\ g_{y2} \\ 0 \\ 0 \end{pmatrix}$$

Pour le deuxième terme (surfacique), dans le cas d'un calcul exact, on obtient :

$$\frac{Aire(T)}{12}\begin{pmatrix} 2 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 0 & 1 \\ 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 & 0 & 1 \\ 1 & 0 & 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 1 & 0 & 2 \end{pmatrix}\begin{pmatrix} f_{x1} \\ f_{y1} \\ f_{x2} \\ f_{y2} \\ f_{x3} \\ f_{y3} \end{pmatrix}$$

Si on utilise la formule d'intégration approchée exacte sur $P1$:

$$\int_T F(x,y)dxdy \approx \frac{Aire(T)}{3}[F(A_1) + F(A_2) + F(A_3)]$$

On obtient une matrice de masse condensée et le second membre surfacique devient :

$$\frac{Aire(T)}{3}\begin{pmatrix} f_{x1} \\ f_{y1} \\ f_{x2} \\ f_{y2} \\ f_{x3} \\ f_{y3} \end{pmatrix}$$

Chapitre 10

Introduction aux problèmes d'évolution : L'équation de la chaleur instationnaire

10.1 Position du problème

La température $u(x,t)$ d'un corps de volume Ω, de densité ρ, de chaleur spécifique c et de conductivité thermique k est régie au cours du temps par l'équation :

$$\rho c \frac{\partial u}{\partial t} = div(k\,\mathbf{grad} u) + f \qquad \forall x \in \Omega \quad \text{et} \quad \forall t \in [0,T]$$

où f représente la puissance volumique fournie au corps Ω.

Si la conductivité k est constante, l'équation se réduit à :

$$\rho c \frac{\partial u}{\partial t} = k\,\Delta u + f \qquad \forall x \in \Omega \quad \text{et} \quad \forall t \in [0,T]$$

Ce problème du premier ordre en temps est le modèle des problèmes paraboliques. La détermination de la solution nécessite de fixer une **condition initiale** en temps : valeur de la température u au temps 0.

$$u(x,0) = u_0(x)$$

On dit que le problème est un problème à valeur initiale ou problème de Cauchy.

D'autre part, un certain nombre de conditions aux limites sur la frontière Γ du domaine peuvent être prises en compte pour déterminer complètement la solution.

— Conditions de type Dirichlet lorsque la température est fixée sur une partie de la frontière
— Conditions de type Neumann si le flux thermique est fixé (nul dans le cas d'un matériau isolé thermiquement)
— Conditions de type Fourier dans le cas le plus général d'un échange convectif avec le milieu extérieur, etc, comme dans le cas stationnaire.

Remarque 10.1.1 (solution stationnaire) *Lorsque la température ne dépend plus du temps (régime permanent ou stationnaire), on retrouve l'équation déjà étudiée :*

$$\begin{cases} -\,div(k\,\mathbf{grad}u) = f \qquad \forall x \in \Omega \\ +\ \textit{Conditions aux limites sur } \Gamma \end{cases}$$

10.2 Étude mathématique de l'équation monodimensionnelle

Nous allons maintenant donner les principales propriétés caractéristiques des problèmes de type paraboliques en nous appuyant, pour simplifier, sur le cas de l'équation de la chaleur monodimensionnelle.

10.2.1 Le modèle de la barre infinie

Considérons tout d'abord le modèle de la barre infinie, sans apport de chaleur et dont la température est initialement donnée. Il est clair que, selon le premier principe de la thermodynamique, la température de la barre doit décroître au cours du temps. Ecrivons l'équation du modèle :

$$\begin{cases} \text{Trouver } u : (x,t) \longrightarrow u(x,t) \text{ telle que :} \\ \dfrac{\partial}{\partial t}\,u(x,t) = \dfrac{\partial^2}{\partial x^2}\,u(x,t) \quad \forall x \in I\!R \text{ et } t \in [0,T] \\ u(x,0) = u_0(x) \quad \text{donnée} \end{cases}$$

Soit $F(\nu,t)$ la transformée de Fourier de u définie par :

$$F(\nu,t) = \int_{-\infty}^{+\infty} exp(-2i\pi\nu x)\,u(x,t)\,dx$$

La transformation de Fourier de l'équation :

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

s'écrit :

$$\frac{\partial}{\partial t} F(\nu, t) + 4\pi^2 \nu^2 F(\nu, t) = 0$$

$F(\nu, 0)$ est la transformée de Fourier de la condition initiale u_0. La résolution de l'équation différentielle du premier ordre en temps ci-dessus donne donc :

$$F(\nu, t) = exp(-4\pi^2 \nu^2 t)\, F(\nu, 0)$$

Or

$$exp(-4\pi^2 \nu^2 t) = F(\frac{1}{\sqrt{4\pi t}}\, exp(-\frac{x^2}{4t}))$$

et

$$F(f * g) = F(f)\, F(g)$$

Donc la transformée de Fourier de la solution u est égale à la transformée du produit de convolution de u_0 et $\frac{1}{\sqrt{4\pi t}}\, exp(-\frac{x^2}{4t})$. Ceci donne, par transformation de Fourier inverse, l'expression classique de la solution u :

$$u(x, t) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{+\infty} exp(-\frac{(x-y)^2}{4t})\, u_0(y)\, dy$$

10.2.2 Propriétés fondamentales de la solution

Nous en déduisons les propriétés fondamentales de la solution u du problème de la chaleur.

1. La solution en un point particulier dépend de la condition initiale en tous les points du domaine. Le domaine de dépendance de la solution s'étend au domaine Ω tout entier.
2. Une perturbation en un point quelconque de la solution initiale influence immédiatement la valeur en tout point de la solution u. On dit que la vitesse de propagation est infinie.
3. La valeur ponctuelle de la solution décroît au cours du temps.
4. t doit être positif. Le phénomène est irréversible, on ne peut pas remonter le temps.
5. Enfin, l'opérateur de la chaleur a un effet régularisant. Pour une condition initiale dans $L^2(\Omega)$, la solution u est C^∞ pour tout temps $t > 0$ (voir un exemple en figure 10.1).

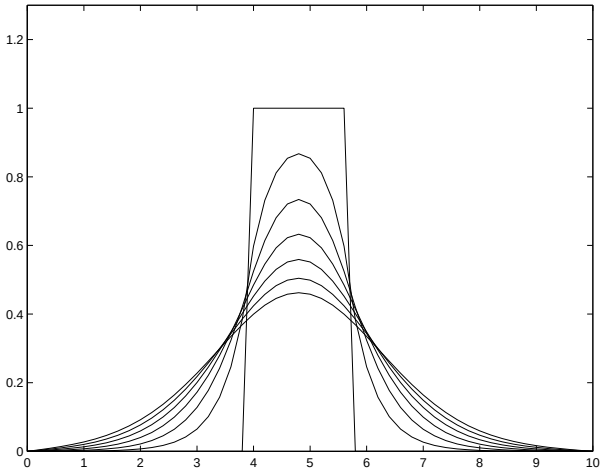


FIGURE 10.1 – Effet régularisant de l'opérateur de la chaleur. On voit ici la suite de solutions en temps à partir d'une condition initiale en fonction porte discontinue

10.2.3 Le modèle de la barre finie avec conditions aux limites de Dirichlet homogènes

On considère une barre de longueur L dont la température est fixée à zéro aux extrémités. L'équation de la température au cours du temps s'écrit :

$$\begin{cases} \dfrac{\partial}{\partial t}\, u(x,t) = \dfrac{\partial^2}{\partial x^2}\, u(x,t) + f(x,t) \quad \forall x \in [0,L] \text{ et } t \in [0,T] \\ u(x,0) = u_0(x) \quad \text{donnée : condition initiale} \\ u(0,t) = u(L,t) = 0 \quad \text{: conditions aux limites de Dirichlet homogènes} \end{cases}$$

Les fonctions ϕ_k définies par

$$\phi_k(x) = sin(\frac{k\pi}{L}x) \quad \text{pour } k = 1, 2, ...n, .. \tag{10.1}$$

sont fonctions propres de l'opérateur

$$-\frac{\partial^2}{\partial x^2} \text{ avec conditions de Dirichlet homogènes}$$

associées aux valeurs propres $\lambda_k = \dfrac{k^2\pi^2}{L^2}$. D'autre part les fonctions ϕ_k forment une famille orthogonale de $L^2[0,L]$.

Exprimons la solution u comme combinaison linéaire des ϕ_k

$$u(x,t) = \sum_k \tilde{u}_k(t)\phi_k(x)$$

et supposons connu un développement de f sous la même forme :

$$f(x,t) = \sum_k \tilde{f}_k(t)\phi_k(x)$$

En reportant ces expressions de u et de f dans l'équation aux dérivées partielles, on obtient un ensemble d'équations différentielles en temps indépendantes pour chaque k.

$$\frac{d\tilde{u}_k}{dt} + \frac{k^2\pi^2}{L^2}\tilde{u}_k = \tilde{f}_k$$

dont la solution s'écrit :

$$\tilde{u}_k(t) = \tilde{u}_k(0)\, exp(-\frac{k^2\pi^2}{L^2}\, t) \; + \; \int_0^t exp(-\frac{k^2\pi^2}{L^2}\, (t-s))\, \tilde{f}_k(s)\, ds$$

Dans le cas particulier $f = 0$, on trouve :

$$\tilde{u}_k(t) = \tilde{u}_k(0)\, exp(-\frac{k^2\pi^2}{L^2}\, t)$$

On admet la convergence dans $L^2[0,L]$ de la série de Fourier de coefficients $\tilde{u}_k(t)$ vers la solution u du problème et ceci $\forall t$. D'où l'expression de la solution

$$u(x,t) = \sum_k \tilde{u}_k(0)\, exp(-\frac{k^2\pi^2}{L^2}\, t)\, sin(\frac{k\pi}{L}x)$$

Remarque 10.2.1 (importante) *La décomposition en modes propres présentée ci-dessus est possible en raison de la propriété essentielle de* **linéarité** *du problème de la chaleur. Elle révèle une propriété fondamentale de l'opérateur de la chaleur : son effet de* **lissage**. *En effet la décroissance des modes en* $exp(-\frac{k^2\pi^2}{L^2}\, t)$ *est d'autant plus rapide que le nombre d'onde* k *est grand. Les hautes fréquences sont donc amorties les premières. Cet effet de lissage a des conséquences positives pour l'approximation numérique, car des erreurs locales d'initialisation ou de calcul, qui correspondront à des modes de fréquence élevée, seront immédiatement amorties.*

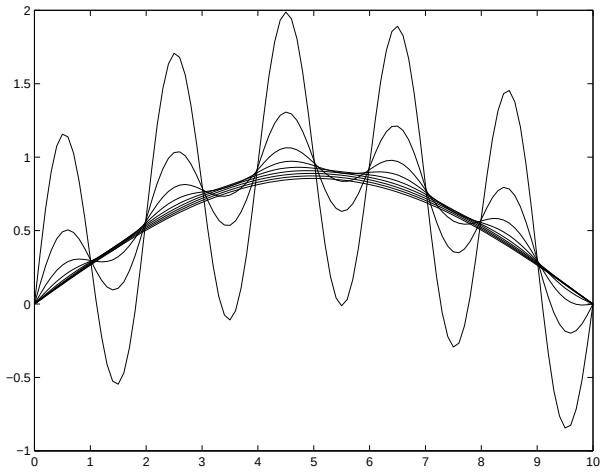


FIGURE 10.2 – Effet de lissage de l'opérateur de la chaleur. La condition initiale est $sin(\pi x) + sin(10\pi x)$. On observe la décroissance immédiate des oscillations en $sin(10\pi x)$ avant que le mode fondamental ne commence à décroître.

10.3 L'équation bi ou tridimensionnelle

Reprenons le problème initial de la chaleur et considérons pour simplifier un problème de Cauchy en temps à conditions aux limites de Dirichlet homogènes en espace.

$$\begin{cases} \dfrac{\partial}{\partial t} u(x,t) = \Delta\, u(x,t) + f(x,t) \quad \forall x \in \Omega \text{ et } t \in [0,T] \\ u(x,0) = u_0(x) \quad \text{donnée : condition initiale} \\ u_{\,|\Gamma}(x,t) = 0 \quad \text{: conditions aux limites de Dirichlet homogènes} \end{cases}$$

10.3.1 Formulation variationnelle

Plaçons-nous, pour simplifier l'exposé, dans le cas d'un domaine plan. La formulation variationnelle du problème s'obtient, comme dans le cas stationnaire, par multiplication par des fonctions tests indépendantes du temps appartenant à l'espace $H_0^1(\Omega)$, compte tenu des conditions aux limites choisies. Après intégration en espace sur le domaine Ω et intégration "par parties" par la formule de Green,

on obtient le problème variationnel :

$$\begin{cases} \text{Trouver pour tout } t \in [0,T], \quad u : (x,y,t) \longrightarrow u(x,y,t) \text{ telle que :} \\ \displaystyle\iint_\Omega \frac{\partial}{\partial t} u(x,y,t)\, v(x,y)\, dxdy + \iint_\Omega \mathbf{grad} u\, \mathbf{grad} v\, dxdy \\ \qquad\qquad = \displaystyle\iint_\Omega f(x,y,t)\, v(x,y)\, dxdy \quad \forall v \in H_0^1(\Omega) \\ u(x,y,0) = u_0(x,y) \quad \text{donnée} \end{cases}$$

On admet le résultat d'existence et d'unicité de la solution de ce problème d'évolution. A chaque instant t, la fonction u considérée comme fonction des variables d'espace x, y appartient alors à l'espace $H_0^1(\Omega)$

Notons $(.,.)$ le produit scalaire de $L^2(\Omega)$, et $a : (u,v) \longrightarrow a(u,v)$ la forme bilinéaire elliptique dans $H_0^1(\Omega)$:

$$a(u,v) = \iint_\Omega \mathbf{grad} u\, \mathbf{grad} v\, dxdy$$

On obtient l'écriture suivante de la formulation variationnelle du problème parabolique ci-dessus :

$$\begin{cases} \text{Trouver pour tout } t \in [0,T], \quad u : (x,y,t) \longrightarrow u(x,y,t) \text{ telle que :} \\ (\dfrac{\partial u}{\partial t}\, ,\, v) + a(u\, ,v) = (f,v) \quad \forall v \in H_0^1(\Omega) \\ u(x,y,0) = u_0(x,y) \quad \text{donnée} \end{cases}$$

On rappelle l'ellipticité de a dans H_0^1. Donc il existe une constante α strictement positive telle que :

$$a(v,v) \geq \alpha \|v\|_{1,2}^2 \qquad \forall v \in H_0^1$$

10.3.2 Propriété de dissipation de l'énergie

En prenant $v = u(x,y,.)$ dans la formulation variationnelle, on obtient

$$\frac{1}{2}\frac{d}{dt}\|u\|_{0,\Omega}^2 + a(u,u) = (f,u)$$

où $\|.\|_{0,\Omega}$ désigne la norme de $L^2(\Omega)$

Utilisons alors l'ellipticité de a et la majoration de la norme de L^2 par la norme H^1

$$a(u,u) \geq \alpha \|u\|_V^2 \geq \alpha \|u\|_{0,\Omega}^2$$

on obtient :

$$\frac{1}{2}\frac{d}{dt}\|u\|^2_{0,\Omega} + \alpha\,\|u\|^2_{0,\Omega} \leq \|f\|_{0,\Omega}\|u\|_{0,\Omega}$$

d'où :

$$\frac{d}{dt}\|u\|_{0,\Omega} + \alpha\,\|u\|_{0,\Omega} \leq \|f\|_{0,\Omega}$$

Multiplions alors les deux membres de l'inégalité par $e^{\alpha t}$, on obtient

$$\frac{d}{dt}[e^{\alpha t}\|u(t)\|_{0,\Omega}] \leq e^{\alpha t}\,\|f\|_{0,\Omega}$$

et en intégrant en temps de 0 à T :

$$\|u(T)\|_{0,\Omega} \leq e^{-\alpha T}\,\|u_0\|_{0,\Omega} + \int_0^T e^{-\alpha(T-t)}\|f(t)\|_{0,\Omega}\,dt$$

On observe à nouveau, dans le cas où f est nulle, c'est à dire en l'absence de source de chaleur, la décroissance en temps de la norme L^2 de la solution. Remarquons que cette décroissance de la norme L^2 de la solution est exponentielle en temps. Cette propriété de décroissance, d'amortissement ou de dissipation d'énergie est caractéristique des problèmes paraboliques. Elle est particulièrement favorable pour l'approximation numérique. En effet, toute perturbation survenant à un instant donné, et en particulier une perturbation due à des erreurs de calcul, est amortie exponentiellement au cours du temps. Il est important que les schémas numériques utilisés respectent ce comportement dissipatif au cours du temps.

Signalons enfin que la technique précédente permet d'obtenir sans difficulté l'unicité de la solution u.

10.4 Étude des schémas de différences finies dans le cas monodimensionnel

10.4.1 Introduction

Une première méthode pour résoudre numériquement les problèmes d'évolution consiste à discrétiser le problème continu par différences finies. Plaçons nous dans le cas monodimensionnel d'une barre de longueur L pour simplifier. On choisit une discrétisation régulière de $[0, L]$ en intervalles de longueur Δx tels que $L = M\Delta x$ et une discrétisation de l'intervalle de temps $[0, T]$ en pas de temps de longueur Δt tels que $T = N\Delta t$. Notons x_j le point $j\Delta x$ et t_n le temps $n\Delta t$. Notons u^n_j la valeur de la solution approchée au point x_j et au temps t_n.

Définition 10.4.1 *Un schéma aux différences finies est dit schéma à p pas en temps si les valeurs u_j^{n+1} des solutions approchées au temps t_{n+1} sont fonctions des valeurs aux p instants précédents, soit aux temps $t_n, t_{n-1}, ...t_{n-p+1}$. En particulier, un schéma est dit à un pas si les u_j^{n+1} ne dépendent que des u_j^n.*

Les deux principales propriétés d'un schéma numérique sont :

l'ordre du schéma qui mesure la précision ou erreur de troncature mathématique commise en remplaçant les dérivées partielles exactes par leurs approximations sous formes de différences divisées. L'ordre est déterminé par des développements de Taylor obtenus en injectant dans l'écriture du schéma numérique la fonction solution continue exacte du problème différentiel.

la stabilité du schéma concerne l'évolution du vecteur des valeurs approchées de la solution aux points x_j au cours des temps t_n (et non plus la solution exacte continue) dans le cas concret où Δt et Δx ne tendent pas vers zéro, mais ont des valeurs fixées. Numériquement, ce critère est relatif à la propagation et l'amplification des erreurs d'arrondis, la condition minimale de stabilité impose que le vecteur de composantes u_j^n reste borné pour tout $n \in [0, N]$. Sinon, il n'est même pas calculable. Si l'on désire de plus que la solution approchée reproduise le comportement de la solution exacte au cours du temps, on devra imposer des conditions plus sévères sur le schéma numérique. Par exemple, dans le cas de l'équation de la chaleur, on cherche à reproduire sur la solution numérique le comportement dissipatif du problème continu. On choisira donc des schémas tels que la solution approchée soit décroissante au cours du temps.

10.4.2 Le Schéma d'Euler explicite

Nous allons préciser les définitions des notions d'ordre et de stabilité en nous appuyant sur l'exemple le plus simple de schéma numérique : le schéma d'Euler explicite (en temps) et centré (en espace).

Considérons le problème

$$\begin{cases} \dfrac{\partial}{\partial t} u(x,t) = \dfrac{\partial^2}{\partial x^2} u(x,t) \quad \forall x \in [0, L] \text{ et } t \in [0, T] \\ u(x,0) = u_0(x) \quad \text{donnée : condition initiale} \\ u(0,t) = u(L,t) = 0 \quad \text{: conditions aux limites de Dirichlet homogènes} \end{cases}$$

et choisissons les approximations classiques suivantes des dérivées première et seconde par différences finies

$$\frac{\partial}{\partial t} u(x_j, t_n) \approx \frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{\Delta t} \quad (\text{ à } O(\Delta t) \text{ près})$$

$$\frac{\partial^2}{\partial x^2}u(x_j,t_n) \approx \frac{u(x_{j+1},t_n)-2u(x_j,t_n)+u(x_{j-1},t_n)}{\Delta x^2} \quad (\text{ à } O(\Delta x^2) \text{ près})$$

Remplaçons les dérivées partielles par leurs approximations en différences finies ci-dessus et la fonction inconnue u par une collection de valeurs discrètes u_j^n pour $j = 0,..M$ et $n = 0,..N$. Nous obtenons un premier exemple de schéma d'approximation en différences finies de l'équation de la chaleur : le schéma d'Euler explicite centré.

$$\begin{cases} \dfrac{u_j^{n+1}-u_j^n}{\Delta t} = \dfrac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2} \\ u_j^0 = u_0(x_j) \quad \text{donnée : condition initiale} \\ u_0^n = u_M^n = 0 \quad \forall n : \text{conditions aux limites de Dirichlet homogènes} \end{cases}$$

Ce schéma est un schéma à un pas, car le vecteur des solutions approchées au temps t_{n+1} ne dépend que des solutions approchées au temps t_n. C'est un schéma explicite car il donne une formule explicite de calcul de la solution au temps t_{n+1} en fonction des valeurs de la solution au temps précédent. Il n'y a pas d'équation à résoudre pour obtenir la valeur au nouvel instant t_{n+1}.

10.4.3 Ordre

Notons $S_{\Delta x,\Delta t}\, u(x_j,t_n)$ l'application d'un schéma aux différences finies à la solution continue u. Par exemple pour le schéma d'Euler explicite centré :

$$S_{\Delta x,\Delta t}\, u(x_j,t_n) = \frac{u(x_j,t_{n+1})-u(x_j,t_n)}{\Delta t} - \frac{u(x_{j+1},t_n)-2u(x_j,t_n)+u(x_{j-1},t_n)}{\Delta x^2}$$

Définition 10.4.2 *Un schéma aux différences finies est d'ordre p en temps et d'ordre q en espace si la différence entre l'équation et le schéma appliqué à la fonction solution du problème continu est un infiniment petit d'ordre p en temps et d'ordre q en espace. C'est à dire si l'on a :*

$$\Big|\, \frac{\partial}{\partial t}\, u(x_j,t_n) \;-\; \frac{\partial^2}{\partial x^2}\, u(x_j,t_n) \;-\; S_{\Delta x,\Delta t}\, u(x_j,t_n)\,]\,\Big| \;=\; O(\Delta t^p) \;+\; O(\Delta x^q)$$

Un schéma **consistant** *est un schéma tel que l'expression ci-dessus tende vers zéro avec Δt et Δx.*

Application : le schéma d'Euler explicite est d'ordre un en temps et d'ordre deux en espace. On montrera en exercice que l'on obtient en effet pour ce schéma

$$\frac{\partial}{\partial t}\,u(x_j,t_n)\;-\;\frac{\partial^2}{\partial x^2}\,u(x_j,t_n)\;-\;[\,\frac{u(x_j,t_{n+1})-u(x_j,t_n)}{\Delta t}\;-\;\frac{u(x_{j+1},t_n)-2u(x_j,t_n)+u(x_{j-1},t_n)}{\Delta x^2}\,]\;=\;-\frac{\Delta t}{2}\frac{\partial^2}{\partial t^2}\,u(x_j,\theta)\;+\;\frac{\Delta x^2}{12}\frac{\partial^4}{\partial x^4}\,u(\xi,t_n)$$

Remarque 10.4.1 *Au point x_j, t_n en dérivant l'équation on a :*

$$\frac{\partial^2}{\partial t^2}\,u(x_j,t_n)\;=\frac{\partial^4}{\partial x^4}\,u(x_j,t_n)$$

on pourrait optimiser l'ordre par un choix de pas de temps et d'espace tel que

$$\frac{\Delta t}{2}=\frac{\Delta x^2}{12}$$

On obtiendrait alors l'ordre 2 en temps et l'ordre 4 en espace. Malheureusement ceci n'est possible que pour des maillages réguliers à pas constants et n'est pas généralisable au cas des éléments finis.

10.4.4 Stabilité

Dans le cas de schémas à un pas appliqués à des problèmes linéaires, le vecteur des solutions approchées au temps t_{n+1} est lié au vecteur des solutions approchées au temps t_n par une relation matricielle. Considérons le problème modèle

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

et appliquons un schéma numérique à un pas. Nous pouvons exprimer le vecteur U^{n+1} des valeurs de la solution au temps t_{n+1} en fonction du vecteur U^n des solutions au temps t_n par :

$$U^{n+1}=C(\Delta t,\Delta x)\,U^n$$

où C est une matrice caractérisant le schéma et dépendant des pas de temps et d'espace.
On en déduit :

$$U^n=C^n\,U^0$$

où U^0 est le vecteur des conditions initiales.

La condition minimale de stabilité s'exprime par le fait que $\|U^n\|$ reste borné quel que soit n. Une condition plus forte impose la décroissance de $\|U^n\|$ quand n augmente.

Condition de stabilité de Von Neumann

Le schéma est stable s'il existe $\tau > 0$ tel que $\|C^n\|$ soit uniformément borné pour tout n et tout Δt vérifiant les conditions :

$$0 < \Delta t < \tau \qquad \text{et} \qquad 0 \leq n\Delta t \leq T$$

Ce critère minimal de stabilité entraîne simplement que la suite U^n ne soit pas explosive. Il est satisfait si l'on a la majoration

$$\|C\| \leq 1 + c\,\Delta t$$

En effet dans ce cas :

$$\|C\| \leq 1 + c\,\Delta t \Longrightarrow \|U^n\| \leq (1 + c\Delta t)^n \,\|U^0\| \leq e^{cn\Delta t}\|U^0\| \leq e^{cT}\|U^0\|$$

U^n reste borné pour tout $n = 0, ..N$. Mais la constante de majoration est exponentielle en la durée d'intégration en temps T et donc devient très grande avec T.

On peut en conséquence préférer des conditions de stabilité plus restrictives telles que :

$$\|C\| \leq 1$$

En effet on a alors

$$\|U^n\| \leq \|U^0\| \quad \forall n = 0, ..N$$

Si l'on veut de plus que la solution numérique reproduise le comportement décroissant de la solution exacte on imposera l'inégalité stricte

$$\|C\| \leq \alpha < 1$$

qui entraîne la décroissance de la norme de U^n.

10.4.5 Étude matricielle de la stabilité

Supposons que le schéma s'exprime sous la forme matricielle présentée plus haut, on déduira la stabilité de majorations de la norme de la matrice C souvent obtenues par le calcul de ses valeurs propres.

Exemple : le schéma d'Euler explicite

Reprenons le problème

$$\begin{cases} \dfrac{\partial}{\partial t}\, u(x,t) = \dfrac{\partial^2}{\partial x^2}\, u(x,t) \quad \forall x \in [0,L] \text{ et } t \in [0,T] \\ \\ u(x,0) = u_0(x) \quad \text{donnée : condition initiale} \\ \\ u(0,t) = u(L,t) = 0 \quad \text{: conditions aux limites de Dirichlet homogènes} \end{cases}$$

et appliquons le schéma d'Euler

$$\begin{cases} \dfrac{u_j^{n+1} - u_j^n}{\Delta t} = \dfrac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} \\ u_j^0 = u_0(x_j) \quad \text{donnée : condition initiale} \\ u_0^n = u_M^n = 0 \quad \forall n \text{ : conditions aux limites de Dirichlet homogènes} \end{cases}$$

On obtient aisément l'écriture matricielle :

$$U^{n+1} = [\, I - \frac{\Delta t}{\Delta x^2}\, A \,]\, U^n$$

où A est la matrice tridiagonale symétrique déjà rencontrée lors de la discrétisation de la dérivée seconde.

$$A = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & \cdots & -1 & 2 \end{pmatrix}$$

Les vecteurs propres de A sont les analogues discrets des fonctions propres ϕ_k (voir équation 10.1). On obtient les vecteurs propres V^k de composantes $V_j^k = sin(\frac{k\pi j}{M})$. Les valeurs propres associées sont :

$$\lambda_k = 4\, sin^2 \frac{k\pi}{2M} \quad \text{pour } k = 1, ...M-1$$

où M dénote le nombre d'intervalles de discrétisation de $[0, L]$ et donc où la dimension de A est égale à $M - 1$.

La matrice $C = I - \dfrac{\Delta t}{\Delta x^2} A$ est une matrice symétrique réelle. Ses vecteurs propres sont ceux de A , ses valeurs propres sont égales à :

$$\mu_k = 1 - \frac{\Delta t}{\Delta x^2} \lambda_k$$

Majorons la norme euclidienne de U^{n+1}

$$\| U^{n+1} \|_2 \leq \| I - \frac{\Delta t}{\Delta x^2} A \|_2 \, \| U^n \|_2$$

Comme la matrice $C = I - \dfrac{\Delta t}{\Delta x^2} A$ est une matrice symétrique, sa norme euclidienne est égale à son rayon spectral

$$\| I - \frac{\Delta t}{\Delta x^2} A \|_2 = \rho(I - \frac{\Delta t}{\Delta x^2} A) = \max_{k=1,..M-1} | 1 - 4 \frac{\Delta t}{\Delta x^2} sin^2(\frac{k\pi}{2M}) |$$

La condition de stabilité $\|C\| \leq 1$ se traduit donc par :

$$\max_{k=1,..M-1} | 1 - 4 \frac{\Delta t}{\Delta x^2} sin^2(\frac{k\pi}{2M}) | \leq 1 \quad \text{soit} \quad 4 \frac{\Delta t}{\Delta x^2} sin^2(\frac{(M-1)\pi}{2M}) \leq 2$$

Ceci sera assuré dès que l'on aura la majoration

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}$$

Cette condition est la condition classique de stabilité du schéma d'Euler explicite pour l'équation de la chaleur. Elle impose des pas de temps très petits, ce qui exclut, dans la plupart des cas, l'usage de ce schéma explicite pour les problèmes paraboliques.

10.4.6 Autres exemples de schémas à un pas

Schéma d'Euler implicite

On considère, pour la discrétisation du même problème, le schéma implicite suivant :

$$\begin{cases} \dfrac{u_j^{n+1} - u_j^n}{\Delta t} = \dfrac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{\Delta x^2} \\ u_j^0 = u_0(x_j) \quad \text{donnée : condition initiale} \\ u_0^n = u_M^n = 0 \quad \forall n : \text{conditions aux limites de Dirichlet homogènes} \end{cases}$$

Ce schéma est dit implicite car le calcul de la solution au pas de temps $n+1$ nécessite la résolution d'un système matriciel.

Ordre du schéma
Un développement de Taylor permet de vérifier simplement que ce schéma est d'ordre un en temps et d'ordre deux en espace comme le schéma explicite.

Stabilité du schéma
La même analyse matricielle conduit au résultat suivant :

$$[\, I + \frac{\Delta t}{\Delta x^2}\, A\,]\, U^{n+1} = U^n$$

La matrice d'itération C est cette fois égale à :

$$C = (\, I + \frac{\Delta t}{\Delta x^2}\ A)^{-1}$$

Ses valeurs propres sont :

$$\mu_k = \frac{1}{1 + \frac{\Delta t}{\Delta x^2}\, \lambda_k}$$

Elles sont donc strictement positives et strictement inférieures à 1 pour tout k. Ce qui entraîne la stabilité inconditionnelle (quels que soient Δt et Δx) du schéma implicite

Observons que l'on a, avec ce schéma, décroissance de la solution approchée au cours des pas de temps.

$$\|U^{n+1}\|_2 \leq \frac{1}{1 + \frac{\Delta t}{\Delta x^2}\, \lambda_1}\, \|U^n\|_2 \quad \text{avec } \lambda_1 = 4\, sin^2(\frac{\pi}{2M})$$

Schéma de Crank-Nicolson ou schéma des trapèzes

On considère, pour la discrétisation du même problème, le schéma implicite suivant :

$$\begin{cases} \dfrac{u_j^{n+1} - u_j^n}{\Delta t} = \dfrac{1}{2}\, \Big[\, \dfrac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} + \dfrac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{\Delta x^2}\, \Big] \\ u_j^0 = u_0(x_j) \quad \text{donnée : condition initiale} \\ u_0^n = u_I^n = 0 \quad \forall n : \text{conditions aux limites de Dirichlet homogènes} \end{cases}$$

Ce schéma est dit implicite car le calcul de la solution au pas de temps $n+1$ nécessite la résolution d'un système matriciel. Il correspond à une intégration en temps approchée selon la formule des trapèzes sur les instants t_n et t_{n+1}.

Ordre du schéma

Ce schéma est d'ordre deux en temps et en espace.

Stabilité du schéma
Le même type d'analyse matricielle que précédemment conduit au résultat suivant :

$$[\, I + \frac{\Delta t}{2\Delta x^2}\, A\,]\, U^{n+1} = [\, I - \frac{\Delta t}{2\Delta x^2}\, A\,] U^n$$

La matrice d'itération C est cette fois égale à :

$$C = (\, I + \frac{\Delta t}{2\Delta x^2}\ A)^{-1} (\, I - \frac{\Delta t}{2\Delta x^2}\ A)$$

Ses valeurs propres sont :

$$\mu_k = \frac{1 - \dfrac{\Delta t}{2\Delta x^2}\, \lambda_k}{1 + \dfrac{\Delta t}{2\Delta x^2}\, \lambda_k}$$

Elles sont donc de module inférieur à 1 pour tout k. Ce qui entraîne la stabilité inconditionnelle (quels que soient Δt et Δx) du schéma de Crank Nicolson.

Remarque 10.4.2 *Attention, la stabilité en norme euclidienne (L^2), considérée ici, n'assure pas la monotonie ou la positivité du schéma. La décroissance en norme L^2 n'implique pas la décroissance de chaque composante du vecteur solution.*

La méthode précédente se complique dans le cas de schémas à plusieurs pas, nous allons présenter ci-dessous une technique de calcul plus simple et adaptable au cas de schémas multipas.

10.4.7 Étude de la stabilité par l'analyse de Fourier

Une technique simple de calcul de la stabilité d'un schéma est donnée dans le cas de problèmes **linéaires** par l'analyse de Fourier. Rappelons l'analyse présentée au paragraphe 10.2.3. Nous avions exprimé la solution u de l'équation de la chaleur

$$\begin{cases} \dfrac{\partial}{\partial t}\, u(x,t) = \dfrac{\partial^2}{\partial x^2}\, u(x,t) \quad \forall x \in [0,L] \text{ et } t \in [0,T] \\ u(x,0) = u_0(x) \quad \text{donnée : condition initiale} \\ u(0,t) = u(L,t) = 0 \quad \text{: conditions aux limites de Dirichlet homogènes} \end{cases}$$

sous la forme du développement :

$$u(x,t) = \sum_{k \in Z} \tilde{u}_k(t) sin(\frac{k\pi}{L} x)$$

En utilisant de nouveau la linéarité du problème et le principe de superposition, nous observons que la solution, dans le cas de conditions aux limites linéaires quelconques, Dirichlet, Neumann, Fourier ou périodiques s'écrit sous la forme générale :

$$u(x,t) = \sum_{k \in Z} \tilde{u}_k(t) e^{ikx}$$

où k est un coefficient réel intégrant le nombre d'onde, le type de conditions aux limites et la dimension du domaine. Les coefficients de Fourier $\tilde{u}_k$ vérifient chacun une équation différentielle en temps dont la solution s'écrit :

$$\tilde{u}_k(t) = \tilde{u}_k(0)\, exp(-k^2\, t)$$

On a donc

$$\tilde{u}_k(t + \Delta t) = exp(-k^2\, \Delta t)\, \tilde{u}_k(t)$$

Faisons la même analyse dans le cas discret. Injectons dans le schéma numérique une suite de solutions de la forme

$$u_j^n = \tilde{u}_k^n\, e^{ikj\Delta x}$$

Ces solutions ont pour conditions initiales

$$u_j^0 = \tilde{u}_k^0\, e^{ikj\Delta x}$$

et correspondent chacune à une composante harmonique. L'étude de la stabilité se ramène à l'étude de l'évolution au cours des pas de temps n des suites $\tilde{u}_k^n$ quand n augmente. La condition minimale de stabilité numérique nécessite que les nombres $\tilde{u}_k^n$ restent bornés $\forall\, k$ et $\forall\, n = 0, ...N$. Si l'on veut de plus décroissance de la solution approchée, on devra avoir décroissance des $\tilde{u}_k^n$ quand n augmente.

Dans les schémas à p pas, on obtient les $\tilde{u}_k^{n+1}$ par multiplication par une matrice d'amplification $G(\Delta t, k)$ selon :

$$\begin{pmatrix} \tilde{u}_k^{n+1} \\ \tilde{u}_k^n \\ : \\ : \\ \tilde{u}_k^{n-p+2} \end{pmatrix} = G(\Delta t, k) \begin{pmatrix} \tilde{u}_k^n \\ \tilde{u}_k^{n-1} \\ : \\ : \\ \tilde{u}_k^{n-p+1} \end{pmatrix}$$

Dans les schémas à un pas, la matrice d'amplification se réduit à un facteur d'amplification $G(\Delta t, k)$ tel que $\tilde{u}_k^{n+1} = G_k(\Delta t)\, \tilde{u}_k^n$.

Nous obtenons alors les conditions de stabilité suivantes :

Condition nécessaire de stabilité de Von Neumann

Pour que le schéma soit stable, il faut qu'il existe $\tau > 0$ tel que les valeurs propres λ_i de la matrice d'amplification $G(\Delta t, k)$ soient toutes majorées en module selon :

$$|\lambda_i| \leq 1 + c\Delta t \quad \forall i = 1, ..p$$

avec $c > 0$, quel que soit k et pour tout $0 < \Delta t < \tau$

Dans le cas de schéma à un pas la matrice d'amplification se réduit à un facteur scalaire et la condition de Von Neuman est suffisante.

Conditions suffisantes de stabilité

1) Si la matrice d'amplification est normale, c'est à dire qu'elle commute avec son adjointe (ou transposée dans le cas réel)

$$G\, G^* = G^*\, G$$

ou bien, ce qui est équivalent, si elle admet une base de vecteurs propres orthonormés, alors la condition de Von Neumann est suffisante.

2) La condition précédente étant parfois difficile à vérifier, on peut utiliser la condition suffisante suivante : le schéma est stable si les coefficients de la matrice $G(\Delta t, k)$ sont bornés et si ses valeurs propres sont toutes de module strictement inférieur à 1 sauf éventuellement une de module égal à 1.

Un premier exemple simple d'application : le schéma d'Euler

Posons

$$u_j^n = \tilde{u}_k^n\, e^{ikj\Delta x}$$

et calculons le facteur d'amplification $G_k(\Delta t)$ tel que $\tilde{u}_k^{n+1} = G_k(\Delta t)\, \tilde{u}_k^n$. On obtient :

$$\frac{\tilde{u}_k^{n+1} - \tilde{u}_k^n}{\Delta t} exp(ikj\Delta x) = \frac{exp(ik\Delta x) - 2 + exp(-ik\Delta x)}{\Delta x^2} exp(ikj\Delta x)\, \tilde{u}_k^n$$

soit :

$$\tilde{u}_k^{n+1} = [1 + \frac{\Delta t}{\Delta x^2}(2\, cos(k\Delta x) - 2]\, \tilde{u}_k^n \;=\; [\, 1 - 4\, \frac{\Delta t}{\Delta x^2} sin^2(\frac{k}{2}\Delta x)\,]\, \tilde{u}_k^n$$

La condition $|G_k(\Delta t)| \leq 1 \quad \forall k$ nécessite $\dfrac{\Delta t}{\Delta x^2} \leq \dfrac{1}{2}$. On retrouve évidemment des calculs analogues et le même résultat que par l'analyse matricielle faite plus haut.

Un exemple simple de schéma implicite à 2 pas : le schéma de Gear

Considérons le schéma suivant pour l'équation de la chaleur monodimensionnelle :

$$\frac{3}{2}\,u_j^{n+1} - 2u_j^n + \frac{1}{2}\,u_j^{n-1} \;=\; \frac{\Delta t}{\Delta x^2}\,[\,u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}\,]$$

Posons comme précédemment

$$u_j^n = \tilde{u}_k^n\, e^{ikj\Delta x}$$

Un calcul simple conduit au résultat suivant

$$\begin{pmatrix} \tilde{u}_k^{n+1} \\ \tilde{u}_k^n \end{pmatrix} = \begin{pmatrix} \frac{2}{a} & -\frac{1}{a} \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \tilde{u}_k^n \\ \tilde{u}_k^{n-1} \end{pmatrix}$$

avec $a = \dfrac{3}{2} + 4\,\dfrac{\Delta t}{\Delta x^2}\,sin^2(\dfrac{k}{2}\Delta x)$

Les valeurs propres de la matrice 2×2 d'amplification ci-dessus sont racines de

$$\lambda^2 - \frac{2}{a}\,\lambda + \frac{1}{2a} = 0$$

On trouve le discriminant $\Delta' = \dfrac{2-a}{2a^2}$. On obtient si $a > 2$ deux racines complexes conjuguées de module $\sqrt{\dfrac{1}{2a}} < 1$ et dans le cas $a \leq 2$ deux racines réelles dont la plus grande en valeur absolue vaut

$$\frac{1}{a} + \sqrt{\frac{2-a}{2a^2}} < 1$$

On a donc stabilité inconditionnelle de ce schéma qui se révèle dans la pratique particulièrement adapté dans le cas d'équations "raides", c'est à dire dans lesquelles on aurait de fortes variations locales des constantes thermiques.

10.5 Méthodes d'éléments finis pour le problème de la chaleur

On considère le problème bidimensionnel suivant :

$$\begin{cases} \dfrac{\partial}{\partial t}\,u(x,y,t) = \Delta\,u(x,y,t) + f(x,y,t) \quad \forall x,y \in \Omega \text{ et } t \in [0,T] \\ u(x,y,0) = u^0(x,y) \quad \text{donnée : condition initiale} \\ u_{\,|\Gamma_0} = u_d \quad \text{: conditions aux limites de Dirichlet} \\ \dfrac{\partial u}{\partial n}_{\,|\Gamma_1} = g \quad \text{: conditions aux limites de Neumann} \end{cases}$$

Formulation variationnelle

La formulation variationnelle du problème s'obtient comme dans le cas stationnaire par multiplication par des fonctions tests indépendantes du temps appartenant, compte tenu des conditions aux limites choisies, à l'espace V des fonctions de $H^1(\Omega)$ nulles sur Γ_0. Après intégration en espace sur le domaine Ω et intégration "par parties" par la formule de Green, on obtient le problème variationnel :

$$\left\{\begin{array}{l} \text{Trouver pour tout } t \in [0,T], \quad u : (x,y,t) \longrightarrow u(x,y,t) \text{ telle que :} \\ \displaystyle\iint_\Omega \frac{\partial}{\partial t} u(x,y,t)\, v(x,y)\, dxdy + \iint_\Omega \mathbf{grad} u\, \mathbf{grad} v\, dxdy \\ \qquad\qquad = \displaystyle\iint_\Omega f(x,y,t)\, v(x,y)\, dxdy + \int_{\Gamma_1} g\, v\, d\gamma \quad \forall v \in V \\ u(x,y,0) = u_0(x,y) \quad \text{donnée} \end{array}\right.$$

10.5.1 Semi-discrétisation en espace par éléments finis

On suppose réalisé un maillage du domaine Ω en éléments finis triangulaires P_k ou quadrangulaires Q_k. Soient w_i les fonctions de base associées aux éléments finis choisis. Le maillage est pris fixe au cours du temps (voir au chapitre 13 une situation dans laquelle le domaine est déformable et donc le maillage change au cours du temps). Les fonctions de base w_i sont donc indépendantes du temps. Notons $\mathbf{I}$ l'ensemble des indices des noeuds du maillage correspondant à une valeur inconnue de la solution u. C'est à dire ici l'ensemble des noeuds n'appartenant pas à Γ_0. Notons $\mathbf{J}$ l'ensemble des indices des sommets du maillage correspondant à une valeur connue de la solution, donc ici appartenant à Γ_0.

Comme dans le cas stationnaire nous décomposerons la solution approchée u_h en somme d'une inconnue auxiliaire $\tilde{u}_h$ et d'une fonction connue u_0 prenant les valeurs imposées sur Γ_0. La solution auxiliaire $\tilde{u}_h$ s'écrira dans la base des w_i pour $i \in \mathbf{I}$ selon :

$$\tilde{u}_h(x,y,t) = \sum_{i \in \mathbf{I}} u_i(t)\, w_i(x,y)$$

La fonction u_0, supposée indépendante du temps pour simplifier, sera approchée par une fonction $u_{0,h}$ prenant les valeurs imposées sur Γ_0 et nulle sur tous les noeuds d'indices $i \in \mathbf{I}$

$$u_{0,h}(x,y) = \sum_{i \in \mathbf{J}} u_d(x_i,y_i)\; w_i(x,y)$$

On en déduit

$$\frac{\partial}{\partial t}u_h(x,y,t) = \sum_{i\in\ \mathbf{I}} u_i'(t)\, w_i(x,y)$$

Notons $N_{\mathbf{I}}$ le nombre de noeuds "inconnus" d'indices $i \in\ \mathbf{I}$. Le problème approché s'écrit sous la forme d'un système différentiel linéaire de $N_{\mathbf{I}}$ équations à $N_{\mathbf{I}}$ fonctions inconnues du temps u_i.

$$\left\{\begin{array}{l} \text{Trouver } \forall t \in [0,T],\ \text{et } \forall j \in \mathbf{I},\ \text{les fonctions } u_j(t)\ \text{telles que, } \forall\, i \in\ \mathbf{I}: \\ \displaystyle\sum_{j\in\ \mathbf{I}} \Big(\iint_\Omega w_j\, w_i\, dxdy \Big) u_j'(t) + \Big(\iint_\Omega \mathbf{grad} w_j\, \mathbf{grad} w_i\, dxdy \Big) u_j(t) = \\ \displaystyle\iint_\Omega f\, w_i\, dxdy + \int_{\Gamma_1} g\, w_i\, d\gamma\ - \sum_{j\in\ \mathbf{J}} \Big(\iint_\Omega \mathbf{grad} w_j\, \mathbf{grad} w_i\, dxdy \Big) u_d(x_j,y_j) \\ u_i(0) = u_{i,0} \quad \text{donnés} \quad \forall\, i \in\ \mathbf{I} \end{array}\right.$$

Ce qui donne matriciellement :

$$\left\{\begin{array}{l} \text{Trouver pour tout } t \in [0,T],\ \text{le vecteur } U(t) \text{ tel que :} \\ MU'(t) + KU(t) = B \\ U(0) = U^0 \quad \text{donné} \end{array}\right.$$

avec M matrice de masse de coefficients

$$m_{i,j} = \iint_\Omega w_j\, w_i\, dxdy$$

K matrice de raideur de coefficients

$$k_{i,j} = \iint_\Omega \mathbf{grad} w_j\, \mathbf{grad} w_i\, dxdy$$

et B vecteur second membre dont les coefficients sont dans ce cas égaux à

$$B_i = \iint_\Omega f(x,y,t)\, w_i\, dxdy + \int_{\Gamma_1} g\, w_i\, d\gamma\ - \sum_{j\in\ \mathbf{J}} k_{i,j}\, u_d(x_j,y_j)$$

10.5.2 Discrétisation complète en espace et en temps

Il nous reste à appliquer les schémas en temps déjà présentés dans le cas de discrétisations en différences finies pour obtenir une discrétisation complète du problème.

Schéma d'Euler explicite

On utilise l'approximation

$$U'(t) \approx \frac{U(t+\Delta t) - U(t)}{\Delta t}$$

ce qui conduit au schéma

$$\begin{cases} \text{Trouver pour tout } n \in [0, N], \text{ la suite de vecteurs } U^n \text{ tels que :} \\ U^0 \text{ donné } : (\, U_i^0 = u^0(x_i, y_i)\,) \\ MU^{n+1} = MU^n - \Delta t K U^n + \Delta t B \end{cases}$$

Remarquons que la dépendance du second membre B par rapport au temps ne poserait pas de problème difficile.

La résolution complète du problème approché nécessite la résolution d'un système matriciel à chaque pas de temps. Nous avons deux possibilités :

1) On factorise une fois pour toute en début de calcul la matrice de masse M qui est symétrique définie positive sous forme LL^T et on a deux systèmes triangulaires à résoudre à chaque pas de temps.

2) On calcule la matrice de masse de façon approchée sous forme d'une matrice de masse condensée diagonale (mass lumping). L'inversion de la matrice de masse est alors immédiate et on obtient véritablement un schéma numérique explicite.

Malheureusement le schéma d'Euler explicite dont la stabilité dépend d'une condition très sévère sur le pas de temps n'est pas adapté à l'équation de la chaleur. On préfèrera utiliser les schémas implicites suivants.

Schéma d'Euler implicite

On utilise l'approximation

$$U'(t) \approx \frac{U(t) - U(t-\Delta t)}{\Delta t}$$

ce qui conduit au schéma

$$\begin{cases} \text{Trouver pour tout } n \in [0, N], \text{ la suite de vecteurs } U^n \text{ tels que :} \\ U^0 \text{ donné } : (\, U_i^0 = u^0(x_i, y_i)\,) \\ MU^{n+1} = MU^n - \Delta t K U^{n+1} + \Delta t B \end{cases}$$

Soit

$$[M + \Delta t K]U^{n+1} = MU^n + \Delta t B$$

Dans ce cas, que l'on ait ou non condensé la matrice de masse sous forme diagonale, on doit résoudre un système matriciel. Ce que l'on fait en factorisant une fois pour toutes au début du calcul, la matrice $M + \Delta t K$ qui est symétrique définie positive, sous forme LL^T puis en résolvant à chaque pas deux systèmes triangulaires.

Schéma de Crank-Nicolson

Le schéma s'écrit

$$\begin{cases} \text{Trouver pour tout } n \in [0, N], \text{ la suite de vecteurs } U^n \text{ tels que :} \\ U^0 \text{ donné } : (\, U_i^0 = u^0(x_i, y_i)\,) \\ MU^{n+1} = MU^n - \dfrac{\Delta t}{2}\, K[U^n + U^{n+1}] + \Delta t B \end{cases}$$

Soit

$$[M + \frac{\Delta t}{2} K]U^{n+1} = [M - \frac{\Delta t}{2} K]U^n + \Delta t B$$

Il est facile de montrer, en reprenant l'analyse matricielle faite en 10.4.6, la stabilité inconditionnelle de ce schéma.

Schéma de Gear

Dans les cas difficiles, en particulier le cas d'équations "raides", si l'on veut l'ordre deux de précision, on utilisera le schéma à deux pas implicite suivant, dit schéma de Gear :

$$\begin{cases} \text{Trouver pour tout } n \in [0, N], \text{ la suite de vecteurs } U^n \text{ tels que :} \\ U^0 \text{et } U^1 \text{ donnés} \\ \dfrac{3}{2} MU^{n+1} = 2MU^n - \dfrac{1}{2} MU^{n-1} - \Delta t\; KU^{n+1} + \Delta t B \end{cases}$$

Soit

$$[\frac{3}{2} M + \Delta t K]U^{n+1} = 2MU^n - \frac{1}{2} MU^{n-1} + \Delta t B$$

Ce schéma, inconditionnellement stable et du second ordre en temps, nécessite au démarrage l'utilisation d'un schéma à un pas (Crank-Nicolson afin de conserver l'ordre 2) pour le calcul de U^1. Les schémas de Runge et Kutta implicites sont un autre choix possible de schémas d'ordre élevé stables avec l'avantage pratique d'être des schémas à un pas.

Chapitre 11

Introduction aux problèmes hyperboliques du second ordre : L'équation des ondes

11.1 Position du problème

Considérons une membrane élastique de surface Ω, plane au repos et fixée sur son bord Γ. Lors de petites vibrations transversales, le déplacement normal au plan d'équilibre en tout point x, y de Ω et à chaque instant t est une fonction $u : x, y, t \longrightarrow u(x, y, t)$ qui vérifie l'équation :

$$\frac{\partial^2 u}{\partial t^2} = c^2 \, \Delta u + f \qquad \forall (x, y) \in \Omega \quad \text{et} \quad \forall t \in [0, T]$$

où c désigne la vitesse des ondes. Ce problème du second ordre en temps est un modèle de problème hyperbolique. La détermination de la solution nécessite de fixer deux **conditions initiales** en temps. En fixant les valeurs du déplacement transversal u et de sa dérivée partielle en temps (voir la remarque 3.6.1), au temps 0 :

$$\begin{cases} u(x, y, 0) = u^0(x, y) \\ \dfrac{\partial}{\partial t} u(x, y, 0) = u^1(x, y) \end{cases}$$

on obtient un problème à valeur initiale ou problème de Cauchy.

Les conditions aux limites choisies, pour la membrane fixée sur son bord Γ, sont des conditions de Dirichlet homogènes mais on pourrait choisir d'autres types de conditions aux limites comme dans les cas stationnaires ou paraboliques.

Remarque 11.1.1 (solution stationnaire) *Lorsque la solution ne dépend plus du temps (régime permanent ou stationnaire) on retrouve une équation déjà*

étudiée de forme :

$$\begin{cases} -\Delta u = f \qquad \forall (x,y) \in \Omega \\ + \textit{ Conditions aux limites sur } \Gamma \end{cases}$$

L'équation des ondes et l'équation de la chaleur ont les mêmes expression dans le cas stationnaire. C'est pourquoi les solutions du problème de Poisson ci-dessus peuvent s'interpréter physiquement, à la fois comme des déplacements d'une membrane élastique ou des températures.

11.2 Étude mathématique de l'équation monodimensionnelle

Nous allons maintenant donner les principales propriétés caractéristiques des problèmes de type hyperboliques en nous appuyant, pour simplifier, sur le cas de l'équation des ondes monodimensionnelle ou équation de la corde vibrante.

11.2.1 Le modèle de la corde infinie

Considérons tout d'abord le modèle de la corde infinie, libre de toute sollicitation. On se donne la position et la vitesse initiale au temps zéro. Ecrivons l'équation du modèle :

$$\begin{cases} \text{Trouver } u : (x,t) \longrightarrow u(x,t) \text{ telle que :} \\ \dfrac{\partial^2}{\partial t^2} u(x,t) = c^2 \dfrac{\partial^2}{\partial x^2} u(x,t) \quad \forall x \in I\!R \text{ et } t \in [0,T] \\ u(x,0) = u^0(x) \quad \text{donnée} \\ \dfrac{\partial}{\partial t} u(x,0) = u^1(x) \quad \text{donnée} \end{cases}$$

Posons

$$y = x + ct \qquad z = x - ct$$

En remplaçant dans l'équation les dérivées partielles par rapport à t et x par leurs expressions en fonctions des dérivées partielles par rapport aux nouvelles variables y et z, on obtient

$$\frac{\partial^2 u}{\partial y \partial z} = 0$$

d'où l'on déduit :

$$\frac{\partial u}{\partial z} = g_1(z)$$

soit :

$$u = g(z) + f(y)$$

Ce qui donne en définitive l'expression remarquable suivante :

$$u(x,t) = f(x+ct) + g(x-ct)$$

Prenant en compte les conditions initiales

$$u(x,0) = f(x) + g(x) = u^0(x) \qquad \frac{\partial}{\partial t}u(x,0) = cf'(x) - cg'(x) = u^1(x)$$

on obtient :

$$f(x) = \frac{1}{2}u^0(x) + \frac{1}{2c}\int_0^x u^1(s)ds$$

et

$$g(x) = \frac{1}{2}u^0(x) - \frac{1}{2c}\int_0^x u^1(s)ds$$

d'où :

$$u(x,t) = f(x+ct) + g(x-ct) = \frac{1}{2}[u^0(x+ct) + u^0(x-ct)] + \frac{1}{2c}\int_{x-ct}^{x+ct} u^1(s)ds$$

11.2.2 Propriétés fondamentales de la solution

1. L'expression $f(x+ct)$ prend au point x et à l'instant t la même valeur qu'au point $x+ct$ au temps zéro. De même $g(x-ct)$ prend au point x et à l'instant t la valeur qu'elle prenait au point $x-ct$ au temps zéro. La solution u au point x et au temps t apparaît comme la somme de deux ondes, l'une f se propageant avec une vitesse $-c$, donc de droite à gauche, l'autre g se propageant avec une vitesse c donc de gauche à droite. c apparaît comme une vitesse de propagation d'onde (mais non comme la vitesse des points de la corde).

2. Domaine de dépendance : la solution au point x et au temps t ne dépend que des valeurs des conditions initiales u^0 et u^1 aux points de l'intervalle $[x-ct, x+ct]$. Inversement les conditions initiales au temps zéro en un point ξ n'influenceront la solution aux instants t que pour les seules abscisses x comprises entre $\xi-ct$ et $\xi+ct$. La figure 11.1 ci-dessous illustre simplement cette notion fondamentale de domaine de dépendance.

   Les droites $x - ct = cste$ et $x + ct = cste$ s'appellent les droites caractéristiques.

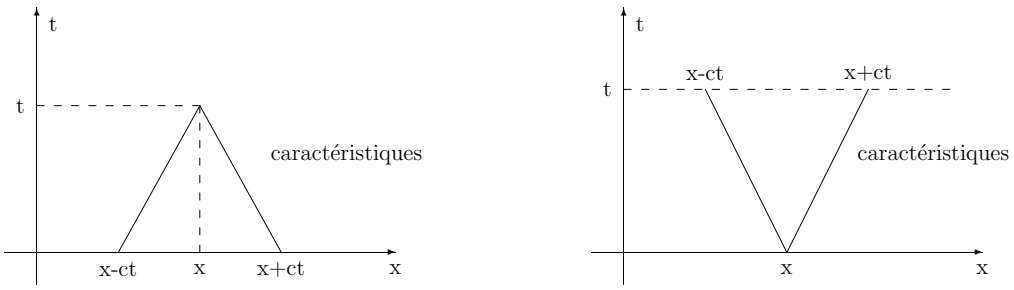


FIGURE 11.1 – Domaine de dépendance et droites caractéristiques.

3. Si $u^1 = 0$ la solution ne dépend que des valeurs de u^0 aux points $x - ct$ et $x + ct$. Une perturbation en un point quelconque de la solution initiale se transmet pour moitié vers la droite avec une vitesse c et pour moitié vers la gauche avec la même vitesse absolue c car dans ce cas

$$u(x,t) = \frac{1}{2}[u^0(x - ct) + u^0(x + ct)]$$

Il est alors facile d'obtenir une solution à partir de conditions initiales simples par transport et en particulier d'observer l'évolution de solutions initiales discontinues de type échelon.

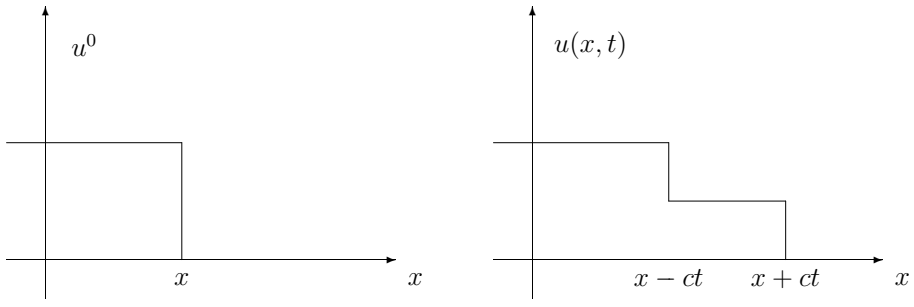


FIGURE 11.2 – Évolution d'une solution initiale de type échelon.

4. t peut être positif ou négatif. Le phénomène est réversible, on peut remonter le temps et retrouver, par résolution de problèmes rétrogrades, la solution à un instant précédent.

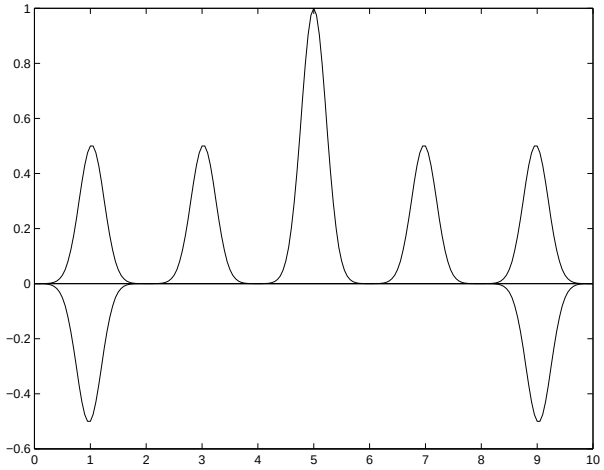


FIGURE 11.3 – Évolution au cours du temps de la solution de l'équation de la corde vibrante avec une initialisation en forme de gaussienne

11.2.3 Le modèle de la corde vibrante finie

On considère une corde de longueur L fixée aux extrémités. L'équation du déplacement transversal au cours du temps s'écrit :

$$\begin{cases} \dfrac{\partial^2}{\partial t^2} u(x,t) = c^2 \dfrac{\partial^2}{\partial x^2} u(x,t) \quad \forall x \in [0,L] \text{ et } t \in [0,T] \\ u(x,0) = u^0(x) \quad \dfrac{\partial}{\partial t} u(x,0) = u^1(x) \text{ données : conditions initiales} \\ u(0,t) = u(L,t) = 0 \quad \text{: conditions aux limites de Dirichlet homogènes} \end{cases}$$

Symétrie, périodicité et réflexion aux bornes

Reprenons la forme générale abstraite de la solution générale

$$u(x,t) = f(x+ct) + g(x-ct) = \frac{1}{2}[u^0(x+ct) + u^0(x-ct)] + \frac{1}{2c}\int_{x-ct}^{x+ct} u^1(s)ds$$

Au point $x = 0$, on a donc : $f(ct) + g(-ct) = 0$ pour tout t. Les fonctions f et g doivent donc vérifier la condition générale : $g(s) = -f(-s) \quad \forall\, s \in I\!R$

On en déduit simplement que la solution u doit être une fonction impaire de x. En effet

$$u(x,t) = f(x+ct) + g(x-ct) = f(x+ct) - f(-x+ct)$$

$$u(-x,t) = f(-x+ct) + g(-x-ct) = f(-x+ct) - f(x+ct)$$

d'où

$$u(-x,t) = -u(x,t) \quad \forall\, x \text{ et } t$$

Ceci entraîne la réflexion avec changement de signe des ondes lorsqu'elles atteignent une extrémité fixe de la corde. Signalons que cette propriété de réflexion sur une frontière Dirichlet introduit une difficulté importante dans la modélisation numérique de propagation d'ondes en milieu infini. En effet on est, dans la pratique, obligé de se limiter à des domaines bornés. Si l'on impose sur les frontières à "l'infini" des conditions de Dirichlet, la réflexion sur ces frontières va entraîner l'apparition d'ondes réfléchies parasites qui vont gravement perturber la solution. De nombreux travaux de recherches sur des conditions aux limites "transparentes" ou "absorbantes" ont été développés pour surmonter cette difficulté. Citons également la méthode PML (perfectly matched layer) qui consiste à ajouter une couche de matériau absorbant à l'infini, ce qui évite le développement de conditions aux limites particulières.

Au point $x = L$ on a : $f(L+ct) + g(L-ct) = 0$ pour tout t. La fonction f doit donc vérifier la condition générale : $f(L+s) = f(-L+s) \quad \forall\, s \in I\!R$. f est donc une fonction périodique de période $2L$. On en déduit que la solution u doit être une fonction périodique en x de période $2L$ et en t de période $\dfrac{2L}{c}$. En effet

$$u(x+2L,t) = f(x+2L+ct) - f(-x+ct+2L) = f(x+ct) - f(-x+ct) = u(x,t)$$

$$u(x,t+\frac{2L}{c}) = f(x+ct+2L) - f(-x-2L+ct) = f(x+ct) - f(-x+ct) = u(x,t)$$

En conséquence, on peut très simplement obtenir la valeur de la solution en un point x et un instant t à partir de la solution initiale au temps zéro complétée par symétrie et périodicité.

Analyse de Fourier

Reprenons les fonctions ϕ_k définies par

$$\phi_k(x) = sin(\frac{k\pi}{L}x) \quad \text{pour } k = 1, 2, ...n, ..$$

fonctions propres de l'opérateur $-\dfrac{\partial^2}{\partial x^2}$ avec conditions de Dirichlet homogènes associées aux valeurs propres $\lambda_k = \dfrac{k^2\pi^2}{L^2}$.

Exprimons la solution u comme combinaison linéaire des ϕ_k

$$u(x,t) = \sum_k \tilde{u}_k(t)\phi_k(x)$$

En reportant cette expression de u dans l'équation aux dérivées partielles, on obtient un ensemble d'équations différentielles du second ordre en temps indépendantes pour chaque k.

$$\frac{d^2\tilde{u}_k}{dt^2} + \frac{k^2\pi^2}{L^2}\tilde{u}_k = 0$$

Dans ce cas sans second membre, la solution s'écrit sous la forme générale :

$$\tilde{u}_k(t) = A_k\, cos(\frac{k\pi}{L}ct) + B_k\, sin(\frac{k\pi}{L}ct)$$

On trouve donc :

$$u(x,t) = \sum_k [\, A_k\, cos(\frac{k\pi}{L}ct) + B_k\, sin(\frac{k\pi}{L}ct)\,]\, sin(\frac{k\pi}{L}x)$$

Remarquons que l'on retrouve ainsi le caractère impair en x et périodique en x et t de la solution.

11.3 L'équation bidimensionnelle

Reprenons le problème initial de membrane vibrante et considérons pour simplifier un problème de Cauchy en temps à conditions aux limites de Dirichlet homogènes en espace.

$$\begin{cases} \dfrac{\partial^2}{\partial t^2}\, u(x,y,t) = c^2 \Delta\, u(x,y,t) + f(x,y,t) \quad \forall x,y \in \Omega \text{ et } t \in [0,T] \\ u(x,y,0) = u^0(x,y) \quad \dfrac{\partial}{\partial t} u(x,y,0) = u^1(x,y) : \text{conditions initiales} \\ u_{\,|\Gamma}(x,y,t) = 0 \quad : \text{conditions aux limites de Dirichlet homogènes} \end{cases}$$

11.3.1 Formulation variationnelle

La formulation variationnelle du problème s'obtient comme dans le cas stationnaire par multiplication par des fonctions tests indépendantes du temps appartenant à l'espace $H_0^1(\Omega)$, compte tenu des conditions aux limites choisies. Après intégration en espace sur le domaine Ω et intégration "par parties" par la formule

de Green, on obtient le problème variationnel :

$$\begin{cases} \text{Trouver pour tout } t \in [0,T], \quad u : (x,y,t) \longrightarrow u(x,y,t) \text{ telle que :} \\ \displaystyle\iint_\Omega \frac{\partial^2}{\partial t^2}\, u(x,y,t)\, v(x,y)\, dxdy + c^2 \iint_\Omega \mathbf{grad} u\, \mathbf{grad} v\, dxdy \\ \qquad\qquad = \displaystyle\iint_\Omega f(x,y,t)\, v(x,y)\, dxdy \quad \forall v \in H_0^1(\Omega) \\ u(x,y,0) = u^0(x,y) \quad \dfrac{\partial}{\partial t} u(x,y,0) = u^1(x,y) \end{cases}$$

On a existence et unicité de la solution de ce problème d'évolution. A chaque instant t, la fonction u considérée comme fonction des variables d'espace x, y appartient alors à l'espace $H_0^1(\Omega)$

11.3.2 Conservation de l'énergie

Considérons le cas d'une équation homogène (second-membre $f = 0$). En prenant $v = \dfrac{\partial u}{\partial t}$ dans la formulation variationnelle on obtient :

$$\iint_\Omega \frac{\partial^2 u}{\partial t^2}\, \frac{\partial u}{\partial t}\, dxdy + c^2 \iint_\Omega \mathbf{grad}(u)\, \mathbf{grad}(\frac{\partial u}{\partial t})\, dxdy = 0$$

D'où

$$\frac{1}{2}\frac{d}{dt}\|\frac{\partial u}{\partial t}\|^2_{0,\Omega} + \frac{c^2}{2}\frac{d}{dt}\|\mathbf{grad} u\|^2_{0,\Omega} = 0$$

où $\|.\|_{0,\Omega}$ désigne la norme de $L^2(\Omega)$

On en déduit la conservation de l'énergie au cours du temps sous la forme :

$$\|\frac{\partial u}{\partial t}\|^2_{0,\Omega} + c^2\|\mathbf{grad} u\|^2_{0,\Omega} = cste$$

Cette propriété de conservation de l'énergie est fondamentale dans le cas des problèmes hyperboliques. L'absence d'amortissement ou de dissipation d'énergie est une des causes principales de la difficulté de la résolution numérique des problèmes hyperboliques. Il est souhaitable que les schémas numériques utilisés respectent ce comportement conservatif au cours du temps. Or la recherche de conservation exacte entraîne le risque d'explosion numérique. On doit donc parfois accepter une légère dissipation dans les schémas numériques pour en assurer la stabilité.

11.4 Étude des schémas de différences finies dans le cas monodimensionnel

11.4.1 Première approche : discrétisation directe de l'équation du second ordre

Une première méthode pour résoudre numériquement ce problème d'évolution consiste à discrétiser l'équation du second ordre par différences finies. Plaçons nous dans le cas monodimensionnel d'une corde de longueur L pour simplifier. On choisit une discrétisation régulière de $[0, L]$ en intervalles de longueur Δx tels que $L = M\Delta x$ et une discrétisation de l'intervalle de temps $[0, T]$ en pas de temps de longueur Δt tels que $T = N\Delta t$. Notons x_j le point $j\Delta x$ et t_n le temps $n\Delta t$. Notons u_j^n la valeur de la solution approchée au point x_j et au temps t_n.

Considérons le problème

$$\begin{cases} \dfrac{\partial^2}{\partial t^2} u(x,t) = c^2 \dfrac{\partial^2}{\partial x^2} u(x,t) \quad \forall x \in [0,L] \text{ et } t \in [0,T] \\ u(x,0) = u^0(x) \quad \text{et } \dfrac{\partial}{\partial t} u(x,0) = u^1(x) \quad \text{données : condition initiale} \\ u(0,t) = u(L,t) = 0 \quad : \text{conditions aux limites de Dirichlet homogènes} \end{cases}$$

et choisissons les approximations classiques suivantes des dérivées secondes par différences finies

$$\frac{\partial^2}{\partial t^2} u(x_j, t_n) \approx \frac{u(x_j, t_{n+1}) - 2u(x_j, t_n) + u(x_j, t_{n-1})}{\Delta t^2} \quad (\text{ à } O(\Delta t^2) \text{ près})$$

$$\frac{\partial^2}{\partial x^2} u(x_j, t_n) \approx \frac{u(x_{j+1}, t_n) - 2u(x_j, t_n) + u(x_{j-1}, t_n)}{\Delta x^2} \quad (\text{ à } O(\Delta x^2) \text{ près})$$

Remplaçons les dérivées partielles par leurs approximations en différences finies ci-dessus et la fonction inconnue u par une collection de valeurs discrètes u_j^n pour $j = 0, ..M$ et $n = 0, ..N$. Nous obtenons un premier exemple de schéma d'approximation en différences finies de l'équation des ondes :

11.4.2 Le schéma explicite (en temps) et centré (en espace)

$$\begin{cases} \dfrac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} = c^2 \dfrac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} \\ u_j^0 = u^0(x_j) \quad \text{et } u_j^1 = u^0(x_j) + \Delta t\, u^1(x_j) \quad \text{déduits des conditions initiales} \\ u_0^n = u_M^n = 0 \quad \forall n : \text{conditions aux limites de Dirichlet homogènes} \end{cases}$$

Ce schéma est un schéma explicite car il donne une formule explicite de calcul de la solution au temps t_{n+1} en fonction des valeurs de la solution au temps précédent. Il n'y a pas d'équation à résoudre pour obtenir la valeur au nouvel instant t_{n+1}.

Ordre

Ce schéma explicite est d'ordre deux en temps et en espace (par développement de Taylor).

Stabilité

Dans le cas de schémas numériques appliqués à des problèmes hyperboliques nous choisirons, comme condition de stabilité, d'imposer au vecteur des solutions approchées d'être conservé ou de décroître en norme au cours du temps.

11.4.3 Étude de la stabilité par l'analyse de Fourier

Reprenons la technique de calcul de la stabilité des schémas par l'analyse de Fourier. Injectons dans le schéma numérique une suite de solutions de la forme

$$u_j^n = \tilde{u}_k^n \, e^{ikj\Delta x}$$

obtenues à partir de conditions initiales

$$u_j^0 = \tilde{u}_k^0 \, e^{ikj\Delta x}$$

et correspondant chacune à une composante harmonique. L'étude de la stabilité se ramène à l'étude de l'évolution au cours des pas de temps n des suites $\tilde{u}_k^n$ quand n augmente. La condition minimale de stabilité numérique nécessite que les nombres $\tilde{u}_k^n$ restent bornés $\forall\, k$ et $\forall\, n = 0, ...N$. Nous imposerons ici que les nombres $\tilde{u}_k^n$ soient conservés ou décroissants en module quand n augmente.

Dans les schémas à p pas, on obtient les $\tilde{u}_k^{n+1}$ par multiplication par une matrice d'amplification $G(\Delta t, k)$ selon :

$$\begin{pmatrix} \tilde{u}_k^{n+1} \\ \tilde{u}_k^{n} \\ : \\ : \\ \tilde{u}_k^{n-p+2} \end{pmatrix} = G(\Delta t, k) \begin{pmatrix} \tilde{u}_k^{n} \\ \tilde{u}_k^{n-1} \\ : \\ : \\ \tilde{u}_k^{n-p+1} \end{pmatrix}$$

Nous choisirons alors les conditions de stabilité suivantes :

Condition de stabilité

Pour que le schéma soit stable il faut qu'il existe $\tau > 0$ tel que les valeurs propres λ_i de la matrice d'amplification $G(\Delta t, k)$ soient toutes majorées en module par 1 selon :

$$|\lambda_i| \leq 1 \quad \forall i = 1, ..p \tag{11.1}$$

quel que soit k et pour tout $0 < \Delta t < \tau$.

Conditions suffisantes de stabilité

1) Si la matrice d'amplification est normale, c'est à dire qu'elle commute avec son adjointe (ou transposée dans le cas réel)

$$G\,G^* = G^*\,G$$

ou bien, ce qui est équivalent, si elle admet une base de vecteurs propres orthonormés, alors la condition précédente (11.1) est suffisante.

2) La condition de normalité n'étant pas toujours vérifiée, on peut utiliser la condition suffisante suivante : le schéma est stable si les coefficients de la matrice $G(\Delta t, k)$ sont bornés et si ses valeurs propres sont toutes de module strictement inférieur à 1 sauf éventuellement une de module égal à 1.

Dans certains cas, comme par exemple le cas du schéma aux différences finies explicite, on est obligé, pour conclure, de faire un calcul complet des vecteurs propres et valeurs propres de la matrice d'amplification $G(\Delta t, k)$.

11.4.4 Application au schéma explicite

Reprenons le schéma explicite :

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

Posons

$$u_j^n = \tilde{u}_k^n \, exp(ikj\Delta x)$$

On obtient :

$$\frac{\tilde{u}_k^{n+1} - 2\tilde{u}_k^n + \tilde{u}_k^{n-1}}{\Delta t^2} = c^2 \frac{exp(ik\Delta x) - 2 + exp(-ik\Delta x)}{\Delta x^2}\, \tilde{u}_k^n$$

soit :

$$\begin{aligned} \tilde{u}_k^{n+1} &= [2 + c^2 \tfrac{\Delta t^2}{\Delta x^2}\big(2\,cos(k\Delta x) - 2\big)]\, \tilde{u}_k^n \; - \tilde{u}_k^{n-1} \\ &= \; [\,2 - 4\,c^2 \tfrac{\Delta t^2}{\Delta x^2} sin^2(\tfrac{k}{2}\Delta x)\,]\, \tilde{u}_k^n - \tilde{u}_k^{n-1} \end{aligned}$$

Notons $\alpha^2 = 4\,c^2\dfrac{\Delta t^2}{\Delta x^2} sin^2(\dfrac{k}{2}\Delta x)$, On obtient l'écriture suivante de la matrice d'amplification :

$$\begin{pmatrix} \tilde{u}_k^{n+1} \\ \tilde{u}_k^n \end{pmatrix} = \begin{pmatrix} 2-\alpha^2 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \tilde{u}_k^n \\ \tilde{u}_k^{n-1} \end{pmatrix}$$

Cette matrice n'est pas une matrice normale.

Les valeurs propres de la matrice 2×2 d'amplification ci-dessus sont racines de

$$\lambda^2 - (2-\alpha^2)\,\lambda \,+\, 1 = 0$$

On trouve le discriminant $\Delta = \alpha^2(\alpha^2 - 4)$.

Si $\alpha^2 > 4$ le discriminant est positif et le trinôme a deux racines réelles distinctes dont le produit vaut 1. L'une des deux est donc forcément de module strictement supérieur à 1 et dans ce cas le schéma est instable.

Si $\alpha^2 < 4$ le trinôme a deux racines complexes conjuguées de module 1 et dans ce cas on ne peut conclure directement car la matrice d'amplification n'est pas normale et qu'alors la condition suffisante de stabilité n'autorise qu'une seule valeur propre de module 1. Un calcul simple permet de vérifier que si λ_1 et λ_2 sont valeurs propres de G, les vecteurs

$$\begin{pmatrix} \lambda_1 \\ 1 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} \lambda_2 \\ 1 \end{pmatrix}$$

sont vecteurs propres de G. On obtient ainsi

$$G = \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} 1 & -\lambda_2 \\ -1 & \lambda_1 \end{pmatrix}$$

D'où

$$G^n = \frac{1}{\lambda_1 - \lambda_2} \begin{pmatrix} \lambda_1 & \lambda_2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{pmatrix} \begin{pmatrix} 1 & -\lambda_2 \\ -1 & \lambda_1 \end{pmatrix}$$

G^n reste donc bornée quel que soit n dans le cas de deux racines complexes conjuguées distinctes, donc à la condition que Δ soit strictement négatif. Une autre manière de montrer la stabilité dans ce cas consiste à remarquer que si l'on dispose de 2 vecteurs propres indépendants on peut exprimer les vecteurs

$$\begin{pmatrix} \tilde{u}_k^{n+1} \\ \tilde{u}_k^n \end{pmatrix}$$

dans la base des vecteurs propres. L'action de la matrice d'itération G se ramène dans cette base à la multiplication des composantes des vecteurs par les valeurs propres λ_1 et λ_2. Comme ces valeurs propres sont de module 1, le vecteur des itérés est borné en module.

Dans le cas $\Delta = 0$ la racine double est -1. Les deux vecteurs propres

$$\begin{pmatrix} \lambda_1 \\ 1 \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} \lambda_2 \\ 1 \end{pmatrix}$$

sont alors confondus. Le sous-espace propre relatif à la valeur propre -1 est de dimension un. La matrice G n'est pas diagonalisable, mais seulement jordanisable sous la forme :

$$\begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}$$

Les puissances nièmes de G tendent vers l'infini avec n. Donc dans ce cas limite le schéma est également instable.

En définitive la stabilité impose $\alpha^2 < 4$, ce qui s'exprime par la condition :

$$c\frac{\Delta t}{\Delta x} < 1$$

dénommée **condition de Courant Friedrichs Lewy** apparaissant ici au sens strict.

11.4.5 Un schéma implicite centré

On considère le schéma implicite suivant directement déduit du schéma explicite précédent :

$$\begin{cases} \dfrac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} = c^2 \dfrac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{\Delta x^2} \\ u_j^0 = u^0(x_j) \quad \text{et } u_j^1 = u^0(x_j) + \Delta t\, u^1(x_j) \quad \text{déduits des condition initiales} \\ u_0^n = u_M^n = 0 \quad \forall n : \text{conditions aux limites de Dirichlet homogènes} \end{cases}$$

Ce schéma n'est plus que d'ordre un en temps. Mais il est inconditionnellement stable, c'est à dire stable quel que soit Δt. Sa matrice d'amplification s'écrit avec les notations précédentes :

$$G(k, \Delta t) = \begin{pmatrix} \dfrac{2}{1+\alpha^2} & -\dfrac{1}{1+\alpha^2} \\ 1 & 0 \end{pmatrix}$$

On peut augmenter la précision en temps, en utilisant le schéma suivant.

11.4.6 Schéma de Newmark implicite d'ordre 2

On considère le schéma implicite suivant déduit des schémas précédents et faisant partie de la famille des schémas de Newmark.

$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} = \frac{c^2}{4}[\delta_j^{2\,n+1} + 2\,\delta_j^{2\,n} + \delta_j^{2\,n-1}]$$

avec

$$\delta_j^{2\,n} = \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2}$$

Ce schéma classique est d'ordre deux et inconditionnellement stable.

11.5 Seconde approche : Système du premier ordre équivalent

L'équation du second ordre

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

se ramène en posant

$$v = \frac{\partial u}{\partial t} \quad \text{et} \quad w = c\frac{\partial u}{\partial x}$$

au système de deux équations du premier ordre :

$$\begin{cases} \dfrac{\partial v}{\partial t} = c\,\dfrac{\partial w}{\partial x} \\ \dfrac{\partial w}{\partial t} = c\,\dfrac{\partial v}{\partial x} \end{cases}$$

pour lequel on doit se donner les deux conditions initiales suivantes au temps zéro pour v et w :

$$v(x,0) = \frac{\partial u}{\partial t}(x,0) = u^1(x)$$

et

$$w(x,0) = c\frac{\partial u}{\partial x}(x,0) = c\frac{d}{dx}\,u^0(x)$$

ceci peut également s'écrire

$$\frac{\partial}{\partial t}\begin{pmatrix} v \\ w \end{pmatrix} = \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix} \frac{\partial}{\partial x}\begin{pmatrix} v \\ w \end{pmatrix}$$

et l'on retrouve sous cette forme la conservation de l'énergie en multipliant scalairement l'équation ci-dessus par le vecteur (v, w) et en intégrant sur $[0, L]$.

Remarque 11.5.1 *Le lien avec l'équation de transport est évident. On retrouvera donc naturellement, dans le chapitre suivant, consacré à la résolution de l'équation de transport, les schémas que nous présentons ici pour résoudre le système du premier ordre équivalent à l'équation des ondes.*

11.5.1 Un premier schéma explicite centré instable

On considère le schéma discret évident suivant :

$$\begin{cases} \dfrac{v_j^{n+1} - v_j^n}{\Delta t} = c\,\dfrac{w_{j+1}^n - w_{j-1}^n}{2\Delta x} \\ \dfrac{w_j^{n+1} - w_j^n}{\Delta t} = c\,\dfrac{v_{j+1}^n - v_{j-1}^n}{2\Delta x} \end{cases}$$

Ce schéma est clairement d'ordre un en temps et deux en espace. Étudions en la stabilité. Posons

$$v_j^n = \tilde{v}_k^n\, e^{ikj\Delta x}$$

et

$$w_j^n = \tilde{w}_k^n\, e^{ikj\Delta x}$$

nous obtenons :

$$\begin{cases} \tilde{v}_k^{n+1} - \tilde{v}_k^n = c\dfrac{\Delta t}{\Delta x}\, i\, sin(k\Delta x)\tilde{w}_k^n \\ \tilde{w}_k^{n+1} - \tilde{w}_k^n = c\dfrac{\Delta t}{\Delta x}\, i\, sin(k\Delta x)\tilde{v}_k^n \end{cases}$$

Soit

$$\begin{pmatrix} \tilde{v}_k^{n+1} \\ \tilde{w}_k^{n+1} \end{pmatrix} = \begin{pmatrix} 1 & ia \\ ia & 1 \end{pmatrix} \begin{pmatrix} \tilde{v}_k^n \\ \tilde{w}_k^n \end{pmatrix}$$

avec

$$a = c\frac{\Delta t}{\Delta x}\, sin(k\Delta x)$$

Les valeurs propres de la matrice d'amplification sont égales à :

$$\lambda = 1 \pm ia \quad \text{donc sont de module} \quad \sqrt{1+a^2} > 1$$

On en déduit l'instabilité de ce schéma quel que soit Δt. Ce résultat négatif que l'on retrouvera dans le cas de l'équation de transport, a fait coulé beaucoup d'encre et a suscité de nombreuses recherches de schémas stables par modifications simples de ce schéma naturel.

11.5.2 Schémas implicites centrés stables

On obtient évidemment des schémas stables en prenant des schémas implicites en temps. Par exemple on peut considérer le schéma de type Euler implicite suivant :

$$\begin{cases} \dfrac{v_j^{n+1}-v_j^n}{\Delta t} = c\,\dfrac{w_{j+1}^{n+1}-w_{j-1}^{n+1}}{2\Delta x} \\ \dfrac{w_j^{n+1}-w_j^n}{\Delta t} = c\,\dfrac{v_{j+1}^{n+1}-v_{j-1}^{n+1}}{2\Delta x} \end{cases}$$

On montrera en exercice que ce schéma d'ordre un en temps et deux en espace. est inconditionnellement stable.

On peut également considérer le schéma de type Crank Nicolson suivant :

$$\begin{cases} \dfrac{v_j^{n+1}-v_j^n}{\Delta t} = \dfrac{c}{2}\,\Big[\dfrac{w_{j+1}^{n}-w_{j-1}^{n}}{2\Delta x} + \dfrac{w_{j+1}^{n+1}-w_{j-1}^{n+1}}{2\Delta x}\Big] \\ \dfrac{w_j^{n+1}-w_j^n}{\Delta t} = \dfrac{c}{2}\,\Big[\dfrac{v_{j+1}^{n}-v_{j-1}^{n}}{2\Delta x} + \dfrac{v_{j+1}^{n+1}-v_{j-1}^{n+1}}{2\Delta x}\Big] \end{cases}$$

Schéma d'ordre deux en temps et en espace inconditionnellement stable et conservatif (les valeurs propres de la matrice d'amplification sont deux complexes conjugués de module un).

11.5.3 Schémas explicites stables

Schéma de Lax

On remplace au premier membre des équations du schéma explicite centré instable v_j^n par $\dfrac{v_{j+1}^n + v_{j-1}^n}{2}$ et de même w_j^n par $\dfrac{w_{j+1}^n + w_{j-1}^n}{2}$. Ceci peut s'interpréter comme un lissage en x ou comme l'ajout d'un terme dissipatif $\dfrac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{2\Delta t}$ approximant $\dfrac{\Delta x^2}{2\Delta t}\dfrac{\partial^2 v}{\partial x^2}$ et de même pour w.

On obtient alors le schéma de Lax suivant :

$$\begin{cases} \dfrac{v_j^{n+1} - \frac{v_{j+1}^n + v_{j-1}^n}{2}}{\Delta t} = c\,\dfrac{w_{j+1}^n - w_{j-1}^n}{2\Delta x} \\ \dfrac{w_j^{n+1} - \frac{w_{j+1}^n + w_{j-1}^n}{2}}{\Delta t} = c\,\dfrac{v_{j+1}^n - v_{j-1}^n}{2\Delta x} \end{cases}$$

On montre que ce schéma est d'ordre un en temps et stable sous la condition de Courant Friedrichs Lewy dite condition CFL

$$\frac{c\Delta t}{\Delta x} \leq 1$$

Remarque 11.5.2 *Voir également, dans le chapitre suivant consacré à la résolution de l'équation de transport, l'interprétation du schéma de Lax comme schéma de caractéristiques.*

Schéma de Lax-Wendroff

On ajoute au schéma explicite instable un terme dissipatif en $O(\Delta t)$ correspondant à une discrétisation de $\frac{c^2\Delta t}{2}\frac{\partial^2}{\partial x^2}$. Une interprétation classique de cette modification consiste à remarquer que le développement de Taylor

$$v(x_j, t_{n+1}) = v(x_j, t_n) + \Delta t \frac{\partial v}{\partial t} + \frac{\Delta t^2}{2}\frac{\partial^2 v}{\partial t^2} + O(\Delta t^3)$$

devient en utilisant l'équation

$$\frac{\partial v}{\partial t} = c\frac{\partial w}{\partial x}$$

$$v(x_j, t_{n+1}) = v(x_j, t_n) + \Delta t \frac{\partial v}{\partial t} + \frac{c^2\Delta t^2}{2}\frac{\partial^2 w}{\partial x^2} + O(\Delta t^3)$$

L'ajout des termes $\frac{c^2\Delta t}{2}\frac{w_{j+1}^n - 2w_j^n + w_{j-1}^n}{\Delta x^2}$ et $\frac{c^2\Delta t}{2}\frac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{\Delta x^2}$ donnera le second ordre en temps au schéma et assurera sa stabilité.

On obtient le schéma de Lax-Wendroff suivant :

$$\begin{cases} \frac{v_j^{n+1} - v_j^n}{\Delta t} = c\,\frac{w_{j+1}^n - w_{j-1}^n}{2\Delta x} + \frac{c^2\Delta t}{2}\frac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{\Delta x^2} \\ \frac{w_j^{n+1} - w_j^n}{\Delta t} = c\,\frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x} + \frac{c^2\Delta t}{2}\frac{w_{j+1}^n - 2w_j^n + w_{j-1}^n}{\Delta x^2} \end{cases}$$

On montrera que la matrice d'amplification de ce schéma s'écrit :

$$G_k = \begin{pmatrix} 1 + \frac{c^2\Delta t^2}{\Delta x^2}(cos(k\Delta x) - 1) & i\frac{c\Delta t}{\Delta x}sin(k\Delta x) \\ i\frac{c\Delta t}{\Delta x}sin(k\Delta x) & 1 + \frac{c^2\Delta t^2}{\Delta x^2}(cos(k\Delta x) - 1) \end{pmatrix}$$

et que ce schéma est stable sous la condition CFL

$$\frac{c\Delta t}{\Delta x} \leq 1$$

Schéma de Courant Friedrichs

On utilise pour stabiliser le schéma explicite un maillage décalé pour w par rapport au maillage de discrétisation de v et un calcul implicite des w en fonction des v. On écrit :

$$\begin{cases} \dfrac{v_j^{n+1} - v_j^n}{\Delta t} = c\,\dfrac{w_{j+\frac{1}{2}}^n - w_{j-\frac{1}{2}}^n}{\Delta x} \\ \dfrac{w_{j-\frac{1}{2}}^{n+1} - w_{j-\frac{1}{2}}^n}{\Delta t} = c\,\dfrac{v_j^{n+1} - v_{j-1}^{n+1}}{\Delta x} \end{cases}$$

Ce schéma est globalement explicite puisque le calcul de v_j^{n+1} et v_{j-1}^{n+1} est effectué avant le calcul de $w_{j-\frac{1}{2}}^{n+1}$. Il est équivalent au schéma explicite du second ordre en posant

$$v_j^n = \frac{u_j^n - u_j^{n-1}}{\Delta t}$$

et

$$w_{j-\frac{1}{2}}^n = c\,\frac{u_j^n - u_{j-1}^n}{\Delta x}$$

On obtient exactement la même condition de stabilité CFL stricte.

$$\frac{c\Delta t}{\Delta x} < 1$$

C'est d'ailleurs à propos de l'étude de ce schéma que Courant Friedrichs et Lewy ont introduit le concept de stabilité.

11.5.4 Interprétation de la condition de Courant-Friedrichs-Lewy

La condition de Courant Friedrichs Lewy souvent mentionnée exprime la compatibilité nécessaire entre domaine de dépendance théorique et domaine de dépendance numérique. Le pas de temps Δt doit rester inférieur à la valeur limite au delà de laquelle des parties du domaine de dépendance théorique ne seraient pas prises en compte dans le schéma numérique. Autrement dit le domaine de dépendance numérique issu du point (x_j, t_{n+1}) doit inclure le domaine de dépendance théorique correspondant, le triangle $\{(x_{j-1}, t_n), (x_{j+1}, t_n), (x_j, t_{n+1})\}$. Une autre façon de dire la même chose consiste à limiter le pas de temps de telle sorte qu'en un pas de temps Δt l'onde ne parcourt pas une distance supérieure à un pas d'espace Δx. On retrouvera cette condition CFL dans le cas de l'équation de transport.

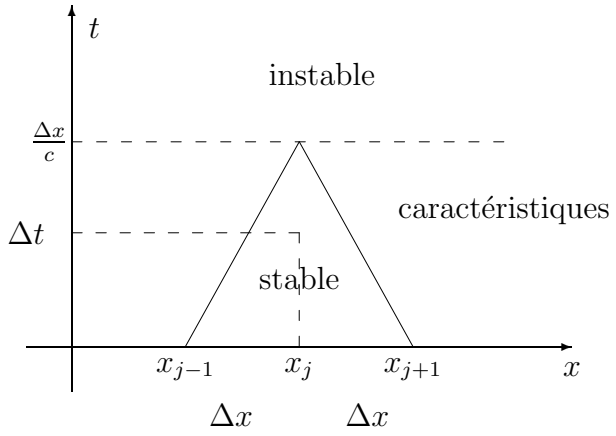


FIGURE 11.4 – Domaine de dépendance et condition de stabilité CFL

11.6 Méthodes d'éléments finis pour le problème des membranes vibrantes

On considère le problème bidimensionnel suivant :

$$\begin{cases} \dfrac{\partial^2}{\partial t^2}\, u(x,y,t) = c^2\, \Delta\, u(x,y,t) + f(x,y,t) \quad \forall x,y \in \Omega \text{ et } t \in [0,T] \\ u(x,y,0) = u^0(x,y) \quad \dfrac{\partial}{\partial t} u(x,y,0) = u^1(x,y) \quad \text{données : condition initiale} \\ u_{\,|\Gamma_0} = u_d \quad : \text{conditions aux limites de Dirichlet} \\ c^2\, \dfrac{\partial u}{\partial n}_{\,|\Gamma_1} = g \quad : \text{conditions aux limites de Neumann} \end{cases}$$

11.6.1 Formulation variationnelle

La formulation variationnelle du problème s'obtient comme dans le cas stationnaire par multiplication par des fonctions tests indépendantes du temps appartenant, compte tenu des conditions aux limites choisies, à l'espace V des fonctions de $H^1(\Omega)$ nulles sur Γ_0. Après intégration en espace sur le domaine Ω et intégration "par parties" par la formule de Green, on obtient le problème variationnel :

$$\left\{\begin{array}{l} \text{Trouver pour tout } t\in[0,T], \quad u:(x,y,t)\longrightarrow u(x,y,t) \text{ telle que :} \\ \displaystyle\iint_\Omega \frac{\partial^2}{\partial t^2}\,u(x,y,t)\,v(x,y)\,dxdy + c^2 \iint_\Omega \mathbf{grad}u\,\mathbf{grad}v\,dxdy \\ \qquad\qquad = \displaystyle\iint_\Omega f(x,y,t)\,v(x,y)\,dxdy + \int_{\Gamma_1} g\,v\,d\gamma \quad \forall v\in V \\ u(x,y,0)=u_0(x,y) \quad \dfrac{\partial}{\partial t}u(x,y,0)=u^1(x,y) \quad \text{données} \end{array}\right.$$

11.6.2 Semi-discrétisation en espace par éléments finis

On suppose réalisé un maillage du domaine Ω en éléments finis triangulaires P_k ou quadrangulaires Q_k. Soient w_i les fonctions de base associées aux éléments finis choisis. Notons $\mathbf{I}$ l'ensemble des indices des noeuds du maillage correspondant à une valeur inconnue de la solution u. C'est à dire ici l'ensemble des noeuds n'appartenant pas à Γ_0. Notons $\mathbf{J}$ l'ensemble des indices des sommets du maillage correspondant à une valeur connue de la solution, donc ici appartenant à Γ_0.

Comme dans le cas stationnaire nous décomposerons la solution approchée u_h en somme d'une inconnue auxiliaire $\tilde{u}_h$ et d'une fonction connue u_0 prenant les valeurs imposées sur Γ_0. La solution auxiliaire $\tilde{u}_h$ s'écrira dans la base des w_i pour $i\in\ \mathbf{I}$ selon :

$$\tilde{u}_h(x,y)=\sum_{i\in\ \mathbf{I}} u_i(t)\,w_i(x,y)$$

La fonction u_0, supposée indépendante du temps pour simplifier, sera approchée par une fonction $u_{0,h}$ prenant les valeurs imposées sur Γ_0 et nulle sur tous les noeuds d'indices $i\in\ \mathbf{I}$

$$u_{0,h}(x,y)=\sum_{i\in\ \mathbf{J}} u_d(x_i,y_i)\ w_i(x,y)$$

On en déduit

$$\frac{\partial^2}{\partial t^2}u_h(x,y,t)=\sum_{i\in\ \mathbf{I}} u_i''(t)\,w_i(x,y)$$

Notons $N_{\mathbf{I}}$ le nombre de noeuds "inconnus" d'indices $i\in\ \mathbf{I}$. Le problème approché s'écrit sous la forme d'un système différentiel linéaire de $N_{\mathbf{I}}$ équations à $N_{\mathbf{I}}$ fonctions inconnues du temps u_i.

$$\left\{\begin{array}{l} \text{Trouver } \forall t \in [0,T], \text{ et } \forall j \in \mathbf{I}, \text{ les fonctions } u_j(t) \text{ telles que,} \forall i \in \mathbf{I} : \\ \displaystyle\sum_{j\in\ \mathbf{I}} \big(\iint_\Omega w_j\, w_i\, dxdy \big) u_j''(t) + \big(c^2 \iint_\Omega \mathbf{grad} w_j\, \mathbf{grad} w_i\, dxdy \big) u_j(t) = \\ \displaystyle\iint_\Omega f\, w_i\, dxdy + \int_{\Gamma_1} g\, w_i\, d\gamma \; - \sum_{j\in\ \mathbf{J}} \big(c^2 \iint_\Omega \mathbf{grad} w_j\, \mathbf{grad} w_i\, dxdy \big) u_d(x_j, y_j) \\ u_i(0) \quad \text{et} \quad u'_i(0) \quad \text{donnés } \forall\, i \in \mathbf{I} \end{array}\right.$$

Ce qui donne matriciellement :

$$\left\{\begin{array}{l} \text{Trouver pour tout } t \in [0,T], \text{ le vecteur } U(t) \text{ tel que :} \\ MU''(t) + KU(t) = B \\ U(0) \text{ et } U'(0) \quad \text{donnés} \end{array}\right.$$

avec M matrice de masse de coefficients

$$m_{i,j} = \iint_\Omega w_j\, w_i\, dxdy$$

K matrice de raideur de coefficients

$$k_{i,j} = c^2 \iint_\Omega \mathbf{grad} w_j\, \mathbf{grad} w_i\, dxdy$$

et B vecteur second membre dont les coefficients sont dans ce cas égaux à

$$B_i = \iint_\Omega f(x,y,t)\, w_i\, dxdy + \int_{\Gamma_1} g\, w_i\, d\gamma \; - \sum_{j\in\ \mathbf{J}} k_{i,j} u_d(x_j, y_j)$$

11.7 Discrétisation complète en espace et en temps

Il nous reste à appliquer les schémas en temps déjà présentés dans le cas de discrétisations en différences finies pour obtenir une discrétisation complète du problème.

11.7.1 Schéma du second ordre explicite

On utilise l'approximation

$$U''(t) \approx \frac{U(t+\Delta t) - 2U(t) + U(t-\Delta t)}{\Delta t^2}$$

ce qui conduit au schéma

$$\begin{cases} \text{Trouver pour tout } n \in [0, N], \text{ la suite de vecteurs } U^n \text{ tels que :} \\ U^0 \text{ et } U^1 \text{ donnés} \\ MU^{n+1} = 2MU^n - MU^{n-1} - \Delta t^2 KU^n + \Delta t^2 B \end{cases}$$

Remarquons que la dépendance du second membre B par rapport au temps ne poserait pas de problème difficile.

La résolution complète du problème approché nécessite la résolution d'un système matriciel à chaque pas de temps. Nous avons deux possibilités :

1) On factorise une fois pour toutes en début de calcul la matrice de masse M qui est symétrique définie positive sous forme LL^T et on a deux systèmes triangulaires à résoudre à chaque pas de temps.

2) On calcule la matrice de masse de façon approchée sous forme d'une matrice de masse condensée diagonale (lumping). Ce qui est facile pour des éléments P1. L'inversion de la matrice de masse est alors immédiate et on obtient véritablement un schéma numérique explicite.

11.7.2 Schéma implicite

Le choix d'une discrétisation implicite en temps conduit au schéma suivant :

$$\begin{cases} \text{Trouver pour tout } n \in [0, N], \text{ la suite de vecteurs } U^n \text{ tels que :} \\ U^0 \text{ et } U^1 \text{ donnés} \\ (M + \Delta t^2 K)U^{n+1} = 2MU^n - MU^{n-1} + \Delta t^2 B \end{cases}$$

Dans ce cas, que l'on ait ou non condensé la matrice de masse sous forme diagonale, on doit résoudre un système matriciel. Ce que l'on fait en factorisant une fois pour toutes au début du calcul, la matrice $M + \Delta t^2 K$ qui est symétrique définie positive, sous forme LL^T puis en résolvant à chaque pas deux systèmes triangulaires.

11.7.3 Schéma de Newmark

On peut, pour une meilleure précision, utiliser le schéma de **Newmark** pour la discrétisation en temps. On aboutit ainsi à une méthode classique et performante pour la résolution temporelle des problèmes de vibrations.

$$\begin{cases} \text{Trouver pour tout } n \in [0, N], \text{ la suite de vecteurs } U^n \text{ tels que :} \\ U^0 \text{ et } U^1 \text{ donnés} \\ (M + \frac{\Delta t^2}{4} K)\, U^{n+1} = 2(M - \frac{\Delta t^2}{4} K)\, U^n - (M + \frac{\Delta t^2}{4} K)\, U^{n-1} + \Delta t^2 B \end{cases}$$

11.8 Analyse modale et décomposition orthogonale propre

Pour résoudre le problème des vibrations de structures, une autre méthode couramment utilisée consiste à rechercher les modes propres de vibrations en résolvant un problème de valeurs propres et à représenter second-membre et solutions dans la base de ces modes propres. C'est ce qu'on appelle l'analyse modale. Ainsi, on ramène la résolution de l'EDP à celle de plusieurs équations différentielles ordinaires pour les coefficients du développement de la solution dans la base des modes propres obtenus.

Le calcul des modes propres est coûteux. La détermination des valeurs propres et des vecteurs propres est un des algorithmes les plus délicats de l'analyse numérique matricielle (voir 2.8). Pour réduire ce coût on choisit une base partielle de modes propres.

On peut utiliser des solutions observées ou calculées pour construire une base orthogonale, voire orthonormale par Gram-Schmidt (voir annexe A), qui sera utilisée à la place de celle des modes propres. C'est ce qu'on appelle la décomposition orthogonale propre (DOP).

Cette réduction du modèle est très utile, en particulier lorsqu'on s'intéresse à plusieurs configurations similaires ou si l'on veut évaluer la sensibilité de la solution aux petites variations des paramètres du problème.

Cette approche sera appliquée en couplage et optimisation aux chapitres 13 et 17.

11.8.1 Cas linéaire

Nous allons illustrer notre propos à travers l'équation des ondes :

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f \quad \text{sur } \Omega, \tag{11.2}$$

avec les conditions initiales et aux limites adéquates. Considérons une semi-discrétisation en temps de l'équation. On pourra ainsi générer une suite $u^n(x)$ de solutions instantanées aux temps t^n vérifiant :

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

On appelle échantillon ou "snapshot" du signal $u(x,t)$ une collection d'instantanés u^n. Cet échantillonnage est satisfaisant si les u^n sont linéairement indépendants et s'ils représentent correctement la dynamique du signal. L'orthonormalisation de Gram-Schmidt permet de garantir l'indépendance linéaire des instantanés retenus. Notons, $\{X_i(x), \quad i = 1, .., m\}$, m snapshots orthonormalisés et $(.,.)$ le produit scalaire euclidien. On aura donc : $((X_i, X_j) = \delta_{i,j}, \text{ pour } i, j = 1, .., m)$.

Considérons une décomposition des solutions sous la forme :

$$u(x,t) = \sum_{j=1}^{m} a_j(t) X_j(x), \quad a_j = (u(x,t), X_j),$$

Les solutions initiales aux deux instants $t = 0$, et $t = \Delta t$ permettront d'identifier les deux conditions initiales nécessaires pour les fonctions a_j.

En portant dans (11.2) et en utilisant la linéarité des dérivations, on obtient :

$$\sum_{j=1}^{m} \frac{d^2(a_j)}{dt^2} X_j - \sum_{j=1}^{m} a_j \Delta X_j = f,$$

Ce qui après multiplication par X_i, pour $i = 1, .., m$ et intégration sur le domaine spatial Ω nous donne :

$$|\Omega| \frac{d^2(a_i)}{dt^2} = \sum_{j=1}^{m} a_j(t) \, (\Delta X_j, X_i) + (f, X_i), \quad i = 1, .., m.$$

Il reste donc à résoudre m équations différentielles ordinaires du second ordre après avoir évalué et stocké les $(\Delta X_j, X_i)$ et (f, X_i).

11.8.2 Cas non-linéaire

Cette technique n'est pas directement utilisable dans le cas de problèmes non-linéaires, car l'opération ci-dessus utilise la linéarité du laplacien. Considérons l'équation non-linéaire suivante :

$$\frac{\partial^2 u}{\partial t^2} + F(u) = 0, \quad \text{avec les conditions aux limites adéquate.}$$

$F(u)$ est un opérateur différentiel spatial non-linéaire en u. Après avoir créé la base des snapshots, on obtient :

$$|\Omega|\frac{d^2(a_i)}{dt^2} = -(F(\sum_{j=1}^{m} a_j(t)X_j), X_i), \quad i = 1, .., m.$$

On ne peut donc pas évaluer la contribution spatiale de façon indépendante du temps. Une solution possible est l'utilisation de la DOP pour l'équation linéarisée.

La solution du modèle linéarisé a de nombreuses applications, notamment en analyse de stabilité où on l'on souhaite que la dépendance de la solution de l'équation linéarisée par rapport aux paramètres du modèle soit continue (i.e. une petite perturbation d'un paramètre implique une variation bornée de la solution). Au voisinage d'une solution u, les petites perturbations vérifient :

$$\frac{\partial^2 v}{\partial t^2} + \frac{\partial F}{\partial u}(u)v = 0,$$

ce qui, en suivant la démarche précédente, aboutit à :

$$|\Omega|\frac{d^2(a_i)}{dt^2} = -\sum_{j=1}^{m} a_j(t)\,(\frac{\partial F}{\partial u}(u)X_j, X_i), \quad i = 1, .., m.$$

L'évaluation spatiale se fait alors indépendamment de l'intégration temporelle.

Chapitre 12

Introduction aux problèmes hyperboliques du premier ordre : l'équation de transport

12.1 Position du problème

Considérons un champ de vitesses $\mathbf{V}(x,t)$ donné et une grandeur scalaire $u(x,t)$ transportée au cours du temps par le champ $\mathbf{V}$ à travers un domaine $\Omega \subset I\!R^d$ de bord Γ. À chaque instant t, u vérifie l'équation de transport (dite aussi, de convection ou d'advection) :

$$\frac{\partial u}{\partial t} + \mathbf{V}.\mathbf{grad}u = 0 \qquad \forall x \in \Omega \quad \text{et} \quad \forall t \in [0,T]$$

Si le champ $\mathbf{V}$ est à divergence nulle (i.e. $div(\mathbf{V}) = \nabla.\mathbf{V} = 0$), on peut écrire cette équation sous forme conservative :

$$\frac{\partial u}{\partial t} + div(\mathbf{V}u) = 0 \qquad \forall x \in \Omega \quad \text{et} \quad \forall t \in [0,T] \tag{12.1}$$

L'équation de transport est un exemple de problème conservatif. En effet, l'utilisation de la formule de Stokes, ou formule de la divergence, nous donne :

$$\int_\Omega \frac{\partial u}{\partial t} dx = \frac{d}{dt}\int_\Omega u dx = -\int_\Gamma (\mathbf{V}.\vec{n}) u d\gamma. \tag{12.2}$$

avec $\vec{n}$ vecteur normal unitaire à Γ orienté vers l'extérieur de Ω.

Si u est une mesure de densité, cette formule lie la variation de la masse contenue dans Ω à l'apport, par le champ $\mathbf{V}$, de u à travers la frontière.

La détermination de la solution de ce problème du premier ordre en temps nécessite de fixer une condition initiale pour u :

$$u(x,0) = u^0(x).$$

On obtient ainsi un problème à valeur initiale ou problème de Cauchy.

Les conditions aux limites choisies sur la frontière Γ doivent être cohérentes avec l'équation de transport qui est du premier ordre en espace. Plus précisément, il convient de laisser libre soit, les valeurs de u sur la frontière où le champ est sortant, soit sur celle où il est entrant. Ce choix dépend des mesures dont on dispose. Notons Γ^- la partie de la frontière où le champ est entrant ($\vec{V}.\vec{n}(x) < 0$) et Γ^+ celle où il est sortant ($\vec{V}.\vec{n}(x) > 0$). Si nous imposons u sur la frontière où le champ $\mathbf{V}$ est entrant :

$$u(x,t) = u_\Gamma(x,t), \quad \text{si} \quad \vec{V}.\vec{n}(x) < 0,$$

on obtient après une semi-discrétisation par Euler implicite en temps de (12.2) :

$$\begin{aligned}(\int_\Omega u^{p+1}dx) - (\int_\Omega u^p dx) &= -\Delta t(\int_\Gamma (\mathbf{V}.\vec{n})u^{p+1}d\gamma) \\ &= -\Delta t(\int_{\Gamma^+} (\mathbf{V}.\vec{n})u^{p+1}d\gamma + \int_{\Gamma^-} (\mathbf{V}.\vec{n})u^{p+1}d\gamma).\end{aligned} \tag{12.3}$$

Ceci implique une contrainte pour $\displaystyle\int_\Omega u^{p+1}dx$. Le système est donc sur-déterminé, la solution de l'équation (12.1) devra vérifier cette contrainte. On choisira les schémas numériques qui la vérifient.

12.2 Discrétisation de l'équation de transport

Plaçons nous dans le cas monodimensionnel d'un tube de longueur L. On choisit une discrétisation régulière de $[0, L]$ en intervalles de longueur Δx tels que $L = M\Delta x$ et une discrétisation de l'intervalle de temps $[0, T]$ en pas de temps de longueur Δt tels que $T = N\Delta t$. Notons x_j le point $j\Delta x$ et t_n le temps $n\Delta t$. Notons u_j^n la valeur de la solution approchée au point x_j et au temps t_n.

Considérons le problème :

$$\begin{cases} \dfrac{\partial}{\partial t}\, u(x,t) + c\, \dfrac{\partial}{\partial x}\, u(x,t) \quad \forall x \in [0,L] \text{ et } t \in [0,T] \\ u(x,0) = u^0(x) \quad \text{données : condition initiale} \\ u(0,t) = a, \quad c > 0. \end{cases} \tag{12.4}$$

La condition aux limites est imposée à la frontière gauche (le point $x = 0$) où le flot est entrant.

La discrétisation de cette équation simple, dont la solution exacte est donnée par

$$u(x,t) = u^0(x - ct)$$

demande quelques précautions car la dérivée première en espace a une structure non symétrique. Physiquement, on modélise un phénomène où le sens de parcours est imposé (penser à l'amont et l'aval dans une rivière). Nous verrons quatre types de discrétisations possibles :

— les schémas centrés,
— les schémas décentrés, dans lesquels le décentrage est introduit soit par l'approximation des dérivées en différences finies, soit par la définition des variables aval et amont en volumes finis, soit par le choix des fonctions de base en éléments finis,
— Les schémas distributifs, qui sont une formulation compacte de schémas éléments finis décentrés.
— Les schémas de caractéristiques.

12.3 Schémas centrés

Les schémas explicites centrés de discrétisation directe de l'équation de transport sont instables. Le choix d'écritures implicites en temps ou l'ajout de termes de viscosité sont absolument nécessaires à la stabilité des discrétisations centrées en espace de l'équation originale. Nous allons illustrer cela en étudiant en particulier :

• Le schéma d'Euler explicite centré instable

• Les schémas d'Euler implicite et de Crank-Nicolson qui sont inconditionnellement stables.

• Les schémas centrés explicites de Lax et de Lax-Wendroff qui sont stables, sous condition CFL, par l'ajout de viscosité numérique.

12.3.1 Un premier schéma explicite centré instable

On considère le schéma discret suivant :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\,\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0. \tag{12.5}$$

Ce schéma est clairement d'ordre un en temps et deux en espace. Étudions en la stabilité par la méthode de Fourier. Posons

$$u_j^n = \tilde{u}_k^n \, e^{ikj\Delta x}$$

nous obtenons :

$$\tilde{u}_k^{n+1} - \tilde{u}_k^n + c\frac{\Delta t}{\Delta x}\, i\, sin(k\Delta x)\tilde{u}_k^n = 0.$$

Soit

$$\tilde{u}_k^{n+1} = (1 - i\alpha)\ \tilde{u}_k^n,$$

avec

$$\alpha = c\frac{\Delta t}{\Delta x}\, sin(k\Delta x).$$

Les coefficients d'amplification G_k sont donc tous de module > 1. On en déduit l'instabilité de ce schéma quel que soit Δt. On dit que le schéma est inconditionnellement instable. Nous allons rechercher des schémas stables par modification de ce schéma naturel.

12.3.2 Schémas implicites centrés stables

On obtient évidemment des schémas stables en prenant des schémas implicites en temps. Par exemple on peut considérer le schéma de type Euler implicite suivant :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\,\frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2\Delta x} = 0. \tag{12.6}$$

La technique de Fourier permet de montrer simplement que son facteur d'amplification est

$$G_k = \frac{1}{1 + ic\frac{\Delta t}{\Delta x} sin(k\Delta x)}$$

et donc que ce schéma, d'ordre un en temps et deux en espace, est inconditionnellement stable. Ce schéma est dissipatif car $|G_k| < 1\, \forall k$.

On peut également considérer le schéma, de type Crank-Nicolson, suivant :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{c}{2}\,\Big[\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} + \frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2\Delta x}\Big] = 0. \tag{12.7}$$

Ce schéma, d'ordre deux en temps et en espace, est inconditionnellement stable et conservatif en norme L^2 car les coefficients d'amplification

$$G_k = \frac{1 - ic\frac{\Delta t}{2\Delta x} sin(k\Delta x)}{1 + ic\frac{\Delta t}{2\Delta x} sin(k\Delta x)}$$

sont des complexes de module un.

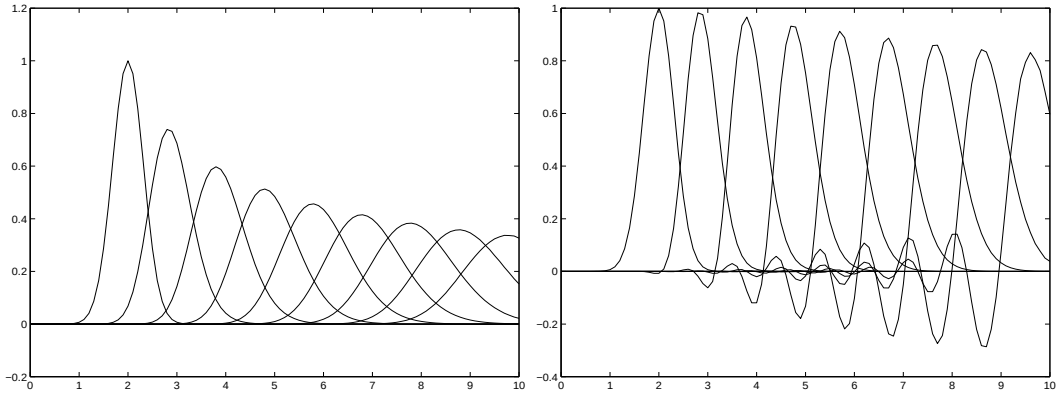


FIGURE 12.1 – Transport d'une gaussienne en utilisant les schémas d'Euler implicite et de Crank-Nicolson avec les mêmes pas d'espace et de temps. On constate l'important amortissement dans le cas du schéma d'Euler implicite (d'ordre un) et la non-positivité du schéma de Crank-Nicolson. Dans les deux cas, la sortie utilise un schéma décentré.

Remarque 12.3.1 *Ainsi, on voit que l'utilisation d'une discrétisation en temps implicite permet d'obtenir des schémas centrés stables pour l'équation d'advection. Cependant, il est parfois coûteux d'utiliser des schémas purement implicites, en particulier dans les cas non-linéaires, les systèmes d'équations et les configurations couplées que nous rencontrerons plus loin. La résolution du schéma implicite est difficile, par exemple, lorsque le système discret est trop grand pour une résolution directe et qu'une résolution itérative s'avère inefficace.*

12.3.3 Schémas explicites centrés stables

Schéma de Lax

On remplace au premier membre du schéma explicite centré instable (12.5), le terme u_j^n par $\dfrac{u_{j+1}^n + u_{j-1}^n}{2}$. On obtient alors le schéma de Lax suivant :

$$\frac{u_j^{n+1} - \frac{u_{j+1}^n + u_{j-1}^n}{2}}{\Delta t} + c\,\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0. \tag{12.8}$$

On peut réécrire le schéma de Lax sous la forme :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\,\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} - \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{2\Delta t} = 0.$$

Ceci peut s'interpréter comme un lissage en x ou comme l'ajout d'un terme dissipatif $-\dfrac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{2\Delta t}$ approximant $-\dfrac{\Delta x^2}{2\Delta t}\dfrac{\partial^2 u}{\partial x^2}$. L'équation équivalente

continue, au deuxième ordre près, est alors une équation d'advection-diffusion :

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = 0, \quad \text{avec} \quad \nu = \frac{\Delta x^2}{2\Delta t}. \tag{12.9}$$

Ainsi, la stabilisation du schéma se fait au prix de l'introduction d'une erreur de consistance. Ce schéma est d'ordre un en temps et stable sous la condition de Courant Friedrichs Levy dite condition CFL :

$$\frac{c\Delta t}{\Delta x} \leq 1. \tag{12.10}$$

On voit que la viscosité numérique ν est uniquement fonction des pas de discrétisation. La viscosité numérique est liée à la vitesse par une expression de la forme : $\nu \sim c\Delta x/2$.

Le schéma de Lax peut également s'interpréter comme le résultat de l'interpo-

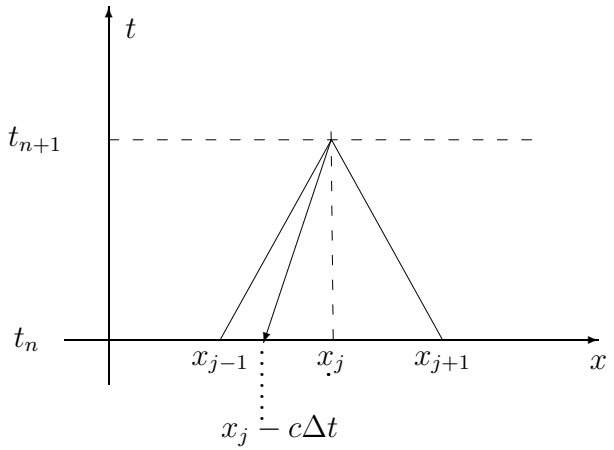


FIGURE 12.2 – Le schéma de Lax est un schéma de caractéristiques

lation sur la grille d'une méthode de caractéristiques. En effet, la méthode des caractéristiques consiste à utiliser l'égalité de la valeur u_j^{n+1} (au point x_j et au temps t_{n+1}) et de la valeur au temps précédent t_n au point $x_j - c\Delta t$. Ce point n'est en général pas un point de la grille (sauf si $\dfrac{c\Delta t}{\Delta x}$ est un entier). On interpole donc cette valeur linéairement selon

$$u_j^{n+1} = \frac{1}{2}[(1 + \frac{c\Delta t}{\Delta x})u_{j-1}^n + (1 - \frac{c\Delta t}{\Delta x})u_{j+1}^n]$$

ce qui redonne (12.8).

Remarque 12.3.2 *Nous retrouverons cette relation entre interpolation et introduction de dissipation dans l'équation en maillage adaptatif (au chapitre 18), lors de l'interpolation d'un champ d'un maillage sur un autre après adaptation.*

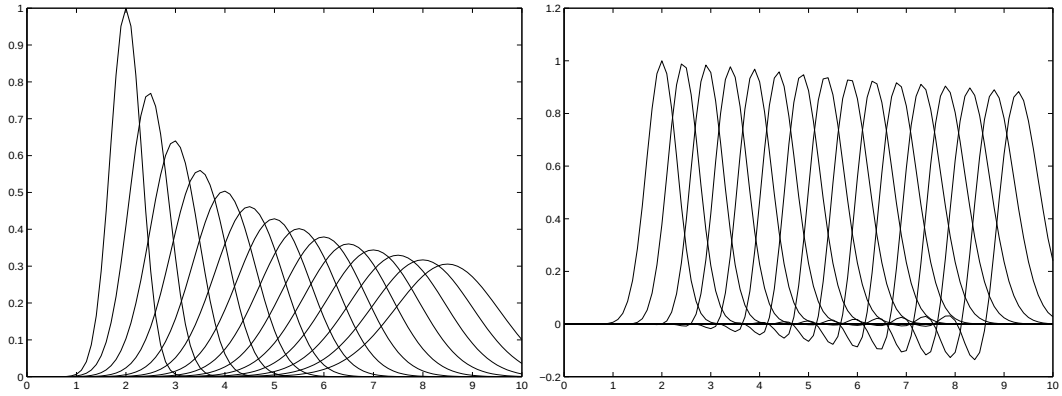


FIGURE 12.3 – Même problème de transport d'une gaussienne que figure 12.1 avec les schémas de Lax et de Lax-Wendroff, tous deux à CFL=0.5. On constate la forte dissipation du schéma de Lax, qui reste cependant positif, et la moindre dissipation, mais la non-positivité de Lax-Wendroff. Là encore, schéma décentré en sortie.

Schéma de Lax-Wendroff

De même, on peut utiliser l'approche Lax-Wendroff comme pour l'équation des ondes au chapitre précédent. Cela revient à ajouter au schéma explicite instable un terme dissipatif en $O(\Delta t)$ correspondant à une discrétisation de $\frac{c^2\Delta t}{2}\frac{\partial^2}{\partial x^2}$. Une interprétation classique de cette modification consiste à remarquer que le développement de Taylor :

$$u(x_j, t_{n+1}) = u(x_j, t_n) + \Delta t \frac{\partial u}{\partial t} + \frac{\Delta t^2}{2}\frac{\partial^2 u}{\partial t^2} + O(\Delta t^3), \tag{12.11}$$

devient, en utilisant $\frac{\partial u}{\partial t} = -c\frac{\partial u}{\partial x}$ et $\frac{\partial^2 u}{\partial t^2} = -c\frac{\partial^2 u}{\partial x \partial t} = c^2\frac{\partial^2 u}{\partial x^2}$:

$$u(x_j, t_{n+1}) = u(x_j, t_n) - c\Delta t \frac{\partial u}{\partial x} + \frac{c^2\Delta t^2}{2}\frac{\partial^2 u}{\partial x^2} + O(\Delta t^3). \tag{12.12}$$

L'ajout du terme

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

donnera le second ordre en temps au schéma et assurera sa stabilité. On construit ainsi le schéma de Lax-Wendroff :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\,\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} - \frac{c^2\Delta t}{2}\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = 0. \tag{12.13}$$

La technique de Fourier permet d'obtenir la stabilité de ce schéma sous la condition CFL $\dfrac{c\Delta t}{\Delta x} \leq 1$.

Remarque 12.3.3 *Bien que les schémas de Lax et de Lax-Wendroff produisent tous deux à $CFL = 1$, la solution exacte (voir aussi la remarque 12.4.1), ils ne donnent pas la même solution lors d'une utilisation sur des maillages à pas variables ou bien avec des vitesses variables en temps et en espace. Leurs extensions aux dimensions supérieures accentuent cet écart. Ceci est dû aux expressions différentes de la viscosité numérique dans ces deux schémas.*

Remarque 12.3.4 *La viscosité numérique a la même dimension qu'une viscosité cinématique physique (vitesse fois longueur). Le concept de viscosité numérique s'étend bien aux dimensions supérieures lors de la résolution d'une équation scalaire (lors de l'advection d'un scalaire en 3D par exemple), par contre son extension n'est pas aisée dans le cas des systèmes d'équations couplées (comme le système d'Euler pour la dynamique des gaz).*

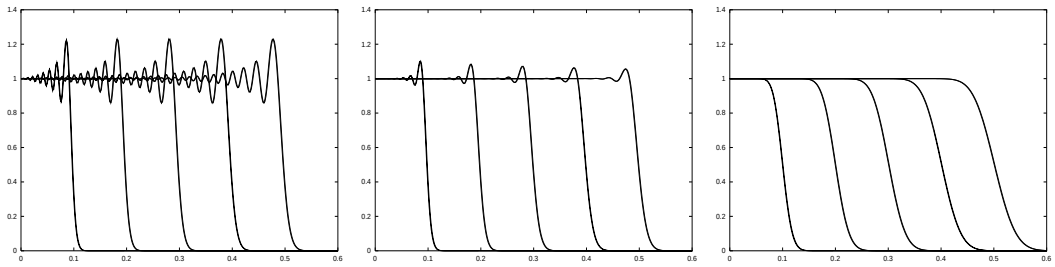


FIGURE 12.4 – Advection d'un front de gauche à droite par un champ de vitesse uniforme sur un maillage uniforme de 200 points par un schéma RK3 explicite sans stockage à $cfl = 0.5$. On utilise une stabilisation par viscosité numérique de la forme $\nu = \alpha c \Delta x$. Les figures de gauche à droite correspondent respectivement à une stabilisation avec $\alpha = 0.005$, $\alpha = 0.05$, $\alpha = 0.5$. On voit que l'introduction de la viscosité numérique supprime les oscillations, mais introduit un étalement trop important du front.

12.4 Décentrage

L'idée du décentrage est naturelle pour l'équation de transport où, comme nous l'avons dit, existe un amont et un aval. Le flot donc l'information proviennent de l'amont. On peut d'ailleurs physiquement expliquer l'instabilité des

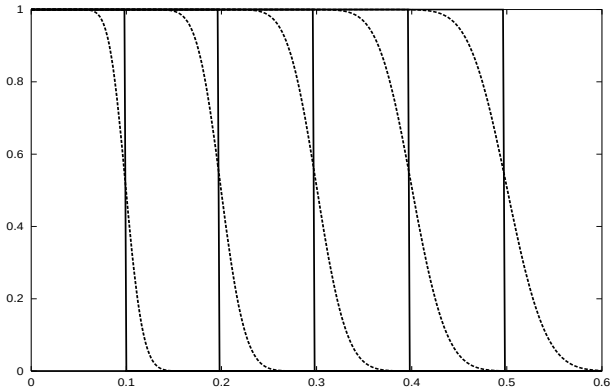


FIGURE 12.5 – Même problème d'advection de front que figure 12.4, mais cette fois par le schéma explicite décentré 12.16, où l'on compare la solution à $cfl = 0.5$ avec celle à $cfl = 1$. Dans ce dernier cas, on constate que l'on n'introduit aucun étalement de la solution (la marche est transportée sans déformation).

schémas centrés par le fait qu'ils recherchent de l'information en aval.

Les schémas décentrés vont permettre l'utilisation de discrétisations explicites en temps. On retrouvera des discrétisations équivalentes à celle des schémas explicites centrés stabilisés ci-dessus. On parlera de viscosité numérique cachée.

Considérons donc de nouveau l'équation (12.4), où l'on suppose cette fois c variable en espace :

$$\frac{\partial u}{\partial t} + c(x)\frac{\partial u}{\partial x} = 0, \quad \text{sur} \quad]0, L[\quad \text{avec} \quad u(0,x) = u^0(x), \tag{12.14}$$

$$u(t,0) = a \quad \text{si} \quad c(0) > 0, \quad \text{et} \quad u(t,L) = b \quad \text{si} \quad c(L) < 0.$$

Nous introduirons le décentrage à trois niveaux qui sont équivalents pour ce problème simple :

• soit à travers des formules de dérivation numérique décentrées,

• soit à travers une formulation volumes finis et l'utilisation du signe des flux entrant et sortant des cellules,

• soit par une modification des fonctions de base dans la formulation variationnelle du problème.

12.4.1 Décentrage par la dérivation

Pour stabiliser le schéma centré (12.5), on remplace l'approximation centrée d'ordre 2 de la dérivée première par une approximation décentrée conditionnelle

d'ordre 1. Plus exactement, on utilise

$$c(x)\frac{\partial u}{\partial x} \approx c_j \frac{u_j - u_{j-1}}{\Delta x}, \quad \text{si} \quad c_j > 0, \quad \text{et} \tag{12.15}$$

$$\approx c_j \frac{u_{j+1} - u_j}{\Delta x}, \quad \text{si} \quad c_j < 0.$$

On obtient le schéma explicite décentré :

$$\begin{array}{l} \dfrac{u_j^{n+1} - u_j^n}{\Delta t} + c_j \dfrac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \quad \text{si} \quad c_j > 0 \\ \dfrac{u_j^{n+1} - u_j^n}{\Delta t} + c_j \dfrac{u_{j+1}^n - u_j^n}{\Delta x} = 0 \quad \text{si} \quad c_j < 0 \end{array} \tag{12.16}$$

qui est d'ordre un en temps et en espace et dont on peut montrer la stabilité sous condition CFL $\leq$ 1 par l'analyse de Fourier habituelle. Au niveau de l'écriture matricielle du schéma, nous avons maintenant un terme non nul sur la diagonale au contraire d'un schéma centré. Le schéma décentré peut, comme le schéma de Lax, s'interpréter comme un schéma de caractéristiques. Mais cette fois l'interpolation du pied de la caractéristique, est faite, selon le signe de c_j entre x_{j-1} et x_j ou entre x_j et x_{j+1}. En recherchant une équation équivalente de la forme advection-diffusion (12.9), on constate que cette opération revient à introduire une viscosité numérique de type $\nu \sim |c(x)|\Delta x/2$.
En effet le schéma décentré 12.16 peut se réécrire sous la forme :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c_j \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} - |c_j| \frac{\Delta x}{2} \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = 0 \tag{12.17}$$

qui correspond à la discrétisation d'une équation d'advection-diffusion.

Remarque 12.4.1 *Il existe cependant un cas particulier où la diffusion numérique peut être complètement éliminée. En considérant le schéma décentré 12.16 dans le cas $c > 0$ constant, le choix d'un $cfl = c\Delta t/\Delta x = 1$ conduit à la solution exacte :*

$$u_j^{n+1} = u_j^n - \frac{c\Delta t}{\Delta x}(u_j^n - u_{j-1}^n) = u_{j-1}^n.$$

On montre figure (12.5) le comportement de notre front dans ce cas particulier. Cependant, cette situation ne se produit qu'en dimension un avec une vitesse c constante et un pas d'espace uniforme.

Il est possible de compenser la perte de précision en utilisant une approximation d'ordre plus élevé, basée sur un nombre plus important de points de discrétisation. La question de savoir quel est le schéma le plus précis que l'on peut construire en utilisant u_j, u_{j+1} et u_{j+2} si $c_j < 0$ et en utilisant u_j, u_{j-1} et

u_{j-2} si $c_j > 0$ se résout par application de la formule de Taylor. Dans ce dernier cas par exemple, par application de la formule de Taylor aux termes u_{j-2} et u_{j-1} on a :

$$\begin{cases} a_0 u_j = a_0 u_j, \\ a_1 u_{j-1} = a_1 (u_j - \Delta x (\frac{\partial u}{\partial x})_j + \frac{\Delta x^2}{2} (\frac{\partial^2 u}{\partial x^2})_j + \Delta x^2 o(1)), \\ a_2 u_{j-2} = a_2 (u_j - 2\Delta x (\frac{\partial u}{\partial x})_j + 4\frac{\Delta x^2}{2} (\frac{\partial^2 u}{\partial x^2})_j + \Delta x^2 o(1)) \end{cases}$$

Après sommation, on cherche à annuler le plus grand nombre de termes possible. On obtient ainsi le système algébrique 3×3 suivant :

$$a_0 + a_1 + a_2 = 0, \quad a_1 + 4a_2 = 0 \quad -(a_1 + 2a_2) = 1.$$

Ce qui permet d'obtenir l'approximation d'ordre deux :

$$(\frac{\partial u}{\partial x})_j \approx \frac{3u_j - 4u_{j-1} + u_{j-2}}{2\Delta x} \quad \text{si} \quad c_j > 0, \tag{12.18}$$

avec une erreur de troncature de $\Delta x^2 u_{xxx}/3$ qui est le double de celle d'une discrétisation centrée ($\Delta x^2 u_{xxx}/6$). Bien entendu, pour appliquer ce schéma, on doit disposer de deux points en aval et amont. Le traitement des frontières est donc une des difficultés majeures lors de la mise en oeuvre des schémas d'ordre élevé. En pratique, pour les points frontières, on utilisera un schéma d'ordre inférieur, n'exigeant qu'un seul point.

En expérimentant ce schéma pour l'advection d'un front (figure (12.6)), on constate qu'il fait apparaître des oscillations à l'aval. On ajoute un point de discrétisation en aval pour les supprimer. On utilise alors, à nouveau, la méthode d'identification par développement de Taylor pour les variables u_{j+1}, u_j, u_{j-1} et u_{j-2} si $c_j > 0$, et on obtient le schéma d'ordre 3 :

$$(\frac{\partial u}{\partial x})_j \approx \frac{2u_{j+1} + 3u_j - 6u_{j-1} + u_{j-2}}{6\Delta x} \tag{12.19}$$

Schémas compacts ou de Padé

On peut améliorer encore la précision des schémas en considérant les dérivées comme des variables indépendantes. Voici, par exemple, la combinaison suivante, appelée construction de Padé :

$$(\frac{\partial u}{\partial x})_j = a_0 u_j + a_1 u_{j+1} + a_2 u_{j+2} + a_3 (\frac{\partial u}{\partial x})_{j+1} + a_4 (\frac{\partial u}{\partial x})_{j+2} + \Delta x^3 o(1).$$

On constate que la construction est implicite. On doit résoudre un système pour déterminer les dérivées $(\frac{\partial u}{\partial x})_j$. Ayant introduit deux nouveaux paramètres, on peut

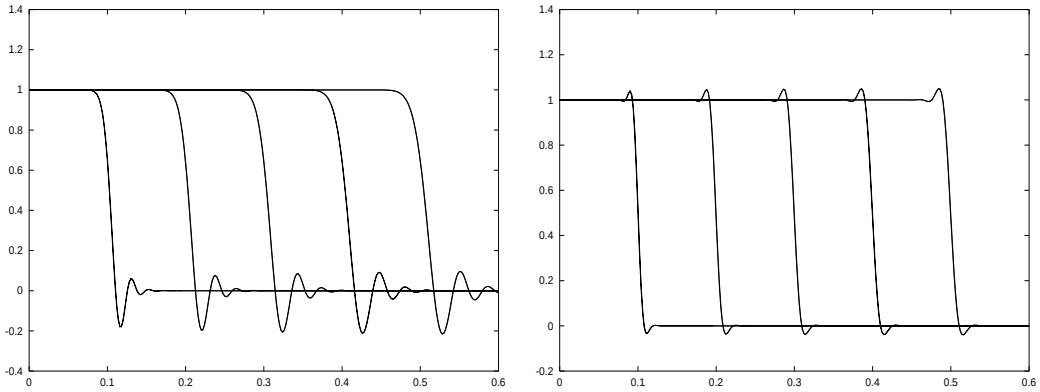


FIGURE 12.6 – Même problème d'advection de front de gauche à droite en utilisant les schémas (12.18) et (12.19). L'étalement du front constaté figure (12.4) a été évité par le contrôle plus rigoureux de la viscosité numérique. Mais avec un schéma à trois points placés uniquement en amont, on obtient des oscillations en aval qui disparaîssent en introduisant un point de discrétisation en aval. On voit que l'ordre plus élevé du schéma n'est pas nécessairement synonyme d'une robustesse moindre.

aboutir à un schéma d'ordre 4 en annulant deux autres termes par substitution dans le développement de Taylor. On trouve en particulier,

$$a_0 = 0, \quad a_1 = \frac{-3}{4\Delta x}, \quad a_2 = \frac{3}{4\Delta x}, \quad a_3 = a_4 = \frac{1}{4},$$

avec une erreur de troncature de la forme $\frac{\Delta x^4}{30}(\partial^5 u/\partial x^5)_j$. Ainsi, après résolution du système algébrique suivant :

$$4(\frac{\partial u}{\partial x})_j - (\frac{\partial u}{\partial x})_{j+1} - (\frac{\partial u}{\partial x})_{j+2} = \frac{3}{\Delta x}(u_{j+2} - u_{j+1}), \tag{12.20}$$

on aboutit directement aux dérivées. On peut alors combiner (12.20), avec l'intégration en temps de (12.4). Ici aussi nous sommes confrontés à la difficulté du traitement des points frontières où l'on est amené à utiliser une discrétisation moins précise.

Remarque 12.4.2 *Une précaution, utile lors de la mise au point des discrétisations par ces méthodes : il faut toujours vérifier que la somme des coefficients s'annule.*

On constate de nombreuses difficultés de résolution pour cette équation si simple. Il faut penser que des discrétisations de types (12.19) sont presqu'impossibles à mettre en oeuvre, en dimension deux et trois, sur des maillages non-structurés. L'adaptation de maillage (chapitre 18) sera alors une aide précieuse, pour aboutir à des solutions précises en utilisant des schémas d'ordre faible.

Remarque 12.4.3 *On est amené, en pratique, à utiliser des discrétisations et des maillages différents selon les quantités à calculer dans les modèles physiques complexes. Par exemple, en aéro-acoustique, le maillage du calcul aérodynamique devra être raffiné dans les couches limites, alors que le maillage sera régulier pour le calcul acoustique.*

12.4.2 Décentrage de la variable en volumes finis

Le principe de la méthode des volumes finis réside dans l'utilisation de la formule de la divergence sur la forme conservative de l'équation de transport. En considérant le domaine discrétisé $\Omega_h = \cup_j C_j$ comme réunion de cellules (volumes finis), ceci nous amène à :

$$\int_{\Omega_h} \frac{\partial u}{\partial t} dx + \int_{\Omega_h} \frac{\partial u\vec{V}}{\partial x} dx = \Sigma_j \int_{C_j} \frac{\partial u}{\partial t} + \Sigma_j \int_{\partial C_j} (u\vec{V}).n d\sigma = 0,$$

où n est la normale unitaire aux bords ∂C_j orientée vers l'extérieur des cellules. L'utilisation de la condensation des masses sur les intégrales de la dérivée temporelle nous donne, en dimension un, avec $\vec{V} = c$, les équations algébriques suivantes pour les noeuds du maillage :

$$|C_j|(\frac{\partial u}{\partial t})_j + ((cu)|_{j+1/2} - (cu)|_{j-1/2}) = 0,$$

où les C_j sont les cellules formées par les intervalles délimités par les milieux des éléments notés $j + 1/2$ et $j - 1/2$. Dans le cas d'un maillage uniforme et l'utilisation d'un schéma d'intégration explicite en temps on obtient :

$$u_j^{n+1} - u_j^n + \frac{\Delta t}{\Delta x}((cu)|_{j+1/2}^n - (cu)|_{j-1/2}^n) = 0. \tag{12.21}$$

Le décentrage se fait alors par le choix de l'approximation en $j + 1/2$ et $j - 1/2$ pour les termes cu. En particulier, l'utilisation de la moyenne correspond à une discrétisation centrée. De même, l'approximation suivante pour $(cu)|_{j+1/2}$ aboutit à un schéma d'ordre un en espace et introduit le décentrage adéquat suivant le signe de la vitesse en $j + 1/2$:

$$\left\{ \begin{array}{l} (cu)|_{j+1/2}^n = \max(0, c_{j+1/2})u_j^n + \min(0, c_{j+1/2})u_{j+1}^n, \quad \text{avec} \quad c_{j+1/2} = \frac{c_{j+1}+c_j}{2} \\ (cu)|_{j-1/2}^n = \max(0, c_{j-1/2})u_{j-1}^n + \min(0, c_{j-1/2})u_j^n, \quad \text{avec} \quad c_{j-1/2} = \frac{c_{j-1}+c_j}{2} \end{array} \right. \tag{12.22}$$

Les schémas d'ordre un introduisent trop de diffusion numérique (figure (12.4)). On peut réduire le niveau de décentrage pour améliorer la précision du schéma,

en considérant par exemple pour $(cu)|_{j+1/2}$ une construction de la forme :

$$\begin{aligned}(cu)|^n_{j+1/2} = &\max(0, c_{j+1/2})(\alpha u^n_j + (1-\alpha)u^n_{j+1}) \\ &+ \min(0, c_{j+1/2})((1-\alpha)u^n_{j+1} + \alpha u^n_j),\end{aligned} \tag{12.23}$$

avec $0.5 \leq \alpha \leq 1$. Le choix $\alpha = 0.5$ correspond à un schéma centré, mais avec une erreur de consistance car $c_{j+1/2}u^n_{j+1/2} \neq (cu)|^n_{j+1/2}$. De même, avec $\alpha = 1$ on retrouve le schéma totalement décentré ci-dessus. On présente, figure 12.7, le résultat de l'advection de notre front pour trois valeurs de α.

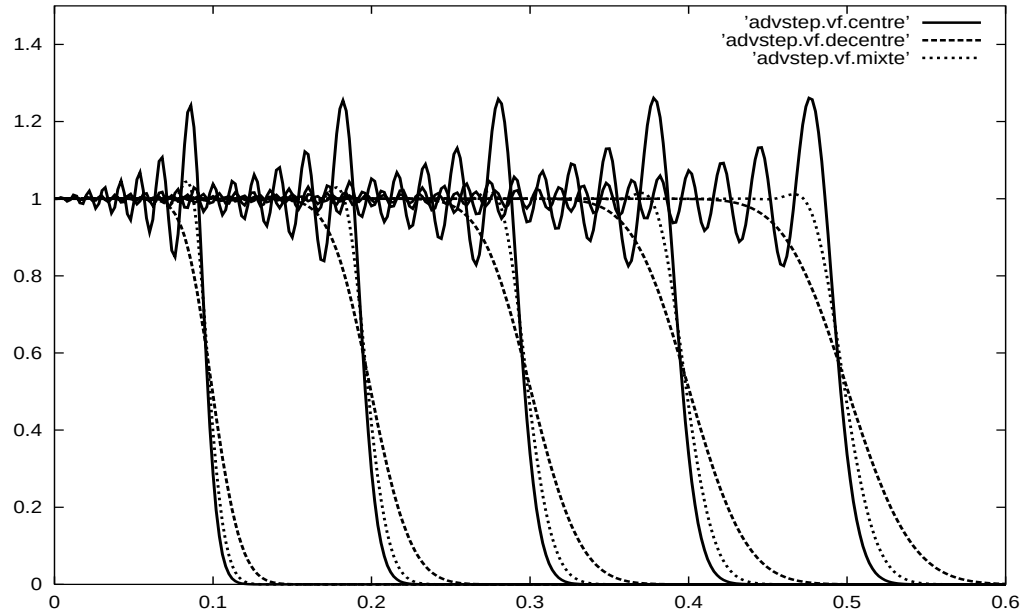


FIGURE 12.7 – Advection d'un front de gauche à droite en volumes finis et en utilisant la construction (12.23) pour respectivement, $\alpha = 0.5$, $\alpha = 1$ et $\alpha = 0.65$. Bien entendu, le choix centré est instable et le décentré total trop dissipatif.

La précision des schémas peut être améliorée par des constructions géométriques en utilisant des développements de Taylor locaux à partir des noeuds voisins. Cependant, ces constructions sont difficiles, surtout pour des maillages non-structurés. De plus, elles font perdre au schéma son caractère compact (utilisant uniquement des informations locales). Ces constructions sont appelées MUSCL (voir bibliographie).

12.4.3 Décentrage par la fonction de base en éléments finis

Pour présenter la technique du décentrage dans une formulation éléments finis, nous allons considérer l'équation continue équivalente (de type advection-diffusion) pour une discrétisation décentrée d'ordre 1, appliquée au schéma

d'Euler explicite :

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} - c\frac{\Delta x}{2}\frac{\partial^2 u}{\partial x^2} = 0.$$

On a vu (équation 12.17) que la discrétisation de cette équation par un schéma centré redonne le schéma de discrétisation décentré d'ordre 1.

En multipliant l'équation équivalente par une fonction test ϕ et en intégrant sur le domaine de résolution on a :

$$\int_\Omega \phi(\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} - c\frac{\Delta x}{2}\frac{\partial^2 u}{\partial x^2})dx = 0.$$

On choisit l'espace des fonctions tests tel que l'intégrale de bord provenant de l'intégration par parties s'annule :

$$\int_\Omega \frac{\partial u}{\partial t}\phi + c\frac{\partial u}{\partial x}\psi = 0, \quad \text{avec} \quad \psi = \phi + c\,\frac{\Delta x}{2}\phi_x.$$

Ainsi, on constate que le décentrage peut être introduit par un changement de fonction de base appliquée à la dérivée première en espace. On a naturellement une perte de consistance du schéma numérique dans la mesure où la même fonction test n'a pas été utilisée pour tous les termes de l'équation.

La mise au point de discrétisations d'ordre élevé se ferait alors en augmentant le degré des fonctions de bases dans la formulation éléments finis. Cette approche est très générale, mais à notre connaissance, les solveurs industriels utilisent presqu'exclusivement des discrétisations d'ordre peu élevé et principalement du $P0$ ou du $P1$ pour la résolution des problèmes hyperboliques d'ordre 1. Ceci est lié au fait que, pour les systèmes notamment et en présence de discontinuités, il existe des résultats fondamentaux qui montrent que la stabilité impose le choix de faibles niveaux de précision pour les schémas (voir Smoller et Godlewski-Raviart). On peut partiellement remédier à ce problème en utilisant les schémas ENO (Essentially non oscillatory) qui permettent d'obtenir une plus grande précision tout en gardant des oscillations bornées.

Schémas distributifs

La généralisation de l'approche précédente a conduit à une nouvelle classe de schémas appelés schémas distributifs. On évalue la contribution de chaque élément et le décentrage se fait par le choix des coefficients de distribution des flux aux noeuds. On obtient :

$$u_j^{n+1} = u_j^n + \frac{\Delta t}{\Delta x}\Sigma_{T, j\in T}\beta_j^T \Phi_T, \tag{12.24}$$

où la sommation porte sur les éléments T ayant le sommet j en commun. Le flux Φ_T est évalué sur l'élément T et β_j^T est le coefficient de distribution de la

contribution de cet élément au noeud j. La conservation des flux impose que les coefficients vérifient :

$$\Sigma_j \beta_j^T = 1, \quad \text{somme sur les sommets } j \text{ de } T.$$

A titre d'exemple, en dimension 1, les éléments ayant le noeud j en commun sont les segments $(j-1, j)$ et $(j, j+1)$ et

$$\Phi_T = -\int_T (c\frac{\partial u}{\partial x})\phi dx \sim \Delta x (c\frac{\partial u}{\partial x})_T.$$

Un schéma centré correspond au choix $\beta_j^T = 1/2$ et un schéma totalement décentré à celui de :

$$\beta_j^{(j-1,j)}\Phi_{(j-1,j)} = 0, \quad \text{et} \quad \beta_j^{(j,j+1)}\Phi_{(j,j+1)} = 1 \quad \text{si} \quad c_{j-1} - c_j > 0,$$

$$\beta_j^{(j-1,j)}\Phi_{(j-1,j)} = 1, \quad \text{et} \quad \beta_j^{(j,j+1)}\Phi_{(j,j+1)} = 0 \quad \text{si} \quad c_j - c_{j+1} < 0.$$

On obtient ainsi des discrétisations compactes, utilisant des structures de données simples. L'implémentation aisée de ces schémas et leur complexité faible les ont popularisés pour les équations scalaires. Cependant le traitement des systèmes reste difficile. En effet, en général on ne peut pas diagonaliser les systèmes. Ce qui ne permet pas de considérer de façon indépendante les variables et donc de se ramener à des problèmes sur des variables scalaires.

12.4.4 Décentrage par les caractéristiques

La méthode des caractéristiques permet l'introduction naturelle du décentrage. En effet, connaissant la caractéristique rétrograde $X(\tau)$ du champ c passant par x, solution de :

$$\frac{dX}{d\tau} = -c, \quad X(\tau = T) = x, \tag{12.25}$$

on peut résoudre (12.4) de façon exacte, car la solution $u(x, T)$ en un point x après un temps T est donnée par la valeur $u(X(\tau = 0), t = 0)$ de la variable au pied $X(\tau = 0)$ de la caractéristique passant par x. On peut généraliser ceci à tous les points et obtenir :

$$u(x, t) = u(X(0)), \quad X(\tau = T) = x, \quad \forall t, \forall x. \tag{12.26}$$

Malheureusement, une implémentation efficace de cette méthode n'est pas aisée. Lorsque l'on discrétise (12.25) et (12.26), on introduit plusieurs approximations. Par exemple, si la discrétisation est linéaire par morceaux, on aboutira à une construction par segments de la caractéristique et si le maillage n'est pas

adapté au champ de vitesse, cette construction introduira une erreur importante. Par ailleurs, pour trouver la valeur au pied de la caractéristique, il faut interpoler en utilisant les valeurs aux noeuds et si localement la qualité du maillage n'est pas bonne, nous transporterons cette erreur loin en aval. Ces deux niveaux d'erreurs font que la méthode des caractéristiques est l'une des rares méthodes à composer les erreurs et donc, souvent, à les amplifier.

On peut atténuer ces deux effets en utilisant l'adaptation de maillages décrite au chapitre 18. Mais l'adaptation devra prendre en compte, à la fois, le champ et la quantité advectée.

La méthode des caractéristiques est, par contre, largement utilisée pour la discrétisation des conditions aux limites en entrée et en sortie. La bonne information étant naturellement prise en compte par la méthode. De plus, l'application de la méthode dans ce cas étant locale (i.e. à la frontière uniquement), les défauts mentionnés ci-dessus ne sont pas pénalisants. C'est donc une bonne méthode de prise en compte des conditions de sortie sur la frontière "infinie" aval en mécanique des fluides.

12.5 Monotonie et positivité

Les quantités advectées représentent souvent des quantités physiques positives (énergie, densité,...). On recherche donc des schémas garantissant cette propriété. Plus exactement, si $u^0(x) \geq 0 \, \forall x$ on demande que $u(x,t) \geq 0, \forall t, \forall x$. On recherche, de façon plus générale, la monotonie des solutions, c'est à dire que l'on veut que le schéma vérifie :

$$\min_j(u_j^n) \leq u_j^{n+1} \leq \max_j(u_j^n) \quad \forall n, \, \forall i$$

Sur la figure (12.6) on voit clairement que, pour certains schémas numériques, le maximum et minimum de la quantité advectée peuvent dépasser les valeurs limites admissibles.

Monotonie par limiteurs ou projection

La première technique pour assurer l'admissibilité de la solution est d'introduire des limiteurs ou "cut-off". 0n projette les prédictions dans le domaine admissible (nous verrons d'autres applications des projections en optimisation au chapitre 17).

Voici l'application du schéma (12.18) ci-dessus à notre équation d'advection, en utilisant une méthode de Runge-Kutta sans stockage à trois pas, où nous avons introduit un opérateur de projection :

$$u_j^0 = \text{donné},$$
$$\text{Pour } n = 0, .., nmax \text{ faire},$$
$$v_j^0 = u_j^n,$$
$$\text{Pour } k = 1, .., 3 \text{ faire},$$

$$(v_x)_j^k = \frac{3v_j^k - 4v_{j-1}^k + v_{j-2}^v}{2\Delta x} \quad \text{si} \quad c_j > 0,$$

$$(v_x)_j^k = \frac{-3v_j^k + 4v_{j+1}^k - v_{j+2}^v}{2\Delta x} \quad \text{si} \quad c_j < 0,$$

$$v_j^{k+1} = u_j^n + \tfrac{\Delta t}{3-k+1}(c(v_x)_j^k),$$

$$v_j^{k+1} = \min(\max(v_j^{k+1}, \min(u^n)), \max(u^n)), \tag{12.27}$$

$$\text{Fin } k,$$
$$u_j^{n+1} = v_j^3,$$
$$\text{Fin } n.$$

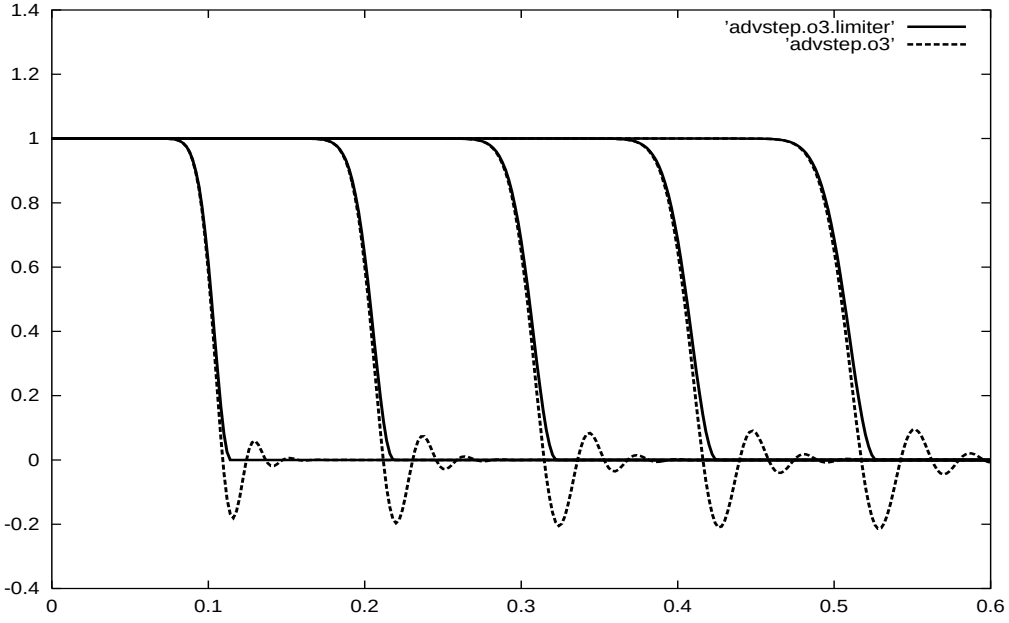


FIGURE 12.8 – Application de l'algorithme ci-dessus à l'advection d'un front de gauche à droite. L'introduction des limiteurs permet de garantir la positivité de la quantité advectée mais s'accompagne d'une légère augmentation de la dissipation.

Cas des volumes finis

Le schéma volumes finis (12.21-12.22) garantit la positivité. En effet, considérons pour simplifier $c > 0$ et constante, on obtient :

$$u_j^{n+1} = u_j^n - \frac{c\Delta t}{\Delta x}(u_j^n - u_{j-1}^n).$$

On retrouve le schéma décentré d'ordre un dont la condition de stabilité est : $CFL = \frac{c\Delta t}{\Delta x} \leq 1$. Ce schéma peut s'écrire

$$u_j^{n+1} = (1 - \frac{c\Delta t}{\Delta x})u_j^n + \frac{c\Delta t}{\Delta x}u_{j-1}^n.$$

Donc avec $0 \leq \frac{c\Delta t}{\Delta x} \leq 1$, on observe que u_j^{n+1} est une combinaison convexe de u_j^n et de u_{j-1}^n. Ce qui donne bien $\min_j(u_j^n) \leq u_j^{n+1} \leq \max_j(u_j^n)$.
Pour le schéma général (12.23), la positivité n'est plus automatique pour les cas non totalement décentrés ($\alpha \neq 1$). La positivité ne sera assurée par la construction (12.23) que sous certaines conditions supplémentaires. Avec $c > 0$ constante, par exemple, il faudrait, en plus de la condition $CFL \leq 1$, vérifier

$$\frac{c\Delta t}{\Delta x}[\alpha(u_j^n - u_{j-1}^n) + (1-\alpha)(u_{j+1}^n - u_j^n)] \leq u_j^n$$

Ceci introduit une difficulté, par exemple, dans le cas de l'advection d'un front. En effet, si le front se trouve en $j-1$, alors $u_j^n = 0$ et le pas de temps s'annule.

Pour éviter ces difficultés, l'approche la plus utilisée consiste à utiliser l'opérateur de projection (12.27).

Cas des caractéristiques

La méthode des caractéristiques assure la positivité de la quantité advectée si l'opérateur d'interpolation utilisé au pied de la caractéristique $X(\tau = 0)$ assure la positivité de l'interpolé. Ce qui est le cas par exemple pour une interpolation de type P1 utilisant les coordonnées barycentriques du pied de la caractéristique dans un élément et les valeurs aux noeuds de la variable. En dimension un, ceci nous donne :

$$u_{X(\tau=0)} = (\zeta u_j + (1-\zeta)u_{j+1}),$$

où $0 \leq \zeta \leq 1$, désigne la coordonnée barycentrique dans l'élément $(j, j+1)$ du pied de la caractéristique. Ainsi, $u_{X(\tau=0)} > 0$ si $u_j > 0$ et $u_{j+1} > 0$.

Par contre, si on utilise une construction plus complexe, basée, par exemple, sur des interpolations de degré plus élevé (P2) ou des développements de Taylor locaux, on n'est plus assuré de cette positivité. Considérons la construction suivante :

$$u_{X(\tau=0)} = 0.5(u_\zeta^- + u_\zeta^+),$$

$$u_\zeta^- = u_j + \zeta\Delta x(\frac{\partial u}{\partial x})_j, \quad u_\zeta^+ = u_{j+1} - (1-\zeta)\Delta x(\frac{\partial u}{\partial x})_{j+1}.$$

Ces deux quantités n'étant pas nécessairement positives, la positivité du schéma n'est pas garantie. Il faudra donc appliquer l'opération de projection (12.27) à u_ζ^+ et u_ζ^-.

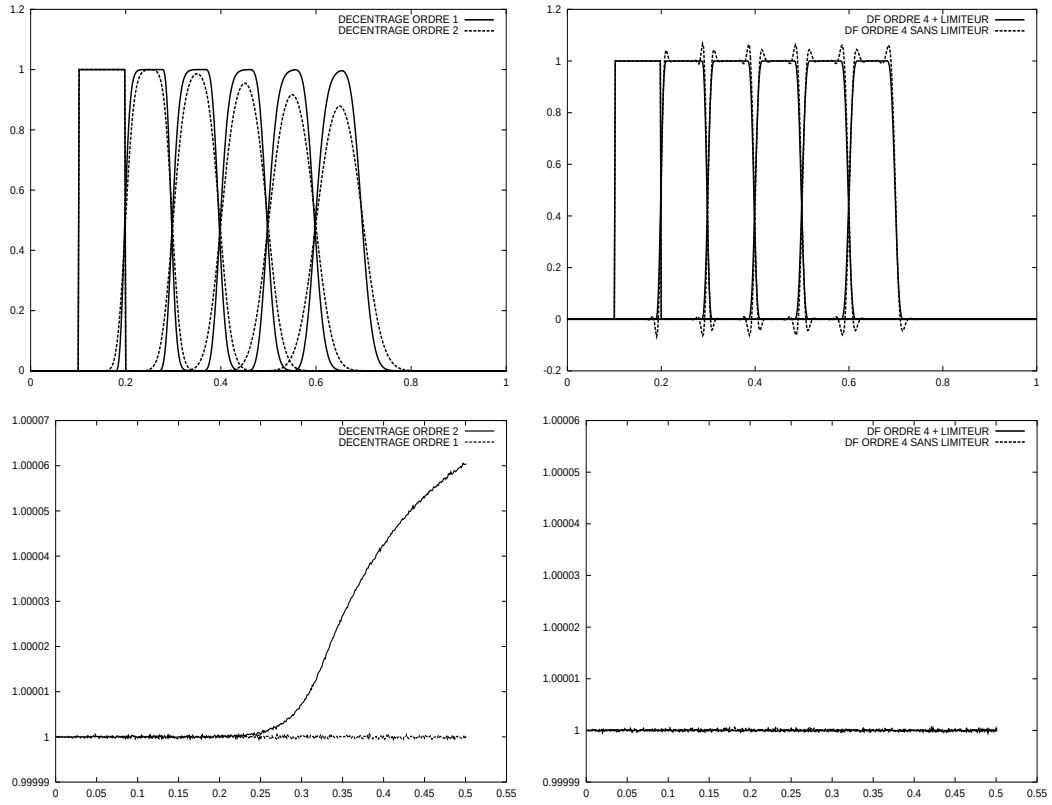


FIGURE 12.9 – Transport d'une porte de gauche à droite par un champ de vitesse uniforme et évolution de $\int_\Omega u dx$ en temps. A gauche, l'utilisation d'un schéma totalement décentré lisse le profil mais est conservatif. L'augmentation de l'ordre du schéma a impliqué l'utilisation de limiteur pour préserver la positivité ; on perd légèrement la conservation (figure en bas - à gauche). Avec un schéma aux différences finis d'ordre 4, la monotonie n'est pas assurée. Après l'introduction des mêmes limiteurs la conservation est assurée (figure en bas - à droite), car la perte de monotonie se fait de façon symétrique. De ce fait, la projection n'a pas d'effet sur la conservation dans ce cas : il n'y a presque pas de déformation de la porte.

Cas des schémas distributifs

De même certains schémas distributifs assurent la positivité. Considérons l'expression (12.24) que l'on peut réécrire sous la forme :

$$u_j^{n+1} = u_j^n + \frac{\Delta t}{\Delta x}\Sigma_{T,j\in T}\beta_j^T\Phi_T = \Sigma_{k_j}c_{k_j}u_{k_j}^n, \tag{12.28}$$

où nous avons exprimé, en introduisant des coefficients c_{k_j} qui dépendent des pas de temps et d'espace, u^{n+1} en fonction de u^n. Les indices k_j désignent les noeuds intervenant dans le schéma de discrétisation de Φ_T. En particulier, si le schéma est compact, ceci fait intervenir uniquement les noeuds de l'élément T. Ainsi, on constate que si la discrétisation est telle que les coefficients vérifient :

$$c_{k_j} \geq 0, \quad \Sigma_{k_j}c_{k_j} = 1, \quad \forall j,$$

la positivité de u^{n+1} sera assurée. Ces relations introduisent donc une condition supplémentaire locale sur le pas de temps.

12.6 Conservation

Une autre propriété fondamentale lors du choix d'une discrétisation de l'équation de transport est la capacité du schéma à assurer la conservation de la quantité advectée et la validité de (12.3).

La recherche de schémas positifs et conservatifs ne concerne pas uniquement les problèmes hyperboliques du premier ordre, mais c'est pour ces équations que leur réalisation est la plus difficile.

Les schémas volumes finis et les schémas distributifs sont naturellement conservatifs. Considérons le transport d'une porte par un champ de vitesse uniforme. Nous comparons (figure 12.9) des schémas volumes finis 12.21-12.22 et 12.23 avec limiteurs (12.27)) et un schéma différences finies d'ordre 4 (12.19).

En conclusion, les volumes finis et les schémas distributifs sont bien adaptés aux problèmes hyperboliques du premier ordre. Les différences finies sont intéressantes si les géométries sont simples. L'utilisation des discrétisations implicites permet souvent de stabiliser les calculs, mais ne garantit en général ni la monotonie, ni la conservation.

12.7 Erreur de phase, erreur de vitesse de groupe

Dans le cas de l'équation de transport, comme dans le cas de l'équation des ondes, il existe une autre source d'erreur et d'instabilité, moins évidente que

l'erreur de troncature ou que l'erreur commise sur la conservation de la quantité transportée. Il s'agit de l'erreur commise sur la vitesse d'advection.

Pour certains problèmes, la solution exacte à un instant précis, n'est pas forcément significative, et, l'utilisateur peut être simplement intéressé par des moyennes temporelles. Par contre, dans d'autres cas, pensons aux prévisions météorologiques, à la dérive d'une pollution, à l'étude fine du décollage ou de l'aterrissage d'un avion, la détermination précise de la solution au temps exact est le résultat recherché.

L'erreur de discrétisation qui entraînerait un décalage sur la vitesse de propagation, produisant un effet de retard ou d'avance, serait alors très génante, d'autant plus que ce type d'erreur est plus difficile à repérer qu'une erreur de module.

Or, les schémas numériques produisent, selon le cas, des résultats en avance ou en retard sur les solutions exactes. Considérons, par exemple, l'équation de transport linéaire en dimension un.

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial t} = 0$$

avec $c > 0$. La solution exacte avance avec la vitesse c.

Pour une solution initiale en $u(x,0) = e^{ikx}$, la solution au temps t vérifie : $u(x,t) = e^{ik(x-ct)} = e^{-ikct}e^{ikx}$. Le facteur multiplicatif faisant passer de la solution au temps t_n à la solution au temps t_{n+1}, sera donc $e^{-ikc\Delta t}$. C'est un complexe de module un et d'argument $-kc\Delta t$. C'est l'erreur sur cet argument qui entraîne l'erreur de vitesse.

Considérons, par exemple, le schéma d'Euler implicite :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\,\frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{2\Delta x} = 0. \tag{12.29}$$

Son coefficient d'amplification

$$G_k = \frac{1}{1 + i\frac{c\Delta t}{\Delta x} sin(k\Delta x)}$$

a un module < 1 et donc le schéma est dissipatif. Mais, de plus, l'argument de G_k

$$Arg(G_k) = -Atan(\frac{c\Delta t}{\Delta x} sin(k\Delta x))$$

n'est pas égal à l'argument exact $-kc\Delta t$. Un développement limité, pour Δx infiniment petit, donne :

$$Arg(G_k) = -[kc\Delta t - \frac{c\Delta t}{\Delta x}\frac{k^3\Delta x^3}{6} - \frac{c^3k^3\Delta t^3}{3}] + ...$$

Tout se passe comme si ce schéma produisait une convection à une vitesse inférieure à la bonne vitesse c. On dit que le schéma d'Euler implicite retarde.

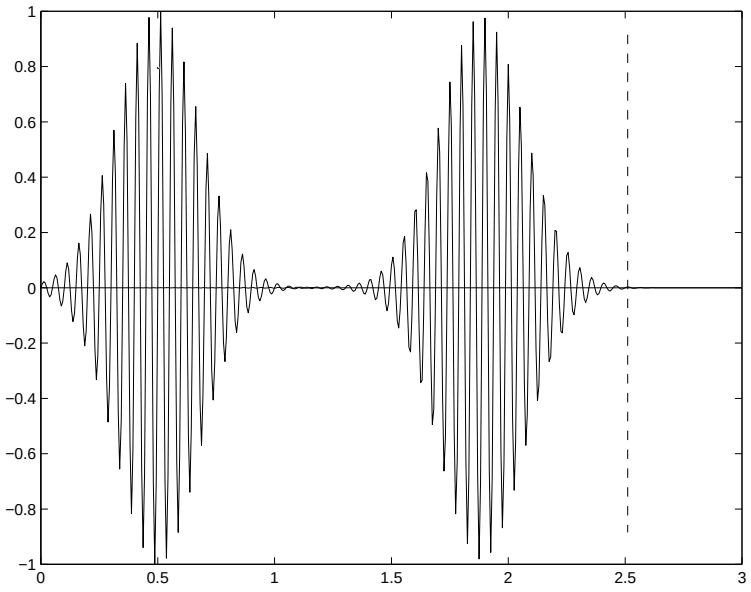


FIGURE 12.10 – Effet retard dû à une erreur de vitesse de groupe

Dans le cas plus complet du transport d'un signal non-monochromatique, deux phénomènes se conjuguent, l'erreur de phase et la dispersion selon les longueurs d'onde, pour produire un effet global d'avance ou de retard.

Il faut alors considérer le phénomène d'erreur de vitesse de groupe. Voici, figure 12.10, un exemple proposé par Lloyd N. Trefethen (SIAM vol 24 n 2, Avril 1982). Le signal

$$u_0(x) = e^{-16(x-\frac{1}{2})^2}\, sin(40\pi x)$$

est transporté avec la vitesse 1. On utilise le schéma de Crank-Nicolson avec $\Delta x = 1/160$ et $\Delta t = 0.4\Delta x$. Le programme a été réalisé en Matlab. On observe un important effet de retard. Au temps $t = 2$, le signal devrait se trouver centré sur la ligne hachurée.

12.8 Équations non-linéaires et linéarisées

Nous allons présenter une extension possible des schémas décentrés à la résolution d'équations non-linéaires. Ce n'est pas la seule approche possible et il existe une littérature abondante sur la résolution des équations hyperboliques non-linéaires du premier ordre (voir Godlewski-Raviart et Smoller), ainsi que sur celle des systèmes d'équations. Cette résolution est souvent délicate.

Considérons l'équation de Burgers ci-dessous, que nous retrouverons au chapitre 17 dans le cas d'un problème inverse.

$$\left\{ \begin{array}{l} \dfrac{\partial u}{\partial t} + 0.5(u^2)_x = 0.3xu, \quad \text{sur} \quad (-1,1) \\ u(t,-1) = 1, \quad u(t,1) = -0.8, \quad u(0,x) = -0.9x + 0.1 \end{array} \right. \tag{12.30}$$

Cette équation est une EDP hyperbolique non linéaire d'ordre 1. Nous allons nous intéresser à la résolution de cette équation et de l'équation linéarisée correspondante. La solution stationnaire exacte de cette équation vérifie

$$\frac{\partial u}{\partial x} = 0.3x$$

dans les zones régulières. Elle est donc quadratique par morceaux et a un saut en x_s :

$$u(x) = 0.15x^2 + 0.85 \quad \text{pour } x < x_s,$$
$$u(x) = 0.15x^2 - 0.95 \quad \text{pour } x > x_s,$$

La position du choc x_s est telle que :

$$u_s^- = -u_s^+ \quad \rightarrow \quad x_s = -\sqrt{1/3}.$$

Elle est obtenue en constatant que pour la solution stationnaire ($\frac{\partial u}{\partial t} = 0$), le saut de (u^2) est nul : $[(u^2)^+ - (u^2)^-] = (u^+ - u^-)(u^+ + u^-) = 0$ en x_s.

La résolution de cette équation peut s'effectuer en considérant la forme non conservative

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0.3xu$$

et en utilisant l'un des schémas stabilisés, vus ci-dessus, avec une intégration en temps explicite RK3 par exemple. On écrit alors l'itération suivante pour $p = 0, 1, 2, \quad n = 1, ...$:

$$\begin{array}{l} u_j^0 = \text{donné}, \\ u_j^{n+1,p+1} = u_j^n - \frac{\Delta t}{3-p+1}\left(u_j^{n,p} \frac{u_{j+1}^{n,p} - u_{j-1}^{n,p}}{2\Delta x} - \nu_j^{n,p} \frac{u_{j+1}^{n,p} - 2u_j^{n,p} + u_{j-1}^{n,p}}{\Delta x^2} - 0.3x_j u_j^{n,p} \right) \\ u_j^{n+1} = u_j^{n,3} \end{array} \tag{12.31}$$

où l'on choisit une viscosité numérique donnée par :

$$\nu_j^{n,p} = \frac{\max(|u_{j-1}^{n,p}|, |u_j^{n,p}|, |u_{j+1}^{n,p}|)\Delta x}{2}.$$

Par ailleurs, on utilise une extension de la condition de stabilité (12.10) prenant en compte la variation de u et la présence du terme source :

$$\Delta t = \min_j \left(\frac{\Delta x}{\max(|u_{j-1}^n|, |u_j^n|, |u_{j+1}^n|, 0.3\Delta x |x_j u_j^n|)}\right).$$

Considérons maintenant l'équation de Burgers linéarisée décrivant les petites

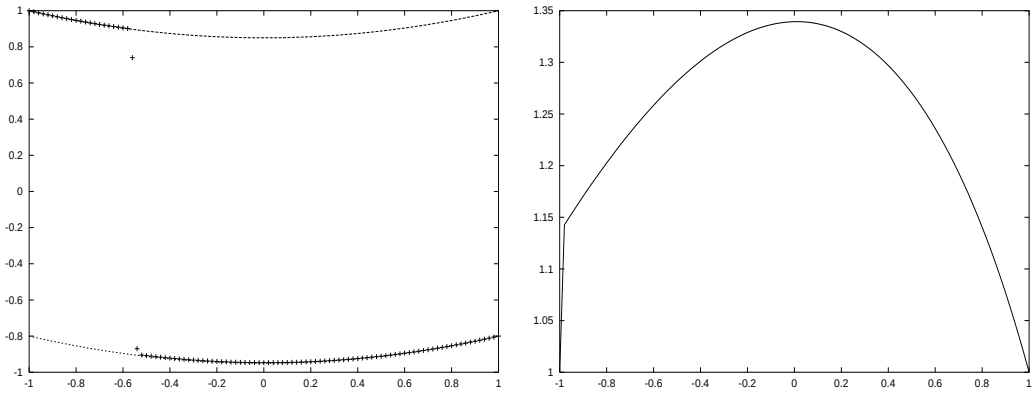


FIGURE 12.11 – A gauche : résolution de l'équation de Burgers (12.30) avec le schéma (12.31) sur un maillage régulier de 100 points avec comparaison avec la solution exacte sur les zones où celle-ci est régulière. Le résultat est assez satisfaisant car on capture le choc sur deux éléments. A droite : solution stationnaire de l'équation de Burgers linéarisée (12.32) au voisinage de la solution de l'équation de Burgers (12.30). La solution reste bornée, l'équation est donc stable.

perturbations v, en partant de la forme conservative (12.30) :

$$(u+v)_t + 0.5((u+v)^2)_x = 0.3x(u+v),$$

u étant solution, en négligeant le terme d'ordre 2, on obtient :

$$v_t + (uv)_x = 0.3xv, \quad v(t, x=-1) = v(t, x=1) = 1, \quad v(t=0, x) = 1. \quad (12.32)$$

Le champ de vitesse u est entrant à gauche et à droite, c'est pourquoi il faut deux conditions aux limites sur v. L'équation de Burgers linéarisée est une équation d'advection linéaire en v avec une vitesse d'advection discontinue u solution de

l'équation de Burgers. Cette équation se réduit dans les zones où la solution de l'équation de Burgers est régulière à $v_t + uv_x = 0$.

La solution couplée d'une équation non-linéaire et de sa version linéarisée a de nombreuses applications, en particulier, lorsque les petites perturbations peuvent donner naissance à des modifications du phénomène principal, soit en raison de la non-linéarité, soit par leur accumulation. On peut citer, par exemple, la simulation d'un jet en combustion turbulente ou la propagation des ondes acoustiques provenant des fluctuations de pressions. Parfois ces perturbations peuvent entrer en résonance avec le phénomène principal. On utilise aussi les perturbations acoustiques en contrôle de combustion.

12.9 Application aux systèmes

Les systèmes d'équations couplées non-linéaires hyperboliques d'ordre 1 interviennent dans de nombreuses applications. Considérons par exemple le système d'Euler en dynamique des gaz en une dimension d'espace, pour les variables densité ρ, quantité de mouvement ρu et énergie totale $\rho E = \rho(C_v T + 0.5u^2)$.

$$\left\{ \begin{array}{c} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2 + p)_x = 0, \\ (\rho E)_t + ((\rho E + p)u)_x = 0. \end{array} \right.$$

où la pression est donnée par la loi des gaz parfaits par exemple ($p = \rho \mathcal{R} T$). Prenons comme domaine de calcul l'intervalle $(0, 1)$. Nous nous intéressons à la solution du problème de Riemann suivant (appelé tube à choc de Sod) pour ce système :

$$(\rho, u, p)(t = 0) = (1, 0, 1), \quad \text{si} \quad x \leq 0.5, \quad = (0,125, 0, 0.1) \quad \text{sinon.}$$

Introduisons le vecteur des variables $w = (\rho, \rho u, \rho E)$, le système peut s'écrire sous la forme :

$$w_t + (f(w))_x = 0.$$

Nous allons utiliser une stabilisation par viscosité numérique pour chaque équation et résoudre le système par un schéma explicite à un pas (on peut bien sur aussi utiliser un schéma de Runge et Kutta pour une meilleure précision en temps).

$$w^0 = \text{donné}, \quad w_j^{n+1} = w_j^n - \Delta t \left(\frac{f(w_{j+1}^n) - f(w_{j-1}^n)}{2\Delta x} - \nu_j^n \frac{w_{j+1}^n - 2w_j^n + w_{j-1}^n}{\Delta x^2} \right),$$

où la viscosité numérique prend aussi en compte la présence d'ondes acoustiques :

$$\nu_j^n = 0.5\, \Delta x \max(|u_{j-1}^n|, c_{j-1}^n, |u_j^n|, c_j^n, |u_{j+1}^n|, c_{j+1}^n).$$

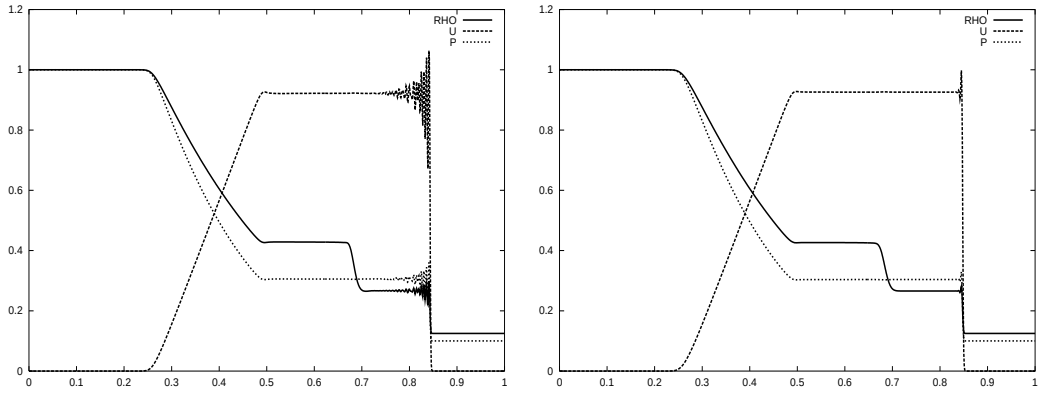


FIGURE 12.12 – Tube à choc de Sod : résolution du système d'Euler en dynamique des gaz avec un schéma d'Euler explicite en temps et $cfl = 0.3$ (à gauche) et avec un schéma RK3 et $cfl = 0.9$ (à droite). Solution pour la densité ρ, la vitesse u et la pression p. On constate que l'augmentation de précision permet de supprimer les oscillations. La qualité de la solution est moindre à travers la discontinuité de contact (discontinuité du milieu). Il existe des schémas numériques permettant une meilleure capture de cette zone.

La vitesse locale du son est donnée par :

$$c_j^n = \sqrt{\gamma \mathcal{R} T_j^n}.$$

Ci-dessus, γ et C_v sont des constantes positives caractéristiques du gaz et $\mathcal{R}$ est la constante des gaz parfaits. De même, la condition de stabilité prend en compte la présence des ondes acoustiques :

$$\Delta t = \min_j \left(\frac{\Delta x}{\max(|u_{j-1}^n| + c_{j-1}^n, |u_j^n| + c_j^n, |u_{j+1}^n| + c_{j+1}^n)} \right).$$

Ce schéma est très élémentaire et provient d'une extension directe de ce que l'on a vu pour les équations scalaires. Il existe des schémas adaptés à la résolution des systèmes, mais ceci dépasse le cadre de cet ouvrage.

12.10 Advection-diffusion-réaction rétrograde

12.10.1 Le modèle de Black et Scholes

Ce modèle prédit le prix d'une option (à l'achat ou à la vente) sur une action ou sur un portefeuille d'actions. Ici l'on considère le cas de l'achat : un industriel, un organisme, une personne s'assure au temps $t = 0$, en payant $u(x, 0)$ (*prime*

d'option), le droit à l'achat d'une action pour un prix K (prix d'exercice) au temps $t = T$ (échéance) quel que soit le prix réel de l'action à $t = T$. On note x (dimension d'espace) $x = x(t)$ le prix de l'action à toute date $0 \leq t \leq T$. Bien sur, on pourra ne pas exercer son option si $x(T) \leq K$. On recherche donc à déterminer $u(0, x)$ pour une condition finale de $u(x, T) = \max(0, x - K)$.

Remarque 12.10.1 *La condition finale peut être plus compliquée. On peut considérer par exemple qu'il n'y aura bénéfice que si la somme K investie sur cette option rapporte plus que la même somme K déposée à la banque (0-risque) sur un compte rémunéré à un taux r.*

Le modèle de Black et Scholes propose une équation parabolique rétrograde en temps pour le comportement de u. Il prend en compte la notion de volatilité σ qui caractérise l'incertitude du comportement : on remplace l'effet des fluctuations par une augmentation de la viscosité du système. Insistons sur le fait que ce modèle prédit le comportement d'une option et non pas d'une action pour laquelle aucune prédiction de comportement n'est possible a priori.

$$\frac{\partial u}{\partial t} + rx\frac{\partial u}{\partial x} + \frac{\sigma^2 x^2}{2}\frac{\partial^2 u}{\partial x^2} = ru, \quad \text{pour} \quad (x, t) \in \mathbf{R}^+ \times \mathbf{R}^+, \tag{12.33}$$

La condition finale est donnée par :

$$u(x, T) = max(0, x - K), K > 0.$$

On limite le domaine de calcul à $x \in (0, L)$ et on introduit les conditions aux limites suivantes :

$$\begin{aligned} &u = 0, \quad \text{en} \quad x = 0, \\ &\frac{\partial u}{\partial t} + rx\frac{\partial u}{\partial x} = ru, \quad \text{en} \quad x = L, \end{aligned}$$

ce qui donne au point L :

$$\frac{\partial u}{\partial t} + rx = ru$$

sachant qu'en $x = L$, $u = x - K$ donc $\dfrac{\partial u}{\partial x} = 1$. Après intégration on obtient :

$$u(L, t) = -K\exp(r(t - T)) + L. \tag{12.34}$$

Les termes retenus étaient faciles à identifier, ce qui n'est pas souvent le cas. Par exemple, si l'on considère uniquement $\frac{\partial u}{\partial t} = ru$ comme modèle réduit, ceci donnerait $u(L, t) = (L - K)\exp(r(t - T))$ comme condition aux limites, qui produit des résultats décevants. Par ailleurs, on peut choisir comme condition aux limites en $x = L$, $\frac{\partial u}{\partial x}(L, t) = 1$, qui est compatible avec la condition finale (i.e. en $t = T$).

De même, on peut utiliser un développement de Taylor pour exprimer le comportement à la frontière :

$$u(L) = u(L-h) + \frac{\partial u}{\partial x}(L-h)h,$$

où h est le pas d'espace. On montre l'effet de diverses conditions aux limites en $x = L$ sur la solution figure (12.13).

Un élargissement du domaine de calcul permet de réduire l'impact de l'erreur commise sur les conditions aux limites, mais ceci implique évidemment un coût de calcul plus important.

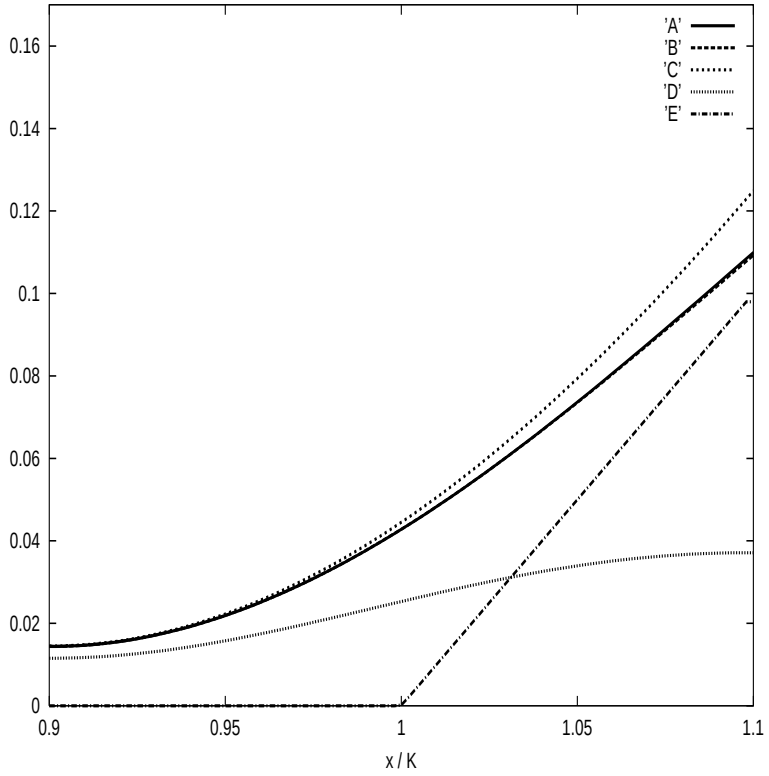


FIGURE 12.13 – Effet du choix de la conditions aux limites en $x = L$ sur la solution du modèle Black et Scholes. On est intéressé par le comportement de l'option au voisinage du prix d'exercice ($x/K = 1$). A : modèle réduit (12.34), B : développement de Taylor d'ordre 1, C : $\frac{\partial u}{\partial x} = 1$, D : Neumann homogène $\frac{\partial u}{\partial x} = 0$, E : condition initiale. Le bon résultat est produit par les conditions A et B.

Remarque 12.10.2 *Ceci est un exemple de conditions aux limites déduites à partir de l'équation. On verra d'autres exemples de conditions aux limites cachées*

et inconnues a priori. Cette démarche est similaire à l'approche "fonctions de paroi" que l'on a déjà vue au chapitre 3.

12.10.2 De Black-Scholes à l'équation de la chaleur

L'équation de Black-Scholes décrivant la valeur C d'une option en fonction du temps t et de la valeur S de l'actif s'écrit :

$$\frac{\partial C}{\partial t} + rS\frac{\partial C}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} = rC$$

où r est le taux de rémunération sans risque et σ la volatilité.
La condition en temps est une condition au temps T (échéance) $C(T, S_T) = max(0, S_T - K)$
Les conditions aux limites sont les suivantes. A priori $S \in [0, +\infty[$.
Soit numériquement $S \in [0, L[$.

$$\text{Pour } S = 0 \text{ on prend } C(t, 0) = 0$$

Pour $S = L$ valeur limite supérieure, on prend

$$C(t, L) = L - Ke^{-r(T-t)}$$

Première étape : on fait le changement de variable $x = ln(S/K)$, on en déduit :

$$\frac{\partial C}{\partial t} + (r - \frac{\sigma^2}{2})\frac{\partial C}{\partial x} + \frac{1}{2}\sigma^2 \frac{\partial^2 C}{\partial x^2} = rC$$

On a donc maintenant une équation linéaire du second ordre à coefficients constants (r et σ sont supposés être des constantes).

Deuxième étape éventuelle : on fait le changement de variable temporelle

$$t = T - \frac{\tau}{\frac{1}{2}\sigma^2}$$

qui ramène l'évolution du temps dans un sens progressif.

$$\frac{\partial C}{\partial \tau} = (k-1)\frac{\partial C}{\partial x} + \frac{\partial^2 C}{\partial x^2} - kC \tag{12.35}$$

avec

$$k = \frac{r}{\frac{1}{2}\sigma^2}$$

Troisième étape : l'équation (12.35) est une équation d'advection-diffusion à coefficients constants. Elle se ramène à une équation de diffusion classique

$$\frac{\partial}{\partial \tau} u(x,\tau) = \frac{\partial^2}{\partial x^2} u(x,\tau)$$

en posant $C(x,\tau) = e^{\alpha x + \beta \tau} u(x,\tau)$, avec α et β bien choisis.

$$\alpha = \frac{1-k}{2} \text{ et } \beta = \alpha^2 + \alpha(k-1) - k$$

12.10.3 Extension aux dimensions supérieures

Considérons le modèle dans le cas où le portefeuille fait intervenir plusieurs actions (par exemple sur le CAC-40). Dans le vocabulaire financier, la dimension s'appelle sous-jacent.

$$\frac{\partial u}{\partial t} + r\vec{x}^T \vec{\nabla} u + \frac{\sigma^2}{2} \nabla.((\vec{x}^T.\vec{x})\vec{\nabla} u) = ru, \quad \text{pour} \quad (\vec{x},t) \in \mathbf{R}^{+40} \times \mathbf{R}^+,$$

La condition finale est donnée par :

$$u(\vec{x},T) = max(0, |\vec{x}| - K), K > 0,$$

où K désigne le prix d'exercice. Ici l'exercice n'est permis qu'en fin de contrat uniquement (i.e. à $t = T$: on appelle ces options européennes ou vanille). Dans ce modèle, la volatilité (i.e. le coefficient de diffusion) permet un retour sur les taux d'intérêt. Ceci permet de modéliser une montée éventuelle des taux si les marchés sont hauts. Ici aussi le domaine de définition est infini (i.e. $\vec{x} \in \mathbf{R}^{+40}$), mais pour effectuer des calculs nous sommes amenés à limiter ce domaine, en introduisant des bornes maximales pour chaque dimension (i.e. $\vec{x} \in \Pi_{i=0}^{40}]0, X_i[$). Les conditions aux limites de ce modèle sont alors :

$$u = 0, \quad \text{sur} \quad \Gamma_1, \quad \text{où} \quad \Gamma_1 = \{\vec{x}, x_i = 0, \forall i\},$$

$$u_{,t} + r\vec{x}^T \vec{\nabla} u = ru, \quad \text{sur} \quad \Gamma_2, \quad \text{où} \quad \Gamma_2 = \{\vec{x}, x_i = X_i, \forall i\}.$$

On voit que la condition équivalente est plus difficile à établir qu'en dimension un. On procède numériquement (c'est à dire que l'on discrétise le modèle réduit). L'autre possibilité est l'utilisation de conditions aux limites équivalentes, basées sur un développement de Taylor, dont on a vu qu'elles produisaient en dimension un, des résultats similaires aux conditions déduites du modèle réduit.

12.10.4 Contraintes d'inégalité

Le modèle de Black et Scholes ci-dessus décrit le prix d'un "call" européen (option à l'achat). En pratique, il existe toujours d'autres contraintes imposées pendant la vie de l'option. Une contrainte possible consiste à demander que le prix de l'option reste toujours inférieur à un certain seuil ψ. Le seuil peut être fonction du prix de l'option et il s'agira alors d'une contrainte d'état.

Le modèle (12.33) devient alors :

$$\min(\frac{\partial u}{\partial t} + rx\frac{\partial u}{\partial x} + \frac{\sigma^2 x^2}{2}\frac{\partial^2 u}{\partial x^2} - ru, u - \psi) = 0, \tag{12.36}$$

Nous verrons au chapitre 17 que les contraintes peuvent être prises en compte de nombreuses façons. La plus simple est la projection. Dans ce cas, à chaque itération de résolution de (12.33), on projette le prix estimé sur le domaine admissible.

Une semi-discrétisation explicite du modèle s'écrit (u^0 donné) :

$$u^{n+1} = \min\left(u^n - \Delta t(rxu_x^n + \frac{\sigma^2 x^2}{2}u_{xx}^n + ru^n), \psi\right).$$

On montre (figure 12.14) l'effet d'une telle contrainte sur la solution du modèle.

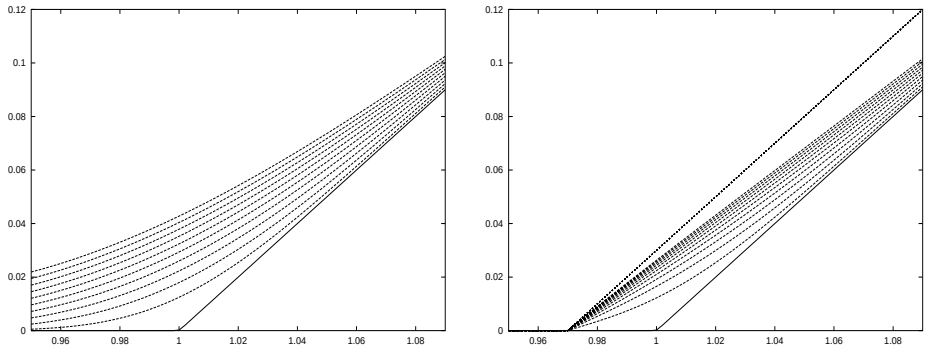


FIGURE 12.14 – Evolution du prix d'une option européenne sans contrainte de seuil (à gauche) et avec contrainte de seuil de déclenchement (à droite) au cours du temps.

Chapitre 13

Couplage de modèles

13.1 Introduction

L'objet de ce chapitre est de présenter quelques configurations faisant intervenir plusieurs modèles physiques discrétisés interagissant entre eux et le couplage entre ces modèles. On s'intéressera aux configurations suivantes :

— couplage hyperbolique-parabolique (structure - fluide),
— couplage elliptique-parabolique-mixte (champs électrique - écoulement d'un fluide - advection-diffusion-réaction d'espèces chimiques),

13.2 Couplage d'EDPs parabolique - hyperbolique

Un premier exemple de couplage concerne l'interaction entre une EDP parabolique représentant le comportement d'un fluide et une EDP hyperbolique modélisant la déformation d'une structure dans le temps, sous l'effet de la solution de l'EDP parabolique. La déformation de la structure modifiant par la suite le domaine géométrique de résolution de l'EDP parabolique. Ce couplage modélise, par exemple, l'interaction entre la structure d'un avion soumis au mouvement de l'air environnant (c'est l'exemple présenté dans les figures 13.1 et 13.2) ou la déformation d'un pont sous l'effet du vent. Un des objectifs principaux de ce type d'étude est l'analyse des possibilités de résonance et l'amélioration de la conception pour éviter ce phénomène et assurer un ensemble stable sur des domaines d'intensités de vent et de fréquences de perturbations les plus larges possibles. D'un point de vue pratique, nous présentons le couplage au travers des EDPs, mais on peut considérer également des équations différentielles, notamment pour l'évolution des modes propres de la structure. Dans ce dernier cas, on doit supposer que les modes propres de la structure couplée restent proches de

ceux de la structure seule. Les figures 13.1 et 13.2 montrent un exemple d'un tel couplage pour le comportement d'une maquette d'avion dans un écoulement instationnaire. On a mis en évidence la possibilité d'une interaction entre la structure et le fluide ; le comportement de la structure devenant instable dans cet exemple.

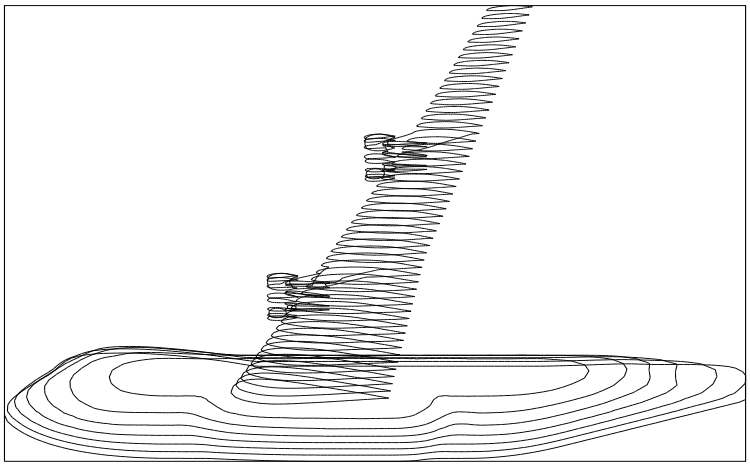


FIGURE 13.1 – Iso-sections développées d'une maquette de 747 servant à la définition de la géométrie utilisée lors de la simulation.

13.2.1 Problème modèle monodimensionnel

Considérons un problème modèle mono-dimensionnel couplant une équation parabolique et une équation hyperbolique. Le domaine de calcul se décompose en $\Omega_1 =]-1, \overline{x}[$ pour l'équation parabolique et $\Omega_2 =]\overline{x}, 1[$ pour l'équation hyperbolique. Le modèle "fluide" (parabolique) est représenté par une équation d'advection-diffusion pour la température T. Les fonctions $c(x) \in I\!R$ et $\mu(x) > 0$ représentent respectivement la vitesse de l'advection et le coefficient de diffusion de la température. Le modèle "structure" (hyperbolique) est représenté par une équation des ondes sur U qui représente un déplacement. $\overline{x}$ est l'interface entre les deux domaines et δ la distribution de Dirac.

$$\frac{\partial T}{\partial t} + \frac{\partial cT}{\partial x} - \frac{\partial}{\partial x}(\mu \frac{\partial T}{\partial x}) = f(x,t), \quad \text{sur } \Omega_1(t) \tag{13.1}$$

$$\frac{\partial^2 U}{\partial t^2} - \frac{\partial^2 U}{\partial x^2} = T(\overline{x}(t))\delta(\overline{x}(t)), \quad \text{sur } \Omega_2(t) \tag{13.2}$$

$$-\frac{\partial^2 X}{\partial x^2} = 0, \quad \text{sur } \Omega_1(t) \tag{13.3}$$

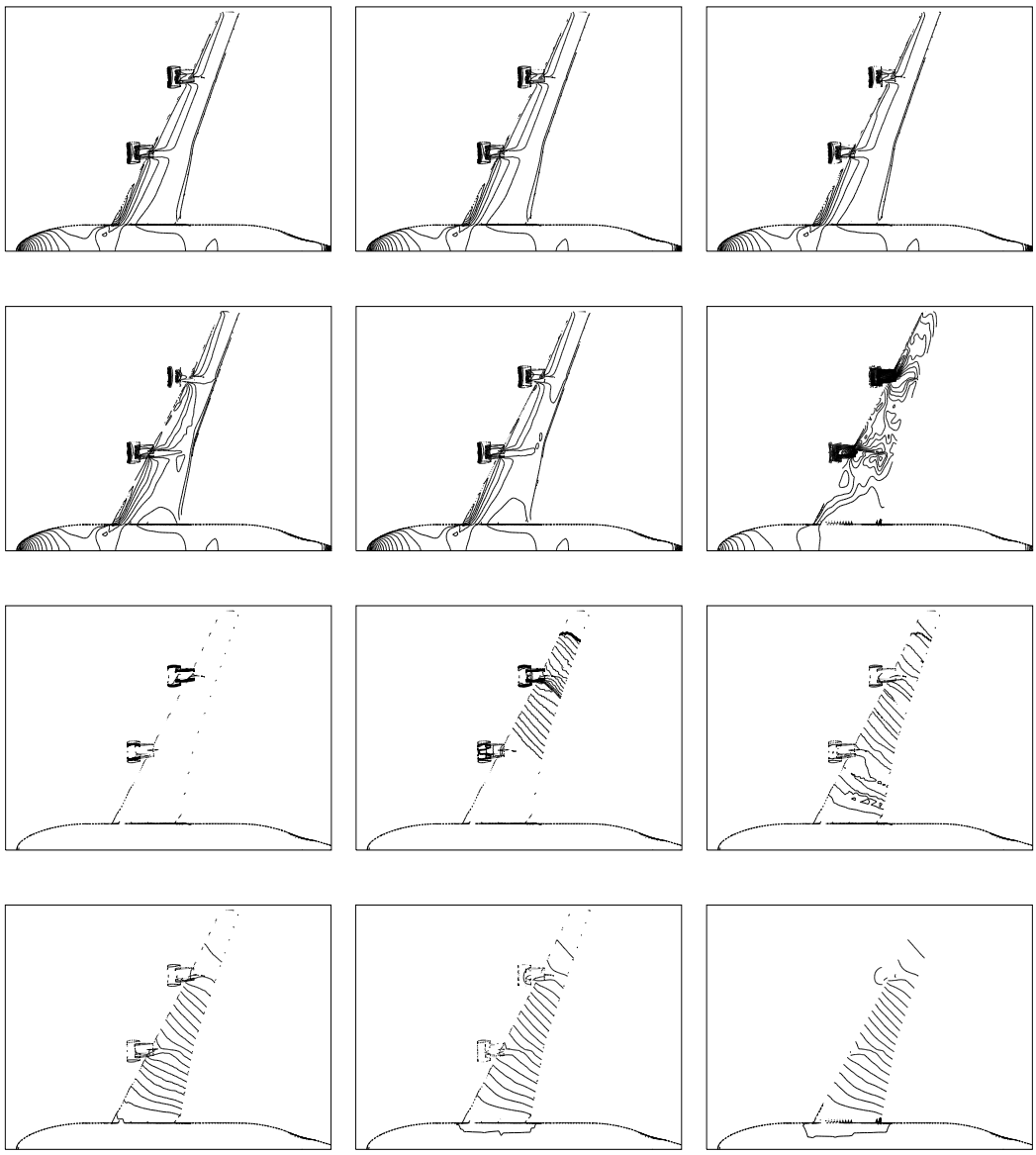


FIGURE 13.2 – Exemple de couplage fluide-structure pour une maquette de 747 (sans la queue). Deux lignes du haut : Iso-pression instantanée de surface (solution du modèle fluide et condition au limite pour le modèle solide élastique). Lignes du bas : Iso-déplacement vertical de la surface (solution du modèle élastique, impliquant un changement de domaine de résolution pour le modèle fluide).

où X représente le déplacement d'un point matériel de $\Omega_1(t)$. L'équation en X permettra de propager dans le domaine Ω et son maillage (dans ce cas simple, selon une loi affine) le déplacement de l'interface.

Les conditions aux limites et initiales sont données par :

$$T(-1,t) = \frac{\partial T}{\partial x}(\overline{x},t) = 0,$$

$$\frac{\partial U}{\partial x}(\overline{x},t) = U(1,t) = 0,$$

$$X(-1,t) = 0, X(\overline{x},t) = U(\overline{x},t).$$

$$T(x,t) = T_0(x) \quad \text{sur } \Omega_1(t),$$

$$U(x,t) = U_0(x), \frac{\partial U}{\partial t}(x,0) = U_1(x) \quad \text{sur } \Omega_2(t).$$

13.2.2 Prise en compte de la déformation du domaine

Afin de prendre en compte des déformations instationnaires de domaine, on utilise la formule de dérivation d'une intégrale sur un domaine dépendant du temps :

$$\frac{d}{dt}\int_{\omega(t)} g(x,t)dx = \int_{\omega(t)} \frac{\partial g(x,t)}{\partial t}dx + \int_{\partial\omega(t)} g\dot{x}.nd\sigma. \tag{13.4}$$

où $\dot{x}$ représente la vitesse de la frontière du domaine. En intégrant (13.1) sur Ω_1 et en utilisant (13.4) et la formule de la divergence, on obtient :

$$\frac{d}{dt}\int_{\omega(t)} Tdx + \int_{\omega(t)} \frac{\partial}{\partial x}(T(c-\dot{x}))dx - \int_{\omega(t)} \frac{\partial}{\partial x}(\mu\frac{\partial T}{\partial x})dx = \int_{\omega(t)} fdx \quad \forall\omega(t) \subset \Omega_1(t). \tag{13.5}$$

Ceci est à la base des techniques ALE(Arbitrary Lagrangian Eulerian).

Discrétisation de l'équation (13.1) :

Ainsi, la déformation du domaine perturbe l'opérateur d'advection. La vitesse de déformation doit être prise en compte lors du décentrage. Sur le maillage $\Omega_{1h}(t^{n+1})$, en dimension un, la vitesse du maillage est définie par :

$$\dot{x}_i = \frac{x_i^{n+1} - x_i^n}{\Delta t}.$$

La discrétisation de (13.5) se fait alors de façon classique par volumes ou éléments finis. En particulier,

$$\frac{d}{dt}\int_{\omega(t)} Tdx \sim \frac{|\omega^{n+1}|T^{n+1} - |\omega^n|T^n}{\Delta t},$$

à l'ordre 1. $|\omega^n|$ représente l'aire du volume de contrôle à l'itération n après condensation de masse (mass lumping).
Discrétisation de l'équation (13.2) :

De même, on discrétise l'équation des ondes sur le domaine $\Omega_{2h}(t)$, donc déformable, en utilisant une formulation variationnelle et des fonctions de base $P1$ (ici en dimension un, voir chapitre 7) :

$$\int_{\Omega_2(t)} \frac{\partial^2 U}{\partial t^2} w_i dx - \int_{\Omega_2(t)} \frac{\partial^2 U}{\partial x^2} w_i = \int_{\Omega_2(t)} T(\overline{x}(t))\delta(\overline{x}(t))w_i dx. \tag{13.6}$$

En exprimant les variables U sur la base w_i, $i = 1, .., N$ avec N le nombre de points de $\Omega_{2h}(t)$, selon

$$U(x,t) = \sum_{i=1}^{N} U_i(t)\, w_i(x)$$

on obtient le système :

$$\begin{aligned} &M_{i,i-1}U''_{i-1}(t) + M_{i,i}U''_i(t) + M_{i,i+1}U''_{i+1}(t) + K_{i,i-1}U_{i-1}(t) + K_{i,i}U_i(t) \\ &\quad + K_{i,i+1}U_{i+1}(t) = T(\overline{x}(t)) \int_{\Omega_{2h}(t)} \delta(\overline{x}(t))w_i dx = T(\overline{x}(t))w_i(\overline{x}(t)), \end{aligned} \tag{13.7}$$

où M et K désigne respectivement les matrices de masse et rigidité :

$$M_{i,j} = \int_{\Omega_{2h}(t)} w_i w_j dx,$$

$$K_{i,j} = \int_{\Omega_{2h}(t)} w'_i(x) w'_j(x) dx.$$

Pour finir, il faut choisir une discrétisation temporelle pour $U''(t)$.

Une discrétisation implicite et précise à l'ordre 2 peut être obtenue avec le schéma de Newmark qui permet d'exprimer U en fonction de U'' :

$$U_i^{n+1} = U_i^n + \Delta t \frac{U'^{n+1}_i + U'^n_i}{2}$$

$$U'^{n+1}_i = U'^n_i + \Delta t \frac{U''^{n+1}_i + U''^n_i}{2}.$$

Dans la formulation ci-dessus, les quantités géométriques sont celles de la configuration $\Omega_{2h}(t^n)$: $M^n_{i,j}, K^n_{i,j}$. La déformation de $\Omega_{2h}(t^n) \to \Omega_{2h}(t^{n+1})$ s'obtient par $x_i^{n+1} = x_i^n + U_i^{n+1}$. Enfin, le couplage par le second membre utilise $T^{n+1}(\overline{x^n})$.

Remarque 13.2.1 *Il est à noter que le nombre de points de discrétisation reste constant dans le temps. En dimension supérieure à un, la connectivité du maillage reste aussi inchangée. Cela peut ne plus être vrai si l'on adapte le maillage entre deux déformations comme nous le verrons au chapitre 18.*

Discrétisation de l'équation (13.3) :

Enfin, pour définir la déformation du domaine $\Omega_{1h}(t)$, nous allons résoudre l'équation (13.3) qui propage la déformation du domaine $\Omega_{2h}(t)$ définie par U à travers $\Omega_{1h}(t)$. La discrétisation se fait par éléments finis $P1$.

En résumé, les étapes de l'algorithme de couplage sont :
— Connaissant $U^n(\overline{x}^n)$.
— Résoudre (13.3) : $\Omega_{1h}(t^n) \to \Omega_{1h}(t^{n+1})$ par $x^{n+1} = x^n + X^{n+1}$.
— Résoudre (13.5) pour obtenir T^{n+1}.
— Utiliser $T^{n+1}(\overline{x}^n)$ pour résoudre (13.7) et évaluer U^{n+1}.

13.2.3 Reformuler avec des systèmes de 1er ordre

Pour mieux comprendre les choix temporels possibles, pour chaque configuration, lors du couplage, on réécrit les équations du modèle sous forme d'un système d'ordre 1 :

$$Z' = f(Z), \quad Z(t=0) = Z_0, \tag{13.8}$$

où $Z = \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} = \begin{bmatrix} U \\ U' \\ T \end{bmatrix}$, et $f(Z) = \begin{bmatrix} Z_2 \\ M^{-1}(T(\overline{x}) - KU) \\ RHS(T,U) \end{bmatrix}$ où RHS désigne les termes de l'équation (13.5). Nous allons utiliser les schémas en temps vus précédemment pour la discrétisation de cet ensemble couplé.

13.2.4 Algorithmes d'ordre un

L'algorithme de couplage le plus simple est basé sur le schéma d'Euler explicite où les deux modèles sont avancés en temps en parallèle et les informations nécessaires communiquées de l'un à l'autre.

$$Z_0 = Z(t=0), \quad Z^{n+1} = Z^n + \Delta t f(Z^n). \tag{13.9}$$

Bien sûr, une condition de stabilité est nécessaire comme pour le cas monomodèle. Pour s'affranchir de cette condition, on peut utiliser une méthode d'Euler implicite (d'ordre 1) :

$$Z^{n+1} = Z^n + \Delta t f(Z^{n+1}).$$

La difficulté ici réside dans un couplage très fort entre les modèles. Ce qui implique la résolution d'un problème de type point fixe. On verra plus loin comment utiliser une prédiction de Z_1^{n+1}.

13.2.5 Améliorer la précision en temps

En utilisant une formule d'intégration de type Crank-Nicolson, on obtient :

$$Z_0 = Z(t=0), \quad Z^{n+1} = Z^n + \Delta t \, \frac{f(Z^n) + f(Z^{n+1})}{2}, \tag{13.10}$$

où l'on peut prendre comme ci-dessus ($Z_1^{n+1} = Z_1^n + \frac{\Delta t}{2}(Z_2^{n+1} + Z_2^n)$) par un schéma de Newmark. Pour réduire le couplage entre les modèles, on peut utiliser la prédiction suivante pour Z_1 :

$$Z_1^{n+1/2} = 2Z_1^n - Z_1^{n-1} = Z_1^n + Z_2^n \Delta t.$$

On voit qu'ici, à chaque itération du couplage, un seul calcul est effectué pour chaque modèle. Les deux modèles peuvent être avancés en parallèle.

13.2.6 Conditions aux limites équivalentes

Si les déformations sont petites, on peut les prendre en compte par une simple modification des conditions aux limites, comme on le précise dans les chapitres 3 et 19. En effet, ceci simplifie grandement la mise en oeuvre du couplage. Pour l'exemple ci-dessus, le couplage de modèles s'écrit alors (on suppose $\overline{x}(0) = 0$) :

$$\frac{\partial T}{\partial t} + \frac{\partial (cT)}{\partial x} - \frac{\partial}{\partial x}(\mu \frac{\partial T}{\partial x}) = f(x,t), \quad \text{sur }]-1,0[\tag{13.11}$$

$$\frac{\partial^2 U}{\partial t^2} - \frac{\partial^2 U}{\partial x^2} = T(0,t)\delta(0), \quad \text{sur }]0,1[\tag{13.12}$$

Les conditions aux limites et initiales sont :

$$T(-1,t) = 0, U(1,t) = 0,$$

$$\overline{x}(t) = U(0,t),$$

$$\frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(\overline{x}(t),t) - \frac{\partial^2 T}{\partial x^2}(0,t)(\overline{x}(t)) = -\frac{\partial^2 T}{\partial x^2}(0,t)(\overline{x}(t)),$$

$$\frac{\partial U}{\partial x}(0,t) = \frac{\partial U}{\partial x}(\overline{x}(t),t) - \frac{\partial^2 U}{\partial x^2}(0,t)(\overline{x}(t)) = -\frac{\partial^2 U}{\partial x^2}(0,t)(\overline{x}(t)),$$

$$T(x,t) = T_0(x) \quad \text{sur }]-1,0[,$$

$$U(x,t) = U_0(x), \frac{\partial U}{\partial t}(x,0) = U_1(x) \quad \text{sur }]0,1[.$$

Ainsi, par des changements de conditions aux limites, le couplage se trouve simplifié dans la mesure où il n'est plus nécessaire de propager la déformation du domaine et de modifier le solveur fluide pour inclure la vitesse du maillage. La déformation conforme de maillages complexes n'est pas une tâche simple, son élimination de la boucle de calcul est très utile.

Au chapitre 17, on présente la relation entre les conditions aux limites équivalentes et l'évaluation des gradients incomplets en optimisation.

13.3 Couplage d'EDPs elliptique - parabolique - mixte

La séparation (une des activités principales de l'homme) se fait souvent en plaçant la quantité contenant les éléments que l'on voudrait séparer dans un champ agissant de façon différente sur les ingrédients présents. On connaît bien la séparation dans un champ gravitationnel ou centrifuge. Une autre façon de séparer les quantités microscopiques, en suspension dans une solution buffer, est de les soumettre à un champ électrique. Avec les nouvelles applications en environnement et santé, cette approche est de plus en plus utilisée car elle permet, au delà de la séparation proprement dite, la création et le contrôle de conditions adéquates pour la réalisation de réactions chimiques. On s'intéresse ainsi aux temps de migrations de quantités diverses ayant des mobilités différentes. Cette technique s'appelle la séparation électro-osmotique.

Le modèle mathématique décrivant le fonctionnement d'un tel dispositif est formé de plusieurs EDP couplées.

13.3.1 Champ électrique

On suppose que les variations de la densité de charges dues aux mouvements des espèces ionisés sont négligeables. Le champ électrique $E(t,x)$ est défini par $E = -\nabla\phi$ (en Volts / m) où $\phi(x,t)$ (Volts), le potentiel électrique, est obtenu par la solution de :

$$\begin{cases} \nabla.E = -\Delta\phi = \frac{F}{\epsilon_r\epsilon_0}\rho_e, \quad \text{in } \Omega \\ \phi(\Gamma_{in}) = \phi_1, \quad \phi(\Gamma_{out}) = \phi_2, \\ \phi = \phi_3 \quad \text{ou} \quad \frac{\partial\phi}{\partial n} = 0 \quad \text{sur les autres frontières.} \end{cases} \tag{13.13}$$

où $\rho_e = \sum_{i=1}^{n} z_i C_i$ est la charge totale (Coulomb/m^3), $z_i \in \mathbf{Z}$ est la valence de l'espèce i de concentration molaire C_i (mol/m^3). F est la constante de Faraday ($F = 96500$), ϵ_r et ϵ_e sont les permitivités relative et absolue (celle du vide). La constante diélectrique $F/(\epsilon_r\epsilon_0) \sim 10^{16}$ est donc très grande et compense la petitesse de la charge totale qui est négligeable. Ceci est une difficulté pour la solution numérique de ce modèle. On verra plus loin comment obtenir une autre équation de diffusion pour ϕ pour les configurations électriquement neutres $\partial_t\rho_e = \rho_e = 0$.

Une possibilité pour la résolution de (13.13) consiste en un calcul séparé, utilisant la linéarité de l'EDP, des deux champs électriques présents : celui dû aux différences de potentiel externe, et le champ créé par la présence des ions.

On a alors, $\phi = \phi_e + \phi_i$, avec

$$\begin{cases} -\Delta\phi_e = 0, \quad \text{in } \Omega \\ \phi_e(\Gamma_{in}) = \phi_1, \quad \phi_e(\Gamma_{out}) = \phi_2, \\ \phi_e = \phi_3 \quad \text{ou} \quad \frac{\partial\phi_e}{\partial n} = 0 \quad \text{sur les autres frontières.} \end{cases} \tag{13.14}$$

et

$$\begin{cases} -\Delta\phi_i = \frac{F}{\epsilon_r\epsilon_0}\rho_e, \quad \text{in } \Omega \\ \phi_i(\Gamma_{in}) = 0, \quad \phi_i(\Gamma_{out}) = 0, \\ \phi_i = 0 \quad \text{ou} \quad \frac{\partial\phi_i}{\partial n} = 0 \quad \text{sur les autres frontières.} \end{cases} \tag{13.15}$$

On peut aussi considérer une densité de charge nulle $\rho_e = 0$. Ce qui est une bonne approximation en dehors des régions où les gradients de concentration sont élevés. Ceci réduit le champ électrique au champ externe et a l'avantage de découpler le potentiel électrique des autres variables du problème (vitesse-pression-concentration). Cette stratégie est intéressante pour certaines applications, notamment lors de l'obtention des sensitivités en optimisation par exemple. Il y a cependant des situations où l'on souhaite une évaluation plus fine des variables et où, malgré cette hypothèse, on peut calculer le champ généré par les ions en utilisant l'hypothèse d'électroneutralité.

13.3.2 Bilan des charges

Il existe deux approches pour définir le potentiel électrique. On peut

— soit résoudre l'équation de Poisson-Boltzmann (13.13)

— soit en utilisant la neutralité électrique, obtenir une nouvelle équation de diffusion pour ϕ.

dans ce cas, la charge électrique totale ρ_e est décrite par l'équation de bilan :

$$\frac{\partial\rho_e}{\partial t} - \nabla.I = 0,$$

$I(= F\sum_{i=1}^{n} z_i j_i)$ désigne le courant de densité de charge, j_i est le flux molaire (s mol / m^2) :

$$j_i = \nu_i z_i F C_i \nabla\phi + D_i \nabla C_i - C_i U, \tag{13.16}$$

où ν_i est la mobilité de l'espèce i (mol s / Kg) et D_i son coefficient de diffusion (m^2/s).

Ainsi, la nouvelle équation de diffusion, au contraire de (13.13), intégre les variations de charges dues aux espèces :

$$F\nabla.((\sum_{i=1}^{n} \nu_i z_i C_i)\nabla\phi) = \nabla.(\sum_{i=1}^{n} D_i z_i \nabla C_i). \tag{13.17}$$

Les conditions aux limites étant les mêmes que pour (13.13). On voit ainsi que l'hypothèse de neutralité électrique est raisonnable dans les régions où $\nabla C_i = 0$. La résolution numérique de cette équation est cependant difficile car, pour une espèce uniformément présente dans une partie du domaine, les régions où $\nabla C_i \neq 0$ sont de mesures nulles et concernent uniquement les frontières de cette région. Le maillage doit donc être adapté et le rester pendant le calcul (i.e. mouvement de la région).

13.3.3 Vitesse de l'écoulement

Le modèle le plus simple pour la vitesse U (m/s) de l'écoulement est donnée par la relation suivante qui relie la vitesse au champ électrique :

$$U = -\mu_{ek} E, \tag{13.18}$$

où μ_{ek} est la mobilité électro-cinétique de la solution qui est fonction des concentrations locales. Des modèles plus complexes peuvent être utilisés. Par exemple, la vitesse et la pression peuvent être obtenues par résolution du système de Stokes.

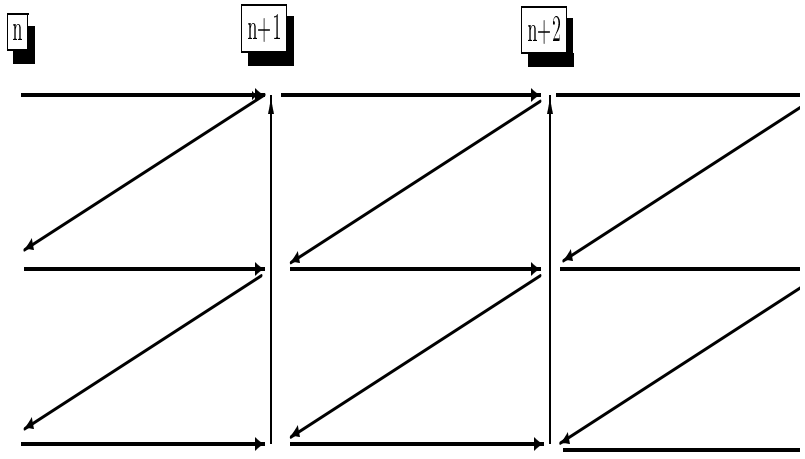


FIGURE 13.3 – Le plus simple schéma de couplage entre trois modèles avec communication d'informations entre modèles à chaque itération.

13.3.4 Advection des espèces

Le modèle est clos par la description du mouvement des espèces et de leurs interactions mutuelles. Les variations temporelles de concentrations C_i sont

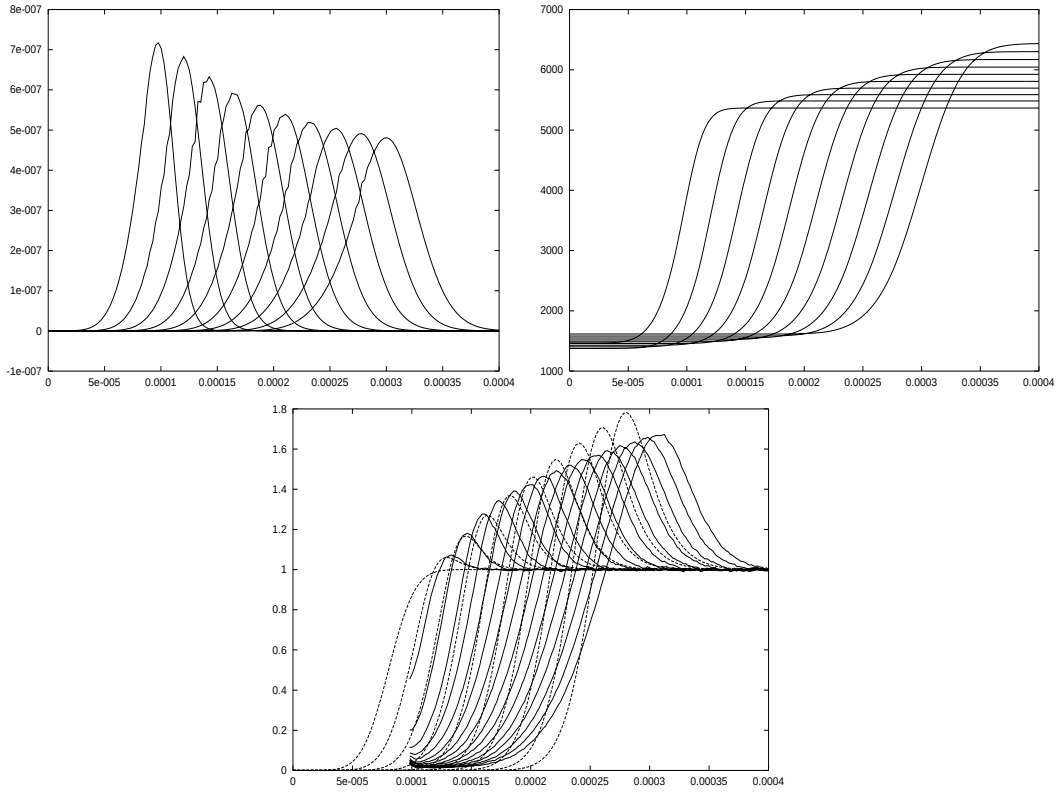


FIGURE 13.4 – Augmentation de la concentration locale (stacking) d'une espèce par mobilité dans un champ électrique. On présente respectivement, la distribution de la charge totale ρ_e, le champ électrique E et la distribution de l'espèce que l'on tente d'identifier après une augmentation de sa concentration molaire avec une comparaison avec l'expérience Calculs réalisés par G. Alexis-Alexandre à Montpellier).

données par le bilan du flux molaire (13.16) :

$$\frac{\partial C_i}{\partial t} - \nabla . j_i = R_i, \tag{13.19}$$

dont la solution requiert la connaissance de U et ϕ. R_i modélise les réactions chimiques entre les espèces. Les conditions aux limites et initiales sont :

$$C_i = \text{donné si} \quad U.n \leq 0 \quad \text{et} \quad \frac{\partial C_i}{\partial n} = 0 \quad \text{sinon,} \quad C_i(x, t = 0) = C_i^0.$$

13.3.5 Algorithme de couplage

On peut utiliser l'un des algorithmes présentés pour le couplage ci-dessus, ou bien l'algorithme implicite par point fixe explicite (2.42) présenté au chapitre 2 qui a l'avantage de ne pas demander de modification des briques de base du programme. En effet, à chaque itération du point fixe pour le modèle couplé, connaissant les concentrations instantanées et les conditions aux limites, les étapes suivantes sont effectuées :

— 1. Calcul de ϕ solution de (13.13) ou (13.17).
— 2. Calcul de la vitesse par (13.18).
— 3. Advection et diffusion des espèces par (13.19).

Le point fixe aura convergé quand les concentrations calculées par les itérations de point fixe tendent vers des distributions limites. On procèdera alors à l'avance en temps.

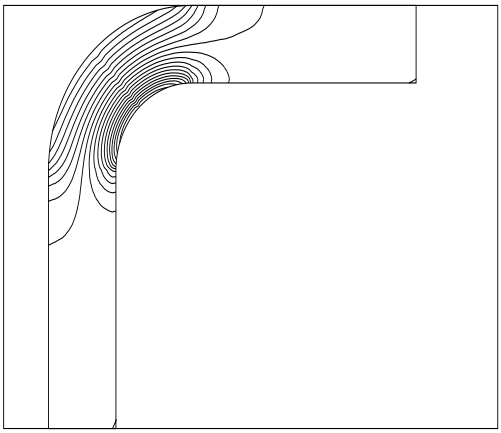


FIGURE 13.5 – Iso-contour de la norme du champ électrique engendré au voisinage d'un coude à 90 degrés.

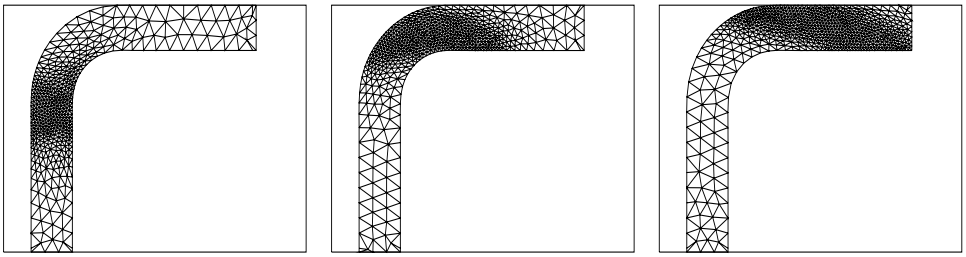


FIGURE 13.6 – Maillage adaptatif pour la capture de l'advection du front de saut de concentration ci-dessous.

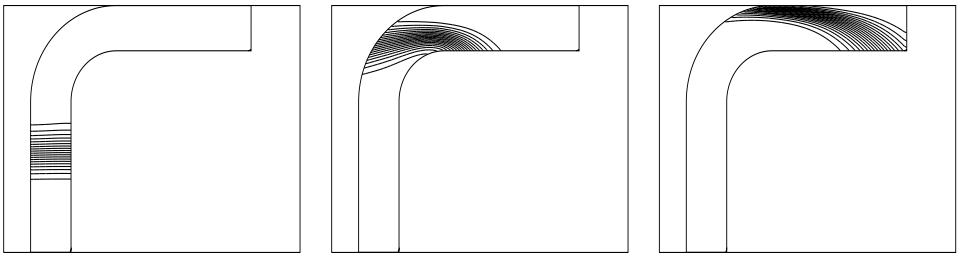


FIGURE 13.7 – Calcul adaptatif d'un front de concentration advecté par le champ de vitesse.

Chapitre 14

Optimisation quadratique et moindres carrés dans $\mathbb{R}^n$

Nous supposons connues les notions de base de l'algèbre linéaire (voir annexe) et nous nous limitons pour la suite au cas d'espaces vectoriels sur le corps des réels.

14.1 Espaces vectoriels

Définition 14.1.1 (Espace vectoriel) *On appelle espace vectoriel réel E, un ensemble muni de deux lois de composition : une loi de composition interne, l'addition et une loi de composition externe, la multiplication par un scalaire réel.*

- *l'addition donne à E une structure de groupe multiplicatif.*
- *la multiplication associe à tout réel λ et tout vecteur (élément de E) x, un vecteur noté λx et vérifie les propriétés suivantes :*

$$\lambda(x+y) = \lambda x + \lambda y \quad (\lambda+\mu)x = \lambda x + \mu x \tag{14.1}$$

$$\lambda(\mu)x = (\lambda\mu)x \quad et \quad 1x = x \tag{14.2}$$

pour tous λ et μ réels et tous x, y de E.

14.1.1 Exemple fondamental : $\mathbb{R}^n$

Soit $\mathbb{R}$ le corps des réels, un élément x de $\mathbb{R}^n$ est une collection de n réels. On note

$$x = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix} \quad \text{vecteur colonne} \tag{14.3}$$

$$x^T \text{ (transposé de } x) = [\, x_1, x_2, ., ., x_n \,] \quad \text{vecteur ligne} \tag{14.4}$$

Les propriétés de l'addition vectorielle et de la multiplication par un scalaire confèrent à $\mathbb{R}^n$ une structure d'espace vectoriel réel.

14.2 Formes linéaires et bilinéaires

14.2.1 Formes linéaires

Définition 14.2.1 *Une forme linéaire l sur un espace vectoriel réel E est une application linéaire de E dans $\mathbb{R}$.*

Si E est de dimension n elle sera donc représentée par un vecteur ligne L de n composantes. Chacune des composantes est l' image par la forme linéaire l d'un vecteur de base de E. L'image d'un vecteur x quelconque de E s'obtient alors par produit scalaire de L par le vecteur X des composantes de x

$$l(x) = (L, X) \tag{14.5}$$

Retenons qu'en dimension finie, toute forme linéaire se représente par un produit scalaire.

14.2.2 Formes bilinéaires

Définition 14.2.2 *Une forme bilinéaire a sur un espace vectoriel réel E est une application de $E \times E$ dans $\mathbb{R}$, linéaire par rapport à chacun de ses deux arguments.*

Soit a la forme bilinéaire, on a donc :

$$a(\lambda_1 u_1 + \lambda_2 u_2, v) = \lambda_1 a(u_1, v) + \lambda_2 a(u_2, v) \tag{14.6}$$

$$a(u, \mu_1 v_1 + \mu_2 v_2) = \mu_1 a(u, v_1) + \mu_2 a(u, v_2) \tag{14.7}$$

Représentation matricielle

Toute forme bilinéaire sur un espace E de dimension finie n se représente, dans une base $\{e_i\}$ par une matrice carrée d'ordre n. Les coefficients A_{ij} de la matrice A représentant l'application a sont donnés par

$$A_{ji} = a(e_i, e_j) \tag{14.8}$$

On a

$$a(u,v) = (AU, V) = (U, A^T V) \tag{14.9}$$

si $(.,.)$ représente le produit scalaire usuel de $\mathbb{R}^n$ et A^T est la matrice transposée de A définie par

$$A^T_{ij} = A_{ji} \quad \forall i,j = 1...N$$

.

14.2.3 Formes bilinéaires symétriques définies positives

Définition 14.2.3 *Une forme bilinéaire a sur un espace vectoriel réel E est symétrique si :*

$$\forall u, v \in E \quad a(u,v) = a(v,u) \tag{14.10}$$

Définition 14.2.4 *Une forme bilinéaire a sur un espace vectoriel réel E est définie positive si :*

$$\forall u \in E \quad a(u,u) \geq 0 \tag{14.11}$$

et

$$a(u,u) = 0 \Longleftrightarrow u = 0 \tag{14.12}$$

Les formes bilinéaires symétriques sont représentées par des matrices symétriques $A_{ij} = A_{ji}$. Les formes bilinéaires symétriques définies positives sont représentées par des matrices symétriques définies positives, qui vérifient donc :

$$(AU, U) \geq 0 \quad \forall U \in \mathbb{R}^N \quad \text{et} \quad (AU, U) = 0 \Rightarrow U = 0 \tag{14.13}$$

.

Théorème 14.2.1 (Résultat important) *Les matrices symétriques réelles ont des valeurs propres réelles, sont diagonalisables et admettent une base de vecteurs propres orthonormés. Les matrices symétriques définies positives ont des valeurs propres strictement positives et donc sont inversibles.*

14.3 Équivalence entre résolution d'un système linéaire et minimisation quadratique

Théorème 14.3.1 *(fondamental pour la suite) Si A est une matrice symétrique définie positive, il y a équivalence entre les trois problèmes suivants :*

$$(1) \qquad \begin{cases} \text{Trouver } X \in \mathbb{R}^N \quad \text{tel que} \\ \\ A\,X = B \end{cases} \tag{14.14}$$

$$(2) \qquad \begin{cases} \textit{Trouver } X \in I\!R^N \quad \textit{tel que} \\ \\ (A\,X, Y) = (B, Y) \quad \forall\, Y \in I\!R^N \end{cases} \tag{14.15}$$

$$(3) \qquad \begin{cases} \textit{Trouver } X \in I\!R^N \quad \textit{tel que} \\ \\ J(X) = \dfrac{1}{2}(A\,X, X) - (B, X) \quad \textit{soit minimal} \end{cases} \tag{14.16}$$

Démonstration
$1 \Longrightarrow 2$ est évident.
$2 \Longrightarrow 1$ en prenant pour Y les vecteurs de base e_i de $I\!R^N$.
$2 \Longrightarrow 3$: On calcule $J(X + \lambda Y)$ pour tout λ réel et tout $Y \in I\!R^N$, on obtient .

$$J(X + \lambda Y) = J(X) + \lambda[(AX, Y) - (B, Y)] + \frac{\lambda^2}{2}(AY, Y)$$

en utilisant la symétrie de la matrice A.

On en déduit, si $(AX, Y) - (B, Y) = 0$ que $J(X + \lambda Y) = J(X) + \frac{\lambda^2}{2}(AY, Y)$
d'où en utilisant le fait que A est définie positive :

$$J(X + \lambda Y) > J(X)$$

si λ et Y sont non nuls. Donc on a montré que si X vérifie (2), X minimise J.
Inversement $3 \Longrightarrow 2$, car si X minimise J, on a

$$\lambda[(AX, Y) - (B, Y)] + \frac{\lambda^2}{2}(AY, Y) \geq 0 \quad \forall \lambda, \forall y$$

Le trinôme en λ ci-dessus doit être toujours positif. Ceci entraîne que son discriminant soit toujours négatif ou nul. Or ce discriminant est

$$\Delta = [(AX, Y) - (B, Y)]^2$$

Ceci implique (2). On a donc démontré les équivalences $1 \Longleftrightarrow 2$ et $2 \Longleftrightarrow 3$ et donc l'équivalence des 3 problèmes.

14.4 Application aux moindres carrés

Considérons le problème général d'un système linéaire sur-déterminé, c'est à dire dans lequel il y a plus d'équations que d'inconnues. C'est en particulier le cas dans le calcul de la droite des moindres carrés ou plus généralement de polynômes

d'approximation au sens des moindres carrés. On ne peut pas obtenir exactement l'égalité

$$AX = B$$

car A est une matrice rectangulaire de N lignes et m colonnes avec $N >> m$. On essaie alors de minimiser l'écart entre les vecteurs AX et B de

$$Minimiser\, J(X) = \|AX - B\|^2 = (AX - B, AX - B)$$

en minimisant la norme euclidienne de leur différence, ou ce qui revient au même le carré de cette norme.

$$Minimiser\, J(X) = \|AX - B\|^2 = (AX - B, AX - B)$$

On utilise les propriétés classiques du produit scalaire $(AU, V) = (U, A^T V)$ pour obtenir :

$$J(X) = (A^T AX, X) - 2(A^T B, X) + (B, B)$$

la matrice $A^T A$ est une matrice carrée $m \times m$ symétrique définie positive dès lors que la matrice rectangulaire A est de rang m. Le théorème (14.3.1) nous donne l'équivalence de ce problème de moindres carrés avec la résolution du système linéaire

$$A^T AX = A^T B$$

On retrouve ainsi le système carré de m équations à m inconnues, dit "système des équations normales".

14.4.1 Droite des moindres carrés

Par exemple, dans le cas de la droite des moindres carrés, il s'agit de trouver la fonction affine $y = a_0 + a_1 x$ qui représente " au mieux " une collection de N valeurs y_i associées aux N abscisses x_i. Au sens des moindres carrés, ceci revient à minimiser la somme

$$\sum_{i=1}^{N} [y_i - a_0 - a_1 x_i]^2$$

donc à minimiser la norme euclidienne de la différence

$$\|Y - Xa\|$$

où l'on a noté Y le vecteur des N valeurs y_i, X la matrice rectangulaire à N lignes et 2 colonnes

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ : & : \\ : & : \\ : & : \\ 1 & x_N \end{pmatrix}$$

et a le vecteur $\begin{pmatrix} a_0 \\ a_1 \end{pmatrix}$ En appliquant les résultats précédents, on obtient la solution en résolvant le système 2×2

$$X^T X a = X^T Y$$

.

14.4.2 Interprétation en terme de projection sur un sous-espace

Reprenons le problème de l'approximation au sens des moindres carrés

$$Minimiser\, J(X) = \|AX - B\|^2$$

On peut interpréter ce problème comme celui de la recherche du vecteur de forme AX le plus proche au sens de la norme euclidienne d'un vecteur B donné dans $\mathbb{R}^N$. Les vecteurs de la forme AX sont des combinaisons linéaires des m vecteurs colonnes de la matrice A. Ces vecteurs colonnes sont indépendants car A est supposée de rang m. Ils engendrent donc un sous-espace F de $\mathbb{R}^N$ de dimension m. Et le problème s'interprète comme la recherche du vecteur du sous-espace F le plus proche (au sens de la norme euclidienne) du vecteur B. On obtient donc X en écrivant que AX est la projection orthogonale de B dans F (voir annexe) donc que

$$(AX - B, V) = 0 \quad \forall V \in F$$

ou ce qui est équivalent, puisque les colonnes de A engendrent F,

$$(AX - B, A_j) = 0 \quad \forall j = 1, N$$

où A_j est le j^{ieme} vecteur colonne de A.

On retrouve ainsi le résultat

$$A^T A X = A^T B$$

Chapitre 15

Calcul différentiel

15.1 Calcul différentiel dans $\mathbb{R}^N$

On rappelle la définition du gradient. Le gradient est un opérateur différentiel linéaire qui, appliqué à une fonction F de N variables $x_1, x_2, ..., x_N$ à valeur scalaire, donne un vecteur : le gradient de F défini par

$$grad(F) = \begin{pmatrix} \frac{\partial F}{\partial x_1} \\ \frac{\partial F}{\partial x_2} \\ \vdots \\ \frac{\partial F}{\partial x_i} \\ \vdots \\ \frac{\partial F}{\partial x_N} \end{pmatrix}$$

On note aussi le gradient ∇F.

Définition 15.1.1 *Une fonction de $\mathbb{R}^N$ dans $\mathbb{R}$ est différentiable (au sens de Fréchet) en $x \in \mathbb{R}^N$ si $\forall h \in \mathbb{R}^N$ on a*

$$F(x+h) = F(x) + (\nabla F(x), h) + \epsilon(h)\|h\|$$

où $(.,.)$ dénote le produit scalaire canonique de $\mathbb{R}^N$ et $\epsilon(h)$ tend vers 0 avec h.

Cela signifie que F peut être approchée au voisinage du point x par une fonction affine.

Définition 15.1.2 *Une fonction de $\mathbb{R}^N$ dans $\mathbb{R}$ est différentiable au sens de Gâteaux au point x si l'on a*

$$\forall h \in \mathbb{R}^N, \quad \frac{d}{dt}F(x+th) = (\nabla F(x), h)$$

L'application linéaire $h \to (\nabla F(x), h)$ que l'on note $DF(x)$ est la différentielle de F au point x et on a, dans le cas de $F \in C^2$, à la fois une différentielle de Fréchet

$$F(x+h) = F(x) + DF(x).h + \epsilon(h)\|h\|$$

et de Gateaux

$$DF(x).h = (\nabla F(x), h) = \lim_{t \to 0} \frac{F(x+th) - F(x)}{t}$$

Remarque 15.1.1 (Un contre-exemple) *Attention, si une fonction différentiable au sens de Fréchet est évidemment différentiable au sens de Gâteaux, l'inverse n'est pas vrai. Dans le cas de la différentielle au sens de Gâteaux, l'accroissement th tend vers 0 avec t donc de façon proportionnelle pour toutes les directions. Considérons la fonction*

$$\begin{cases} f(x,y) = \frac{x^3 y}{x^4+y^2} \text{ pour } x \text{ et } y \neq 0 \\ f(0,0) = 0 \end{cases}$$

On peut montrer en exercice que f est Gâteaux-différentiable en $(0,0)$, mais non Fréchet-différentiable (prendre un accroissement $h = \begin{pmatrix} \epsilon \\ \epsilon^2 \end{pmatrix}$)

15.1.1 Dérivée directionnelle

Le gradient permet de calculer la dérivée d'une fonction sur une courbe quelconque $x(\lambda)$, par dérivation d'une fonction composée, selon

$$\frac{d}{d\lambda} F(x(\lambda)|_{\lambda=0} = \sum_i \frac{\partial F(x(0))}{\partial x_i} = (\nabla F(x(0)), x'(0))$$

En particulier, la **dérivée directionnelle** de la fonction F le long de la droite d' équation paramétrique $x(\lambda) = x + \lambda h$ est

$$\frac{d}{d\lambda} F(x+\lambda h)|_{\lambda=0} = (\nabla F(x(0)), h)$$

15.1.2 Matrice Jacobienne

Si F est une fonction différentiable de $I\!R^N$ dans $I\!R^N$, on appelle matrice jacobienne de F, la matrice symétrique de coefficients

$$J_{i,j} = \frac{\partial F_i(x)}{\partial x_j}$$

.

15.1.3 Matrice Hessienne

La matrice hessienne ou le hessien d'une fonction F deux fois différentiable de $\mathbb{R}^N$ dans $\mathbb{R}$ est par définition la matrice jacobienne de son gradient, c'est à dire la matrice symétrique $HF(x)$ de coefficients :

$$HF(x)_{i,j} = \frac{\partial^2 F(x)}{\partial x_i \partial x_j}$$

.

15.1.4 Formule de Taylor

Une fonction F de $\mathbb{R}^N$ dans $\mathbb{R}$ de classe C^2, c'est à dire deux fois différentiable admet le développement de Taylor à l'ordre 2 suivant

$$F(x+h) = F(x) + (\nabla F(x), h) + \frac{1}{2}(HF(x)h, h) + \epsilon(h)\|h\|^2$$

où $\epsilon(h)$ tend vers 0 avec h.

$(.,.)$ est le produit scalaire canonique de $\mathbb{R}^N$, $\nabla F(x)$ est le gradient de F au point $x \in \mathbb{R}^N$, $HF(x)$ est la matrice hessienne ou Hessien de F au point x.

$$HF(x)_{ij} = \frac{\partial^2 F}{\partial x_i \partial x_j}$$

Remarque 15.1.2 *Dans le cas de la fonction de $\mathbb{R}^N$ dans $\mathbb{R}$, définie par*

$$J(x) = \frac{1}{2}(AX, X) - (B, X)$$

dans le théorème (14.3.1), on vérifie simplement que l'on a $\nabla J(x) = AX - B$ et $HJ(x) = A$. On observe, ce qui sera justifié plus bas, que la minimisation de J équivaut d'après le théorème (14.3.1) à la recherche de zéros du gradient de J.

15.2 Généralisation aux espaces de Hilbert

On se place, dans ce qui suit, dans des espaces vectoriels normés. Par la suite on considérera surtout des Hilberts, espaces vectoriels normés complets dont la norme dérive d'un produit scalaire. On a les définitions suivantes, généralisations des définitions dans $\mathbb{R}^N$

Définition 15.2.1 *Une fonction d'un espace vectoriel normé E dans un espace vectoriel normé F est différentiable (au sens de Fréchet) en $x \in E$, s'il existe une application linéaire continue de E dans F, notée $DF(x)$ telle que $\forall h \in E$ on ait*

$$F(x+h) = F(x) + DF(x).h + \epsilon(h)\|h\|$$

où $DF(x).h$ dénote l'image dans F de l'application $DF(x)$ appliquée à h, et $\epsilon(h)$ tend vers 0 avec $\|h\|$.

Cela signifie que F peut être approchée au voisinage du point x par une fonction affine.

Définition 15.2.2 *Une fonction de E dans $\mathbb{R}$ est différentiable au sens de Gâteaux au point $x \in E$, s'il existe une application linéaire continue de E dans F, notée $DF(x)$ telle que $\forall h \in E$ on ait*

$$\forall h \in E, \quad \frac{d}{dt}F(x+th)_{|t=0} = DF(x).h$$

15.3 Formulaire

— $D(f+g)(x) = Df(x) + Dg(x)$, évident d'après la définition
— Si $l : x \to l(x)$ est une application linéaire continue, on a $Dl(x).h = l(h) \quad \forall x$, en effet $l(x+h) = l(x) + l(h)$

— Si $a : x, y \to a(x,y)$ est une application bilinéaire de $V \times V$ dans W ($\mathbb{R}$ par exemple), on a $Da(f(x), g(x)).h = a(Df(x).h, g(x)) + a(f(x), Dg(x).h)$.

— Cas du produit

$$D(f(x)g(x)).h = g(x)Df(x).h + f(x)Dg(x).h$$

— Inverse algébrique

$$D(\frac{1}{f(x)}.h = -\frac{Df(x).h}{f(x)^2}$$

— Composée

$$D(g \circ f)(x) = Dg(f(x)) \circ Df(x)$$

— Fonction inverse

$$Df^{-1}(y) = (Df(f^{-1}(y)))^{-1}$$

15.4 Applications

— Soit a est une forme bilinéaire symétrique continue et l une forme linéaire continue sur un Hilbert H, et soit F la fonctionnelle définie de H dans $\mathbb{R}$ par

$$F(v) = \frac{1}{2}a(v,v) - l(v)$$

On a

$$DF(v).w = a(v,w) - l(w) \quad \forall w \in H$$

— Par exemple, pour

$$F(v) = \frac{1}{2}[\int_0^L v'^2(x)dx + \int_0^L v^2(x)dx] - \int_0^L f(x)v(x)dx$$

on obtient

$$DF(v).w = \int_0^L v'(x)w'(x)dx + \int_0^L v(x)w(x)dx - \int_0^L f(x)w(x)dx$$

— Différentielle de la longueur d'un arc. Pour

$$F(v) = \int_0^L \sqrt{1 + v'^2(x)}dx$$

on obtient

$$DF(v).w = \int_0^L \frac{v'(x)w'(x)}{\sqrt{1 + v'^2(x)}}dx$$

Chapitre 16

Convexité et optimisation

La notion de convexité permet d'obtenir des résultats d'existence et d'unicité de problèmes d'optimisation à la fois dans le cas où ces problèmes conduisent à la résolution de systèmes linéaires à matrice symétriques et dans le cas de certains problèmes non-linéaires.

16.1 Ensembles convexes

On se place dans un espace vectoriel E sur le corps des réels, le cas le plus courant étant l'espace $\mathbb{R}^n$.

Définition 16.1.1 *Un sous-ensemble C d'un espace vectoriel E est convexe si pour tout couple x, y d'éléments de C, et $\forall \lambda \in [0,1]$,*

$$\lambda x + (1-\lambda)y \in C$$

Exemples : L'espace E tout entier est convexe. Un demi-espace, le cône positif de $\mathbb{R}^n$ sont convexes.

Définition 16.1.2 *On appelle combinaison convexe de n éléments x_i de E, tout élément*

$$x = \sum_i \lambda_i x_i$$

avec les poids $\lambda_i \geq 0$ et $\sum_i \lambda_i = 1$.

Définition 16.1.3 *L'enveloppe convexe d'un ensemble d'éléments x_i de E est l'ensemble des combinaisons convexes des x_i. C'est aussi le plus petit convexe contenant les x_i.*

Définition 16.1.4 *Un point d'un convexe est extremal s'il n'est pas le milieu de 2 points du convexe. Par exemple les sommets d'un polyèdre convexe.*

16.1.1 Projection sur un convexe

On se place désormais dans des espaces de Hilbert, c'est à dire des espaces vectoriels munis d'un produit scalaire, normés par la norme déduite de ce produit scalaire et complet pour cette norme. En dimension finie, il s'agit des espaces euclidiens dont le prototype et l'exemple le plus utile est l'espace $\mathbb{R}^n$ muni du produit scalaire canonique $(x, y) = \sum_i x_i y_i$ et de la norme euclidienne associée.

On admettra le théorème suivant :

Théorème 16.1.1 *Soit K un sous-ensemble convexe fermé non vide d'un espace euclidien ou de Hilbert H, pour tout $u \in H$ il existe un élément unique $\bar{u} \in K$ tel que*

$$\|u - \bar{u}\| = \min_{v \in K} \|u - v\| \tag{16.1}$$

On note ce projeté $\bar{u}$ de u dans K : $\Pi_K u$.

Propriété caractéristique

Le projeté $\Pi_K u$ de u sur K est caractérisé par la propriété :

$$(u - \Pi_K u, w - \Pi_K u) \leq 0 \quad \forall w \in K \tag{16.2}$$

Démonstration.

Par définition $(u - \Pi_K u, u - \Pi_K u) \leq (u - v, u - v) \quad \forall v \in K$. Soit w quelconque dans K on considère la combinaison convexe $tw + (1-t)\Pi_K u$, avec $0 \leq t \leq 1$ qui appartient également à K. Donc, par définition du projeté :

$$(u - \Pi_K u, u - \Pi_K u) \leq (u - tw - (1-t)\Pi_K u, u - tw - (1-t)\Pi_K u) \quad \forall w \in K$$

On développe et on obtient

$$2\,(u - \Pi_K u, w - \Pi_K u) \leq t\,(\Pi_K u - w, \Pi_K u - w)$$

et le résultat en faisant tendre t vers zéro dans l'inégalité.

Autres propriétés.

1) Π_K est idempotente, i.e. $\Pi_K^2 = \Pi_K$
2) Π_K est monotone, i.e. $(\Pi_K u - \Pi_K v, u - v) \geq 0 \quad \forall u, v \in H$.
3) Π_K est faiblement contractante, i.e. $\|\Pi_K u - \Pi_K v\| \leq \|u - v\| \quad \forall u, v \in H$.

Démonstration.

1) $\Pi_K u \in K$, donc évidemment $\Pi_K(\Pi_K u) = \Pi_K u$

2 et 3) On utilise la propriété caractéristique :

$$(u - \Pi_K u, w - \Pi_K u) \leq 0 \quad \forall w \in K$$

$$(v - \Pi_K v, w - \Pi_K v) \leq 0 \quad \forall w \in K$$

On choisit $w = \Pi_K v$ dans la première inégalité et $w = \Pi_K u$ dans la seconde, et on obtient par addition

$$\|\Pi_K u - \Pi_K v\|^2 \leq (u - v, \Pi_K u - \Pi_K v)$$

ceci entraîne 2) et on obtient 3) par Schwarz.

$$(u - v, \Pi_K u - \Pi_K v) \leq \|u - v\| \, \|\Pi_K u - \Pi_K v\|$$

16.1.2 Projection sur un sous-espace vectoriel fermé

Un sous-espace est un sous ensemble convexe, donc les résultats précédents s'appliquent. On a de plus le théorème suivant :

Théorème 16.1.2 *Soit F un sous-espace fermé non vide d'un espace euclidien ou de Hilbert H, on note Π_F la projection sur F. On a les résultats suivants :*

1) $\forall u \in H, \quad \Pi_F u$ est caractérisé par

$$(u - \Pi_F u, v) = 0 \quad \forall v \in F \tag{16.3}$$

2) L'application projection Π_F est linéaire, son noyau est l'orthogonal de F dans H noté $F^\perp$
3) Il existe un couple unique d'applications linéaires Π_F et $\Pi_{F^\perp}$ qui appliquent respectivement H dans F et H dans $F^\perp$ telles que :

$$u = \Pi_F u + \Pi_{F^\perp} u \tag{16.4}$$

Démonstration

1) Conséquence directe de la propriété caractéristique de la projection sur un convexe en prenant successivement $w = \Pi_K u + v$ et $w = \Pi_K u - v$.

Mais on peut aussi en faire une démonstration directe en montrant comme pour le théorème 14.3.1, l'équivalence entre la minimisation de $\|u - v\|$, $\forall v \in F$ et les relations d'orthogonalité $(u - \Pi_F u, v) = 0 \, \forall v \in F$
2) La linéarité se déduit simplement de la propriété caractéristique précédente. Enfin

$$(u, v) = (\Pi_F u, v) \quad \forall v \in F$$

entraîne $(\Pi_F u, v) = 0 \quad \forall v \in F$ si $u \in F^\perp$, d'où le résultat que le noyau de Π_F est $F^\perp$
3) L'existence de Π_F est acquise, on montre simplement par

$$(u - \Pi_F u, v) = 0 \quad \forall v \in F$$

que $(u - \Pi_F u) \in F^\perp$ et le résultat car $u - (u - \Pi_F u) = \Pi_F u$ est orthogonal à $F^\perp$.

16.2 Minimisation de fonctions quadratiques

Soit a une forme bilinéaire symétrique sur un espace euclidien ou de Hilbert E, b un vecteur donné de E, c un réel et $(.,.)$ le produit scalaire dans E. On considère la fonction quadratique sur E : application de $E \times E$ dans $\mathbb{R}$ de la forme :

$$J(v) = \frac{1}{2} a(v, v) - (b, v) + c \tag{16.5}$$

Théorème 16.2.1 *Si la forme bilinéaire a est symétrique définie positive, il existe un élément unique de E qui minimise J et il y a équivalence entre le problème de minimisation de J dans E et le problème :*

$$\begin{cases} \text{Trouver } u \in E \quad \text{tel que} \\ \\ a(u, v) = (b, v) \quad \forall\, v \in E \end{cases} \tag{16.6}$$

16.3 Fonctions convexes

On se place dans un convexe C d'un espace vectoriel E.

Définition 16.3.1 *Une fonction F de C dans $\mathbb{R}$ est convexe si*

$$\forall x, y \in C, \forall \lambda \in [0, 1] \quad F(\lambda x + (1 - \lambda) y) \leq \lambda F(x) + (1 - \lambda) F(y) \tag{16.7}$$

F est strictement convexe si

$$\forall x, y \in C, x \neq y \, \forall \lambda \in]0, 1[\quad F(\lambda x + (1 - \lambda) y) < \lambda F(x) + (1 - \lambda) F(y) \tag{16.8}$$

Comme on le voit dans la définition, une fonction est convexe ou strictement convexe si elle l'est sur tout segment. Ceci permet de ramener les démonstrations de convexité en dimension un.
Les propriétés classiques des fonctions convexes en dimension un résumées sur la figure 16.1 se généralisent ainsi au cas de fonctionnelles convexes sur des espaces

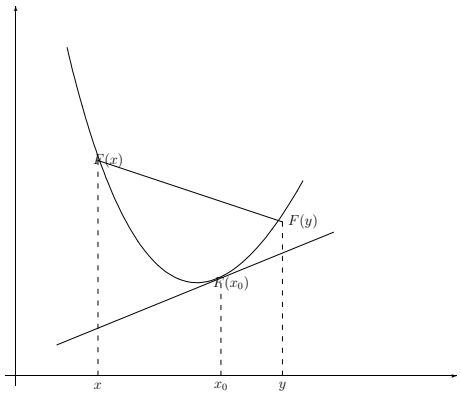


FIGURE 16.1 – Le graphe d'une fonction strictement convexe est au-dessus de ses tangentes et en dessous de ses cordes

vectoriels normés.

Exemple important : considérons une forme bilinéaire symétrique définie positive a et une forme linéaire l sur un espace de Hilbert H. La fonctionnelle J de H dans $\mathbb{R}$ définie par

$$J(v) = \frac{1}{2}a(v,v) - l(v)$$

est strictement convexe sur H (ou tout convexe de H) car sur chaque droite $u+\lambda v$ avec $\lambda \in \mathbb{R}$ et u, v quelconques dans H, , la fonction $f(\lambda) = J(u+\lambda v)$ est une fonction trinôme de la forme

$$f(\lambda) = J(u+\lambda v) = a\lambda^2 + b\lambda + c$$

avec $a > 0$.

16.3.1 Propriétés caractéristiques des fonctions convexes

1) Si F est une fonction convexe sur un sous-ensemble convexe C d'un espace vectoriel, on a l'inégalité de convexité générale :

$$F(\sum \lambda_i x_i) \leq \sum \lambda_i \, F(x_i) \quad \forall x_i \in C, \forall \lambda_i \geq 0 : \sum \lambda_i = 1$$

2) Le graphe d'une fonction convexe est au-dessus de ses tangentes. De façon générale, si $DF(x_0)$ est la différentielle de la fonction convexe F au point x_0, on a l'inégalité :

$$F(x) \geq F(x_0) + \, DF(x_0)\,(x - x_0)$$

l'inégalité étant stricte si F est strictement convexe.

Définition 16.3.2 *Une application f d'un espace de Hilbert H dans lui même est dite monotone si pour tout couple x, y d'éléments de H on a :*

$$(f(x) - f(y), x - y) \geq 0$$

où $(.,.)$ est le produit scalaire dans H.

Théorème 16.3.1 *Une fonctionnelle différentiable F sur un espace de Hilbert est convexe si et seulement si son gradient ∇F est monotone.*

Ce résultat généralise le fait qu'une fonction dérivable réelle de la variable réelle est convexe si et seulement si sa dérivée est monotone croissante.

Théorème 16.3.2 *Une fonctionnelle deux fois différentiable F sur un espace de Hilbert est convexe si et seulement si sa hessienne $HF(x)$ est semi-définie positive et strictement convexe si sa hessienne est définie positive en tout point $x \in H$.*

Ce résultat généralise le fait qu'une fonction deux fois dérivable réelle de la variable réelle est convexe si et seulement si sa dérivée seconde est positive.

16.4 Convexité et optimisation

Définition 16.4.1 *Une fonction F d'un espace de Hilbert H sur $\mathbb{R}$ est dite coercive si*

$$\lim_{\|x\| \to \infty} F(x) = +\infty$$

Une fonction continue admettant un minimum sur tout ensemble compact, on en déduit qu'en particulier, en dimension finie une fonction continue de $\mathbb{R}^n$ dans $\mathbb{R}$ admet au moins un minimum sur tout fermé borné de $\mathbb{R}^n$ et qu'une fonction continue coercive de $\mathbb{R}^n$ dans $\mathbb{R}$ admet au moins un minimum.

Un minimum local d'une fonction est nécessairement un point qui annule son gradient. Inversement, si la fonction est convexe et différentiable sur un Hilbert, un point qui annule son gradient est un minimum global.

En résumé voici les deux théorèmes les plus utiles qui synthétisent les propriétés de minimisation des fonctionnelles strictement convexes.

Théorème 16.4.1 *Une fonctionnelle strictement convexe sur un convexe compact d'un espace vectoriel normé admet un minimum unique*

Théorème 16.4.2 *Une fonctionnelle strictement convexe et coercive sur un convexe d'un espace de Hilbert admet un minimum unique.*

16.5 Optimisation sans contraintes

Il s'agit ici de chercher le minimum d'une fonctionnelle J de $\mathbb{R}^n$ dans $\mathbb{R}$. Nous supposerons la fonctionnelle J, appelé souvent "coût", strictement convexe et suffisamment régulière (en général, deux fois continûment différentiable donc de classe C^2)

16.5.1 Optimisation quadratique

Nous envisageons le cas particulier important où J est une fonctionnelle quadratique de la forme

$$J(x) = \frac{1}{2}(Ax, x) - (b, x)$$

avec A, matrice symétrique définie positive $n \times n$, x et $b \in \mathbb{R}^n$, et $(.,.)$ est le produit scalaire euclidien de $\mathbb{R}^n$.

Dans ce cas la fonctionnelle J est strictement convexe sur $\mathbb{R}^n$ et nous avons obtenu l'équivalence (voir le théorème 14.3.1) entre minimisation de la J et résolution du système linéaire

$$Ax = b$$

On peut donc appliquer à ce problème toutes les méthodes de résolution directes ou itératives vues en 2.7 adaptées au cas de matrice symétriques définies positives, donc la factorisation de Choleski, les méthodes itératives de gradient et de gradient conjugué.

16.5.2 Optimisation convexe

Si la fonctionnelle J est strictement convexe et différentiable, mais non quadratique, on a montré l'équivalence entre recherche du minimum et recherche du zéro du gradient de J. On pourra utiliser l'algorithme de gradient appliqué au cas non-linéaire (2.64), les méthodes de Newton ou de Quasi-Newton pour les systèmes (2.4).

16.6 Optimisation sous contraintes égalité

L'étude de l'optimisation sous contraintes égalité va nous conduire à la notion fondamentale de dualité. Nous introduirons les multiplicateurs de Lagrange et le Lagrangien outils généraux de la minimisation sous-contraintes.

16.6.1 Quelques exemples simples

1. Trouver le rectangle d'aire donnée de périmètre minimal. Donc soient a et b les mesures des côtés du rectangle, trouver le couple a, b qui rélaise le minimum de $J(a, b) = a + b$ sous la contrainte $ab = S$. (Observons que le coût J est linéaire mais que l'ensemble des a, b vérifiant la contrainte n'est pas ici un convexe de $I\!R^2$) On peut résoudre de façon élémentaire

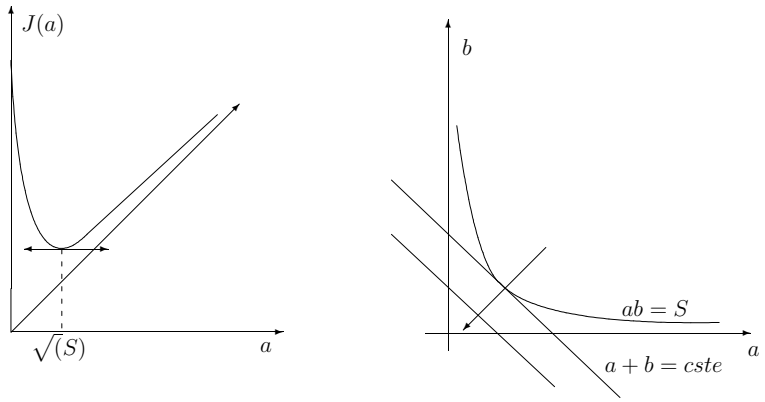


FIGURE 16.2 – Deux représentations graphiques du minimum sous contraintes

ce problème en éliminant une des variables, b par exemple. Une autre approche, plus intéressante pour la suite, consiste à tracer sur un même graphique, les lignes iso-J et l'ensemble des contraintes. On observe alors que le minimum de J se produit en un point a, b où le gradient de J est colinéaire au gradient de la courbe $ab = S$ représentant l'ensemble des contraintes. Un raisonnement intuitif conduit à remarquer que si le gradient de J n'était pas orthogonal à la courbe des contraintes, on pourrait diminuer le coût J en restant sur cette courbe, ce qui est contradictoire avec l'hypothèse : a, b réalise le minimum de J sous la contrainte $ab = S$. Le gradient de J est égal à

$$\nabla J = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

La normale à la courbe des contraintes est colinéaire à

$$\begin{pmatrix} b \\ a \end{pmatrix}$$

La colinéarité de ces deux vecteurs entraîne bien $a = b = \sqrt{S}$.

2. Minimiser la fonction quadratique $J(x, y) = x^2 + 2y^2$ sous la contrainte affine $x + y = 1$. Observons que cette fois on cherche bien à minimiser une fonction strictement convexe sur un convexe. On est donc assuré de l'existence et de l'unicité du minimum. On vérifie à nouveau la colinéarité

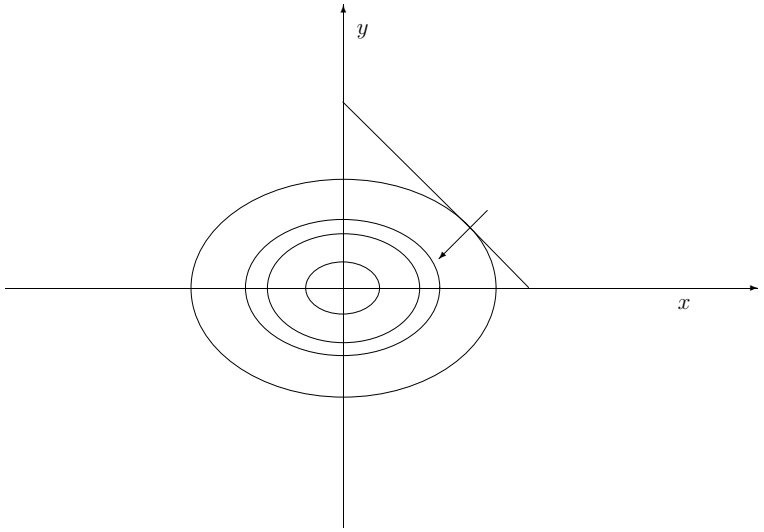


FIGURE 16.3 – Minimum d'une fonction quadratique sous contrainte affine

du vecteur gradient de la fonction coût J avec le vecteur gradient de la courbe des contraintes.

3. Un exemple plus général : la minimisation d'une fonction quadratique

$$J(x) = \frac{1}{2}(Ax, x) - (b, x)$$

avec A, matrice symétrique définie positive $n \times n$, sous l'ensemble de m contraintes égalités affines $Bx = c$, où B est une matrice $m \times n$ et c un vecteur donné de $\mathbb{R}^m$. On suppose la matrice B de rang m. On est ici dans le cas de la minimisation d'une fonctionnelle quadratique strictement convexe sur un convexe donc on est assuré de l'existence et de l'unicité de la solution. La solution x^* peut s'interpréter comme la projection sur le convexe K défini comme l'ensemble des x tels que $Bx = c$ du point $\bar{x}$ réalisant le minimum de J sur l'espace $\mathbb{R}^n$ tout entier et donc solution de $A\bar{x} = b$. En effet

$$J(x) = \frac{1}{2}(Ax, x) - (b, x) = \frac{1}{2}(A(x - \bar{x}), x - \bar{x}) - (b, \bar{x})$$

Donc minimiser J sur K équivaut à minimiser $(A(x - \bar{x}), x - \bar{x})$ donc à projeter $\bar{x}$ au sens du produit scalaire $(x, y) \to (Ax, y)$ sur K.

Le théorème de projection nous donne la caractérisation (voir 16.2) suivante

$$(A(\bar{x} - x^*), x - x^*) \leq 0 \quad \forall x \in K$$

Examinons K. L'ensemble des x tels que $Bx = c$ est l'ensemble des x tels que $B(x - x^*) = 0$. Donc x appartient à K si et seulement si $x - x^*$ appartient au noyau de B, ou ce qui revient au même, est orthogonal à l'ensemble des vecteurs lignes de B donc à l'ensemble des vecteurs colonnes de B^T. [On retrouve ici un résultat classique d'algèbre dans $\mathbb{R}^n$: l'orthogonal du noyau d'une application linéaire est égal à l'image de sa transposée]. Comme l'ensemble des $x - x^*$ est un sous espace de $\mathbb{R}^n$ on a :

$$(A(\bar{x} - x^*), x - x^*) \leq 0 \quad \forall x \in K \iff (A(\bar{x} - x^*), x - x^*) = 0 \quad \forall x \in K$$

On en déduit que $A(\bar{x} - x^*)$ est orthogonal à $x - x^*$ pour tout $x \in K$, Donc $A(\bar{x} - x^*) = b - Ax^*$ appartient au sous-espace de $\mathbb{R}^n$ engendré par les vecteurs colonnes de B^T. En résumé, il existe m scalaires $p_1, p_2, ...p_m$ tels que

$$Ax^* - b + \sum_{i=1}^{m} p_i B_i^T = 0 \tag{16.9}$$

Ce résultat se généralise selon le théorème suivant

Théorème 16.6.1 (de Lagrange) *On considère le problème de minimisation d'une fonctionnelle coût $J(x)$ pour $x \in \mathbb{R}^n$, sous un ensemble de m contraintes*

$$E_i(x) = 0 \quad \forall i = 1, ..m$$

On suppose qu'il existe une solution x^* à ce problème et que les fonctions J et E_i soient continûment dérivables en x^*. On suppose que la matrice jacobienne des contraintes soit de rang m au point x^*, c'est à dire que les m vecteurs gradients des E_i au point x^* soient linéairement indépendants. Alors il existe m scalaires $p_1^*, p_2^*, ...p_m^*$ tels que

$$\nabla J(x^*) + \sum_{i=1}^{m} p_i^* \nabla E_i(x^*) = 0 \tag{16.10}$$

Remarque 16.6.1 *Ce théorème appliqué au cas d'une fonctionnelle J quadratique et de contraintes affines redonne le résultat (16.9).*

16.6.2 Le Lagrangien

Les conditions de Lagrange (16.10) sont des conditions nécessaires, mais non suffisantes en général. On peut les exprimer sous une forme plus concise et plus

commode à l'aide du **Lagrangien**. Reprenons le problème de minimisation d'une fonctionnelle coût $J(x)$ pour $x \in I\!R^n$, sous un ensemble de m contraintes

$$E_i(x) = 0 \quad \forall i = 1, ..m$$

On appelle Lagrangien de ce problème d'optimisation sous contraintes, l'expression

$$\mathcal{L}(x, p) = J(x) + \sum_{i=1}^{p} p_i E_i(x) \tag{16.11}$$

Les conditions d'optimalité de Lagrange s'écrivent donc simplement de la façon suivante

Théorème 16.6.2 *Sous les hypothèses du théorème (16.6.1), si x^* est solution du problème de minimisation sous contraintes et p^* le vecteur de $I\!R^m$ formé par les multiplicateurs de Lagrange, le couple (x^*, p^*) est un point stationnaire du Lagrangien, donc vérifie*

$$\begin{cases} \nabla_x \mathcal{L}(x^*, p^*) = \nabla_x J(x^*) + \sum_{i=1}^{m} p_i^* \nabla E_i(x^*) = 0 \\ \nabla_p \mathcal{L}(x^*, p^*) = 0 \Rightarrow E_i(x) = 0 \quad \forall i = 1, ..m \end{cases}$$

16.6.3 Interprétation des multiplicateurs de Lagrange

Les multiplicateurs de Lagrange ne sont pas des artifices de calcul. Ils représentent chacun la sensibilité du coût optimal J à la contrainte associée. Plus précisément supposons que l'on modifie la contrainte $E_j(x) = 0$ selon $E_j(x) = \epsilon$, soit

$$\mathcal{L}_\epsilon(x, p) = J(x) + p_j(E_j(x) - \epsilon) + \sum_{i=1, i\neq j}^{p} p_i E_i(x)$$

On obtient alors à l'optimum

$$p_j = -\frac{dJ(x_\epsilon^*)}{d\epsilon}_{|\epsilon=0}$$

Application

1) Reprenons l'exemple 1. Le lagrangien du problème s'écrit

$$\mathcal{L}(a, b, p) = a + b + p(ab - S)$$

On obtient les conditions d'optimalité

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial a} = 1 + pb = 0 \\ \frac{\partial \mathcal{L}}{\partial b} = 1 + pa = 0 \\ \frac{\partial \mathcal{L}}{\partial p} = ab - S = 0 \end{cases}$$

d'où la solution $a^* = b^* = \sqrt{S}$ et $p^* = -\frac{1}{\sqrt{S}}$. Si on modifie la contrainte S en $S + \epsilon$, on obtient la solution $a_\epsilon = b_\epsilon = \sqrt{S+\epsilon}$ et donc le coût J devient $J_\epsilon = a_\epsilon + b_\epsilon = 2\sqrt{S+\epsilon}$. On a bien

$$\frac{dJ}{d\epsilon}_{|\epsilon=0} = \frac{1}{\sqrt{S}} = -p^*$$

2) Considérons l'exemple de la minimisation quadratique sous contraintes égalités affines. Le lagrangien s'écrit

$$\mathcal{L}(x,p) = \frac{1}{2}(Ax,x) - (b,x) + (p, Bx - c)$$

et on retrouve les conditions d'optimalité de Lagrange

$$\left\{ \begin{array}{l} \nabla_x \mathcal{L}(x^*,p^*) = Ax^* - b + B^T p^* = 0 \\ \nabla_p \mathcal{L}(x^*,p^*) = Bx^* - c = 0 \end{array} \right.$$

Soit sous forme matricielle

$$\left(\begin{array}{cc} A & B^T \\ B & 0 \end{array} \right) \left(\begin{array}{c} x \\ p \end{array} \right) = \left(\begin{array}{c} b \\ c \end{array} \right) \tag{16.12}$$

Le problème est bien posé si la matrice B est de rang m. On peut résoudre d'abord le problème dual en p, en utilisant $x = A^{-1}B^T p + A^{-1}b$, puis $Bx = c$, ce qui donne le problème dual :

$$BA^{-1}B^T p = BA^{-1}b - c \tag{16.13}$$

qui est bien posé car la matrice $BA^{-1}B^T$, avec les hypothèses faites, est symétrique définie positive.

On obtient ensuite x comme solution du problème primal

$$Ax = b - B^T p$$

A titre d'exercice, on pourra appliquer la démarche précédente au cas de la minimisation de $J(x,y) = x^2 + 2y^2$ sous la contrainte affine $x + y = 1$.

16.6.4 Point-selle du Lagrangien

Définition 16.6.1 *Le couple $\bar{x}, \bar{p}$ est un point-selle du lagrangien $\mathcal{L}(x,p)$ si*

$$\sup_{p \in I\!R^m} \mathcal{L}(\bar{x},p) = \mathcal{L}(\bar{x},\bar{p}) = \inf_{x \in I\!R^n} \mathcal{L}(x,\bar{p})$$

Ceci signifie que $\bar{x}$ réalise le minimum de la fonction $x \to \mathcal{L}(x,\bar{p})$ et que $\bar{p}$ maximise la fonction $p \to \mathcal{L}(\bar{x},p)$.

Théorème 16.6.3 *Si $\bar{x}, \bar{p}$ est un point-selle du lagrangien, on a*

$$\sup_{p\in I\!R^m} \inf_{x\in I\!R^n} \mathcal{L}(x,p) = \mathcal{L}(\bar{x},\bar{p}) = \inf_{x\in I\!R^n} \sup_{p\in I\!R^m} \mathcal{L}(x,p)$$

Démonstration :

On a évidemment

$$\mathcal{L}(x,p) \leq \sup_{p\in I\!R^m} \mathcal{L}(x,p) \quad \forall x, \forall p$$

donc

$$\inf_{x\in I\!R^n} \mathcal{L}(x,p) \leq \inf_{x\in I\!R^n} \sup_{p\in I\!R^m} \mathcal{L}(x,p) \quad \forall p$$

et donc toujours

$$\sup_{p\in I\!R^m} \inf_{x\in I\!R^n} \mathcal{L}(x,p) \leq \inf_{x\in I\!R^n} \sup_{p\in I\!R^m} \mathcal{L}(x,p) \tag{16.14}$$

Inversement, considérons les fonctions $G(x) = \sup_{p\in I\!R^m} \mathcal{L}(x,p)$ et $H(p) = \inf_{x\in I\!R^n} \mathcal{L}(x,p)$.

On a d'une part

$$\mathcal{L}(\bar{x},\bar{p}) = \inf_{x\in I\!R^n} \mathcal{L}(x,\bar{p}) = H(\bar{p})$$

d'autre part

$$\mathcal{L}(\bar{x},\bar{p}) = \sup_{p\in I\!R^m} \mathcal{L}(\bar{x},p) = G(\bar{x})$$

d'où

$$\inf_{x\in I\!R^n} G(x) \leq G(\bar{x}) = \mathcal{L}(\bar{x},\bar{p}) = H(\bar{p}) \leq \sup_{p\in I\!R^m} H(p)$$

Donc

$$\inf_{x\in I\!R^n} \sup_{p\in I\!R^m} \mathcal{L}(x,p) \leq \mathcal{L}(\bar{x},\bar{p}) \leq \sup_{p\in I\!R^m} \inf_{x\in I\!R^n} \mathcal{L}(x,p) \tag{16.15}$$

En rapprochant (16.14) et (16.15) on obtient les égalités annoncées.

16.6.5 Problème dual

La fonction $H(p)$ est la fonction duale du problème d'optimisation, p est la variable duale. Le problème dual du problème de la recherche de la valeur $\bar{x}$ qui minimise le coût $J(x)$ est le problème de la recherche de $\bar{p}$ qui maximise $H(p)$. Une fois le problème dual résolu, donc $\bar{p}$ obtenu, on obtient $\bar{x}$ comme solution du problème de minimisation **sans contraintes**

$$\mathcal{L}(\bar{x},\bar{p}) = \inf_{x\in I\!R^n} \mathcal{L}(x,\bar{p})$$

Théorème 16.6.4 *Si $\bar{x}, \bar{p}$ est point-selle du lagrangien $\mathcal{L}(x,p) = J(x) + (p, E(x))$, alors $\bar{x}$ est solution du problème de minimisation de $J(x)$ sous les contraintes $E(x) = 0$.*

Ce théorème d'optimalité du point-selle est une conséquence directe du précédent. En effet

$$\mathcal{L}(\bar{x}, p) \leq \mathcal{L}(\bar{x}, \bar{p}) \quad \forall p$$

entraîne

$$(p, E(\bar{x})) \leq (\bar{p}, E(\bar{x})) \quad \forall p$$

donc nécessairement $E(\bar{x}) = 0$, les contraintes sont vérifiées.
D'autre part

$$\mathcal{L}(\bar{x}, \bar{p}) \leq \mathcal{L}(x, \bar{p}) \quad \forall x \quad \text{tel que } E(x) = 0$$

entraîne $J(\bar{x}) \leq J(x) \quad \forall x \quad$ tel que $E(x) = 0$

Application : Le cas d'une fonctionnelle quadratique sous contraintes affines

$$\mathcal{L}(x,p) = \frac{1}{2}(Ax, x) - (b, x) + (p, Bx - c)$$

Par définition la fonction duale $H(p) = \inf_x \mathcal{L}(x,p)$. On calcule

$$\nabla_x \mathcal{L}(x,p) = Ax - b + B^T p = 0$$

La valeur x^* minimisant $\mathcal{L}(x,p)$ annule son gradient. On obtient $x^* = A^{-1}(b - B^T p)$. La fonction duale $H(p)$ s'écrit :

$$H(p) = \frac{1}{2}(Ax^*, x^*) - (b, x^*) + (p, Bx^* - c)$$

Tous calculs faits, on obtient

$$H(p) = -\frac{1}{2}(A^{-1}(B^T p - b), B^T p - b) - (p, c)$$

H est une fonction strictement concave qui admet un maximum unique obtenu en écrivant

$$\nabla H(p) = -BA^{-1}B^T p + BA^{-1}b - c = 0$$

Noter que l'on retrouve bien (16.13).

16.6.6 Algorithme d'Uzawa

L'algorithme d'Uzawa utilise la dualité introduite ci-dessus. On ramène le calcul du minimum de la fonctionnelle coût J sous les contraintes $E(x) = 0$, à une double itération successive de recherche de maximums de la fonction duale H et de minimums du problème primal. Le maximum de H étant obtenu par une méthode de gradient à pas fixe.

Initialisation : p^0 donné, $\rho > 0$ assez petit

Pour $k \to k+1$ faire

Calculer x^k solution du problème primal $\min_x \mathcal{L}(x, p^k)$

Tester $J(x^k) \leq J(x^{k-1})$ sinon diminuer ρ

Calculer $p^{k+1} = p^k + \rho E(x^k)$

Tant que $\|E(x^k)\| > \epsilon \|E(x^0)\|$

Cet algorithme ramène la résolution d'un problème de minimisation sous contraintes à la résolution d'une suite de problèmes d'optimisation sans contraintes. Il a le défaut d'être assez lent. Une amélioration possible consiste à remplacer le Lagrangien par un "Lagrangien augmenté" selon

$$\mathcal{L}^*(x, p) = \mathcal{L}(x, p) + \frac{c}{2}\|E(x)\|^2$$

Ceci présente le double avantage de produire des problèmes mieux conditionnés et de résoudre le problème délicat du choix du paramètre ρ. En effet, on peut montrer que l'algorithme d'Uzawa appliqué au Lagrangien augmenté converge avec $\rho = c$.

16.6.7 Convergence de l'algorithme d'Uzawa dans le cas d'une fonctionnelle coût quadratique sous contraintes affines

Soit $J(x) = \frac{1}{2}(Ax, x) - (b, x)$ pour $x \in I\!R^n$ et A symétrique définie positive. Et l'ensemble de m contraintes affines $Bx = c$, avec B matrice $m \times N$ de rang m. Le couple (x^*, p^*) est point-selle du Lagrangien

$$\mathcal{L}(x, p) = \frac{1}{2}(Ax, x) - (b, x) + (p, Bx - c)$$

si

$$\left\{ \begin{array}{l} \nabla_x \mathcal{L}(x^*, p^*) = Ax - b + B^T p = 0 \\ \nabla_p \mathcal{L}(x^*, p^*) = Bx - c = 0 \end{array} \right.$$

L'algorithme d'Uzawa construit l'itération

$$\left\{ \begin{array}{l} Ax^k = b - B^T p^k \\ p^{k+1} = p^k + \rho(Bx - c) \end{array} \right.$$

On en déduit $A(x^k - x^*) = -B^T(p^k - p^*)$ donc $x^k - x^* = -A^{-1}\, B^T(p^k - p^*)$ et

$$p^{k+1} - p^* = p^k - p^* - \rho(BA^{-1}\, B^T(p^k - p^*)$$

soit

$$p^{k+1} - p^* = [Id - \rho BA^{-1}\, B^T](p^k - p^*)$$

La suite p^k converge donc vers p^* si $\|Id - \rho BA^{-1}\, B^T\| < 1$, soit, en considérant la norme euclidienne de cette matrice symétrique, si son rayon spectral, c'est à dire le maximum en module des valeurs propres, est inférieur à 1. Un calcul simple donne, en notant classiquement λ la valeur propre d'une matrice : $\lambda(Id - \rho BA^{-1}B^T) = 1 - \rho\lambda(BA^{-1}B^T)$ et

$$\lambda(BA^{-1}B^T) \leq \frac{\lambda_{max}(BB^T)}{\lambda_{min}(A)}$$

en rappelant que les matrices A et BB^T sont symétriques définies positives.
On en déduit la convergence sous la condition

$$\rho < 2\, \frac{\lambda_{min}(A)}{\lambda_{max}(BB^T)}$$

16.6.8 Pénalisation

Une autre technique possible, pour ramener simplement un problème de minimisation sous contraintes à la résolution d'un problème d'optimisation sans contraintes, consiste à pénaliser la contrainte. On ajoute à la fonction coût à minimiser un terme de pénalisation qui devient très grand quand les contraintes ne sont pas satisfaites.

Par exemple, on peut remplacer le problème de minimisation de $J(x)$ sous l'ensemble de contraintes $E(x) = 0$ par la minimisation du coût pénalisé

$$J_\epsilon(x) = J(x) + \frac{1}{2\epsilon}\|E(x)\|^2$$

avec ϵ petit. Cette technique à l'avantage d'être très simple à utiliser mais l'inconvénient de conduire à des systèmes mal conditionnés.

16.7 Optimisation sous contraintes inégalités

On considère maintenant le problème d'optimisation avec contraintes inégalités. Plus précisément il s'agit ici de chercher le minimum d'une fonctionnelle J de $\mathbb{R}^n$ dans $\mathbb{R}$, sous un ensemble de contraintes égalités et inégalités

$$E_i(x) \leq 0 \quad \forall i = 1, ..m$$

Nous ferons l'hypothèse que la fonction coût J et les contraintes sont convexes et différentiables. Le problème s'énonce alors comme un problème de minimisation d'une fonctionnelle convexe sur un convexe.

Définition 16.7.1 (contraintes actives) *On dit que la contrainte $E_i(x) \leq 0$ est active ou saturée au point x si $E_i(x) = 0$.*

Définition 16.7.2 *On dit qu'un point x de l'ensemble admissible*

$$C = \{x | E_i(x) \leq 0 \quad \forall i = 1, ..m\}$$

est un point régulier si les gradients des contraintes actives en ce point sont linéairement indépendants.

On peut avoir pour un même problème un ensemble de contraintes composé de contraintes égalités et de contraintes inégalités.

16.7.1 Théorème de Kuhn et Tucker

Le théorème d'optimalité de Kuhn et Tucker étend au cas de contraintes inégalité le théorème d'optimalité de Lagrange (16.6.1).

Théorème 16.7.1 (Kuhn et Tucker) *Si x^* est un point régulier de l'ensemble des contraintes et si x^* est un minimum local de J alors il existe m réels p_i^* positifs ou nuls tels que*

$$\begin{cases} \nabla_x J(x^*) + \sum_{i=1}^m p_i^* \nabla E_i(x^*) = 0 \\ \\ p_i^* E_i(x^*) = 0 \quad \forall i = 1, ..m \end{cases}$$

Si la fonction coût J est strictement convexe et l'ensemble des contraintes est convexe ce minimum est unique.

La deuxième condition $p_i^* E_i(x^*) = 0 \quad \forall i = 1, ..m$ est plus compliquée qu'il ne semble à première vue. Elle présente un aspect combinatoire qu'illustre l'exemple suivant tiré du livre de Jean-Christophe Culioli : Introduction à l'optimisation.

Un exemple

Considérons le problème simple suivant. Trouver le vecteur (x_1, x_2) de $\mathbb{R}^2$ qui minimise le coût

$$J(x_1, x_2) = \frac{1}{2}(x_1 - 1)^2 + \frac{1}{2}(x_2 - 2)^2$$

sous l'ensemble de contraintes

$$\begin{cases} x_1 - x_2 = 1 \\ x_1 + x_2 \leq 2 \\ -x_1 \leq 0 \\ -x_2 \leq 0 \end{cases}$$

Ecrivons les conditions d'optimalité de Kuhn et Tucker :

$$\begin{cases} \frac{\partial}{\partial x_1}[J(x) + \sum_{i=1}^{m} p_i E_i(x)] = x_1 - 1 + p_1 + p_2 - p_3 = 0 \\ \frac{\partial}{\partial x_2}[J(x) + \sum_{i=1}^{m} p_i E_i(x)] = x_2 - 2 - p_1 + p_2 - p_4 = 0 \\ x_1 - x_2 = 1 \\ p_2(x_1 + x_2 - 2) = 0 \\ p_3(x_1) = 0 \\ p_4(x_2) = 0 \end{cases}$$

Comment peut-on résoudre ce problème ? La technique des contraintes actives consiste à commencer par résoudre le problème avec les seules contraintes égalité. Ici $x_1 - x_2 = 1$. On obtient le système

$$\begin{cases} x_1 - 1 + p_1 = 0 \\ x_2 - 2 - p_1 = 0 \\ x_1 - x_2 = 1 \end{cases}$$

La résolution de ce système donne $x_1 = 2$, $x_2 = 1$. Ce qui contredit la contrainte $x_1 + x_2 \leq 2$. On introduit alors cette contrainte comme une contrainte active donc sous forme égalité. On obtient le nouveau système

$$\begin{cases} x_1 - 1 + p_1 + p_2 = 0 \\ x_2 - 2 - p_1 + p_2 = 0 \\ x_1 - x_2 = 1 \\ x_1 + x_2 = 2 \end{cases}$$

On obtient alors $x_1 = \frac{3}{2}$, $x_2 = \frac{1}{2}$. Et ici toutes les contraintes sont vérifiées. On a la solution.

Mais on imagine bien qu'avec un grand nombre de variables et de contraintes ce petit jeu peut devenir très long.

16.7.2 Lagrangien généralisé

On peut reformuler les conditions d'optimalité de Kuhn et Tucker à l'aide du Lagrangien comme dans le cas des contraintes égalités. Mais ici, les multiplicateurs de Lagrange sont astreints à être positifs ou nuls. On appelle Lagrangien de ce problème d'optimisation sous contraintes, l'expression

$$\mathcal{L}(x,p) = J(x) + \sum_{i=1}^{p} p_i E_i(x) \tag{16.16}$$

où les multiplicateurs de Lagrange sont cette fois positifs ou nuls : $p_i \geq 0$. Les conditions d'optimalité de Lagrange s'écrivent donc simplement de la façon suivante

Théorème 16.7.2 *Si x^* est solution du problème de minimisation sous contraintes inégalités et p^* le vecteur de $\mathbb{R}^m_+$ formé par les multiplicateurs de Lagrange, le couple (x^*, p^*) vérifie*

$$\begin{cases} \nabla_x \mathcal{L}(x^*,p^*) = \nabla_x J(x^*) + \sum_{i=1}^m p_i^* \nabla E_i(x^*) = 0 \\ p_i E_i(x) = 0 \quad \forall i = 1,..m \\ E_i(x) \leq 0, \quad p_i \geq 0 \quad \forall i = 1,..m \end{cases}$$

On peut alors reprendre les définitions des points-selles et les résultats d'optimalité obtenus dans le cas de contraintes égalités en restreignant les multiplicateurs à des réels positifs ou nuls.

16.7.3 Point-selle du Lagrangien généralisé

Définition 16.7.3 *Le couple $\bar{x}, \bar{p}$ est un point-selle du lagrangien généralisé $\mathcal{L}(x,p)$ si*

$$\sup_{p \in \mathbb{R}^m_+} \mathcal{L}(\bar{x},p) = \mathcal{L}(\bar{x},\bar{p}) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x,\bar{p})$$

Ceci signifie que $\bar{x}$ réalise le minimum de la fonction $x \to \mathcal{L}(x,\bar{p})$ et que $\bar{p}$ maximise la fonction $p \to \mathcal{L}(\bar{x},p)$.

Théorème 16.7.3 *Si $\bar{x}, \bar{p}$ est un point-selle du lagrangien, on a*

$$\sup_{p \in \mathbb{R}^m_+} \inf_{x \in \mathbb{R}^n} \mathcal{L}(x,p) = \mathcal{L}(\bar{x},\bar{p}) = \inf_{x \in \mathbb{R}^n} \sup_{p \in \mathbb{R}^m_+} \mathcal{L}(x,p)$$

Démonstration : La démonstration se fait comme dans le cas de contraintes égalités.

16.7.4 Problème dual

La fonction $H(p) = \inf_{x \in I\!R^n} \mathcal{L}(x,p)$ est encore la fonction duale du problème d'optimisation, p est la variable duale. Le problème dual du problème de la recherche de la valeur $\bar{x}$ qui minimise le coût $J(x)$ est le problème de la recherche de $\bar{p}$ qui maximise $H(p)$. Une fois le problème dual résolu, donc $\bar{p}$ obtenu, on obtient $\bar{x}$ comme solution du problème de minimisation **sans contraintes**

$$\mathcal{L}(\bar{x},\bar{p}) = \inf_{x \in I\!R^n} \mathcal{L}(x,\bar{p})$$

Théorème 16.7.4 *Si $\bar{x},\bar{p}$ est point-selle du lagrangien $\mathcal{L}(x,p) = J(x) + (p,E(x))$, alors $\bar{x}$ est solution du problème de minimisation de $J(x)$ sous les contraintes $E_i(x) \leq 0$.*

Ce théorème d'optimalité du point-selle est une conséquence directe du précédent. En effet

$$\mathcal{L}(\bar{x},p) \leq \mathcal{L}(\bar{x},\bar{p}) \quad \forall p$$

entraîne

$$(p,E(\bar{x})) \leq (\bar{p},E(\bar{x})) \quad \forall p \in I\!R^m_+$$

donc nécessairement $(p-\bar{p},E(\bar{x}) \leq 0$. Prenons $p = \bar{p} + e_i$, on en déduit $E_i(x) \leq 0 \quad \forall i$, les contraintes sont vérifiées. Et de plus, en prenant successivement $p = 0$ et $p = 2\bar{p}$, on obtient la relation $(\bar{p},E(\bar{x}) = O$
D'autre part

$$\mathcal{L}(\bar{x},\bar{p}) \leq \mathcal{L}(x,\bar{p}) \quad \forall x$$

entraîne $J(\bar{x}) \leq J(x) \quad \forall x \quad$ tel que $E(x) \leq 0$

16.7.5 Quelques algorithmes

Voici quelques algorithmes classiques pour résoudre les problèmes d'optimisation avec contraintes inégalités.

Le gradient projeté

Cet algorithme est très simple à utiliser si l'on peut facilement projeter un point dans le convexe des contraintes admissibles. C'est le cas si les contraintes sont du type assez fréquent en pratique $x_i \geq 0$. Le projeté se fait alors simplement en prenant $\bar{x}_i = max(x_i,0)$. La méthode du gradient projeté s'obtient alors facilement à partir de l'algorithme du gradient selon

$$\left\{ \begin{array}{l} \tilde{x}^{k+1} = x^k - \rho \nabla J(x^k) \\ x^{k+1} = P_K(\tilde{x}^{k+1}) \text{ selon } x_i^{k+1} = max(0,\tilde{x}_i^{k+1}) \end{array} \right.$$

L'algorithme d'Uzawa

On reprend l'algorithme (16.6.6) en l'adaptant. L'algorithme d'Uzawa utilise la dualité introduite ci-dessus. On ramène le calcul du minimum de la fonctionnelle coût J sous les contraintes $E_i(x) \leq 0$, à une double itération successive de recherche de maximums de la fonction duale H et de minimum du problème primal. Le maximum de H étant obtenu par une méthode de gradient à pas fixe.

Initialisation : $p^0 \in I\!R^m_+$ donné , $\rho > 0$ assez petit

Pour $k \to k+1$ faire

Calculer x^k solution du problème primal $\min_x \mathcal{L}(x, p^k)$

Tester $J(x^k) \leq J(x^{k-1})$ sinon diminuer ρ

Calculer $p^{k+1} = P_K(p^k + \rho E(x^k))$ où P_K est la projection sur les $p_i \geq 0$

Tant que $\|E(x^k)\| > \epsilon \|E(x^0)\|$

16.7.6 Convergence de l'algorithme d'Uzawa dans le cas d'une fonctionnelle coût quadratique sous contraintes inégalités affines

Soit $J(x) = \frac{1}{2}(Ax, x) - (b, x)$ pour $x \in I\!R^n$ et A symétrique définie positive. Et l'ensemble de m contraintes affines $Bx \leq c$, avec B matrice $m \times N$ de rang égal au nombre de contraintes actives. Le couple (x^*, p^*) est point-selle du Lagrangien

$$\mathcal{L}(x, p) = \frac{1}{2}(Ax, x) - (b, x) + (p, Bx - c)$$

si

$$\left\{ \begin{array}{l} \nabla_x \mathcal{L}(x^*, p^*) = Ax^* - b + B^T p^* = 0 \\ E_i(x^*) \leq 0, \quad p_i^* \geq 0 \quad \forall i = 1, ..m \end{array} \right.$$

L'algorithme d'Uzawa construit l'itération

$$\left\{ \begin{array}{l} Ax^k = b - B^T p^k \\ p^{k+1} = P_K(p^k + \rho(Bx - c)) \end{array} \right.$$

On en déduit $A(x^k - x^*) = -B^T(p^k - p^*)$ donc $x^k - x^* = -A^{-1}\, B^T(p^k - p^*)$ et

$$p^{k+1} - p^* = P_K(p^k - p^* - \rho(BA^{-1}\, B^T(p^k - p^*)))$$

soit

$$p^{k+1} - p^* = P_K([Id - \rho BA^{-1}\, B^T](p^k - p^*))$$

La projection sur un convexe est une application contactante. Donc, on peut majorer la norme de $p^{k+1} - p^*$ par celle de $[Id - \rho BA^{-1}\, B^T](p^k - p^*)$. On se ramène ainsi à la démonstration du cas des contraintes égalités. La suite p^k converge donc vers p^* si $\|Id - \rho BA^{-1}\, B^T\| < 1$, soit, en considérant la norme euclidienne de cette matrice symétrique, si son rayon spectral, c'est à dire le maximum en module des valeurs propres, est inférieur à 1. Un calcul simple donne, en notant classiquement λ la valeur propre d'une matrice : $\lambda(Id - \rho BA^{-1}B^T) = 1 - \rho\lambda(BA^{-1}B^T)$ et

$$\lambda(BA^{-1}B^T) \leq \frac{\lambda_{max}(BB^T)}{\lambda_{min}(A)}$$

en rappelant que les matrices A et BB^T sont symétriques définies positives.
On en déduit la convergence sous la condition

$$\rho < 2\,\frac{\lambda_{min}(A)}{\lambda_{max}(BB^T)}$$

Pénalisation extérieure

On reprend l'idée de la pénalisation des contraintes selon

$$J_\epsilon(x) = J(x) + \frac{1}{2\epsilon}\|E^+(x)\|^2$$

avec $E_i^+(x) = max(0, E_i(x))$. Ici encore la méthode est très simple à mettre en œuvre mais souffre d'un mauvais conditionnement.

Pénalisation intérieure

On introduit dans ce cas des fonctions barrières qui tendent vers l'infini lorsque l'on s'approche des frontières du convexe des contraintes admissibles. On peut choisir par exemple des fonctions logarithmes en pénalisant le coût selon

$$J_{\epsilon_k}(x) = J(x) - \epsilon_k \sum_i ln(-E_i(x))$$

Il faut alors définir une stratégie adaptée de la variation du paramètre ϵ_k au cours des itérations. ϵ_k doit tendre vers zéro à la convergence des itérations. La barrière créée par la pénalisation collant de plus en plus près de la frontière du convexe des points admissibles.

Chapitre 17

Optimisation et problèmes inverses

17.1 Introduction

L'optimisation est un sujet trop vaste pour que nous puissions en donner un exposé exhaustif dans le cadre de cet ouvrage. C'est aussi un sujet en évolution constante et ses applications se développent de plus en plus. L'idée est naturelle, une fois assurés d'outils de simulation satisfaisants, de les utiliser de façon inverse comme une aide à la conception. Par exemple, une fois que l'on dispose d'un bon modèle de simulation d'un écoulement autour d'une aile d'avion, on en vient naturellement à essayer de résoudre le problème de la conception de la "meilleure" aile possible par rapport à un ensemble de critères donnés.

Nous ne ferons pas un exposé complet des méthodes et des algorithmes d'optimisation, ce qui nécessiterait un ouvrage spécifique de plusieurs volumes (nous renvoyons à la littérature citée dans la bibliographie).

Mais nous concentrerons notre exposé sur notre expérience numérique avec un souci constant de privilégier les techniques simples et réellement utiles. En particulier, nous présentons une méthode nouvelle d'optimisation, basée sur l'analogie entre algorithmes de minimisation et résolution numérique d'équations différentielles par une méthode de tir. Nous détaillerons les différentes étapes nécessaires à la mise au point d'une chaîne d'optimisation, de contrôle ou de résolution d'un problème inverse. Lors de la résolution de ces problèmes par une méthode de minimisation, on est amené à aborder les points suivants :

- Définition du problème de minimisation, de la fonctionnelle, des contraintes et de l'équation d'état.
- Définition de l'espace de contrôle ou paramétrisation.
- Choix d'une approche pour la prise en compte des contraintes.
- Évaluation des gradients des fonctionnelles et des contraintes.

— Choix d'une méthode de minimisation.
— Choix de stratégies minimisant la complexité de l'ensemble ci-dessus.

Nous commencerons par un rappel succinct des résultats théoriques. Les théorèmes classiques d'existence, d'unicité et de convergence des méthodes de résolution supposent que la fonctionnelle coût à minimiser et le domaine admissible soient convexes. Ce contexte favorable n'est pas toujours, on le verra, celui des applications où l'on pourra au mieux parler d'une convexité locale. On cherchera donc souvent des solutions sous-optimales aux problèmes d'optimisation, plutôt que des optima globaux. De même, les méthodes déterministes de résolution nécessitent souvent le calcul du gradient de la fonction coût à minimiser. Ceci est impossible si la fonctionnelle est non-différentiable. Pourquoi alors utiliser ces méthodes ? On constate que, malgré l'existence d'opérations non-différentiables dans la boucle de simulation, la correspondance paramétrisation-fonctionnelle est assez lisse. Cette observation donne une certaine légitimité aux méthodes déterministes.

D'autre part, le souci de simplicité de mise en oeuvre, l'utilisation croissante d'outils boîte-noire, ainsi que la nécessité d'être parfois réalisable expérimentalement, font que les différences finies restent encore la méthode la plus utilisée pour l'évaluation des gradients, et ceci malgré leur forte sensibilité au choix des incréments et malgré leur coût en stockage (taille de l'espace de contrôle). Nous présenterons d'autres approches permettant de contourner ces difficultés et surtout de réduire la complexité des problèmes d'optimisation.

17.2 Paramétrisation

On appelle **paramétrisation** l'ensemble des **variables indépendantes** d'un problème. En général la paramétrisation d'un problème n'est pas unique : pensez aux différentes façon de décrire un cercle : en polaire, en cartésiennes, mais aussi en se donnant une spline, deux splines, ..., ou plutôt un ensemble de points reliés entre eux (formant des segments). Bien sûr les deux premières paramétrisations limitent la description des formes aux cercles et si on utilise ces paramétrisations dans un problème d'optimisation, on ne peut espérer trouver que des cercles (ou plutôt, le cercle le plus intéressant). Inversement, en prenant des splines cubiques (courbes $C^2(s)$ avec s l'abscisse curviligne), on pourra atteindre d'autres formes. Cependant, une spline seule ne pourra pas atteindre une forme ayant des discontinuités de normales. Ainsi, le nombre de pointes que l'on sera susceptible de capturer sera égal aux nombres de splines. Les choses seraient simples si la complexité des calculs ne dépendait pas, souvent de façon croissante, de la dimension de l'espace de paramétrisation. Ceci suggère parfois, comme on le verra plus loin d'utiliser des méthodes différentes pour la résolution des problèmes

directs et inverses.

Cette étape de paramétrisation est donc importante car elle caractérise la taille du problème d'optimisation, mais aussi définit l'espace dans lequel on peut trouver des solutions optimales.

17.3 Définition du Problème

Considérons une boucle de simulation liant une paramétrisation x à une fonctionnelle positive J, mesurant un coût ou une contre-performance à minimiser, via la solution d'une équation d'état :

$$J(x) : x \to q(x) \to U(q(x)) \to J(x, q(x), U(q(x))),$$

où q et U sont des variables intermédiaires où dépendantes. L'état U est solution de l'équation d'état $(E(x, q(x), U(q(x))) = 0)$ et représente, en général, la quantité la plus complexe à évaluer.

Le problème de minimisation pour J sur un espace admissible pour la paramétrisation $\mathcal{O}_{ad} \subset I\!R^N$ s'écrit :

$$\inf\{J(x, q(x), U(q(x))), \quad x \in \mathcal{O}_{ad}\}. \tag{17.1}$$

Nous commencerons par un rappel des résultats théoriques classiques d'existence d'unicité et de convergence des méthodes de résolution du problème (17.1).

17.4 Résultats de base de l'optimisation sans contraintes

Nous avons déjà utilisé le résultat le plus simple de l'optimisation lors de la présentation de la méthode des éléments finis comme méthode de minimisation de l'énergie. Généralisé, ce résultat s'énonce ainsi.

17.4.1 Théorème général

Théorème 17.4.1 *Soit J une fonction donnée d'un espace de Hilbert H dans $I\!R$, couramment nommée fonctionnelle sur H en optimisation.*

Sous les hypothèses :

1. *J est strictement convexe et dérivable sur H.*
2. *J est coercive sur H : $(\lim_{\|x\| \to \infty} J(x) = +\infty)$,*

alors J admet un minimum unique dans H.

Ce minimum est solution du problème

$$J'(x).d = \lim_{\theta \to 0} \frac{J(x+\theta d) - J(x)}{\theta} = 0, \quad \forall d \in H,$$

où J' est la dérivée de la fonctionnelle J.

[En dimension finie, si $H = \mathbb{R}^N$, le minimum est le zéro du gradient ∇J.]

17.4.2 Projection sur un sous-espace fermé

Un exemple fondamental de minimisation est donné par la recherche de la meilleure approximation dans un sous espace F d'un Hilbert H au sens de sa norme. Il s'agit de trouver l'élément du sous-espace le plus proche, au sens de la norme de l'espace de Hilbert, d'un élément u quelconque du Hilbert. On doit donc minimiser la fonctionnelle strictement convexe $\|u-v\|$ sur F (voir annexe B). Les applications sont nombreuses et variées. Exprimée dans une bas orthogonale, la meilleure approximation coïncide avec le développement de Fourier. On retrouve également toutes les techniques d'approximation au sens des moindres carrés. La méthode des éléments finis est une méthode d'approximation au sens des moindres carrés en considérant comme norme, une norme déduite de l'énergie minimisée et comme espace de projection, le sous-espace obtenu par la méthode d'interpolation choisie.

17.4.3 Équivalence entre résolution d'un système linéaire et minimisation quadratique

Un cas particulier important de fonctionnelle strictement convexe sur un Hilbert est le suivant :

$$J(v) = \frac{1}{2}a(v,v) - l(v) \tag{17.2}$$

avec a forme bilinéaire symétrique continue elliptique dans V et l forme linéaire continue dans V. C'est cette fonctionnelle d'énergie qui est minimisée dans les méthodes variationnelles de résolution d'équations aux dérivées partielles linéaires elliptiques comme la méthode des éléments finis.

Dans le cas simple où $V = \mathbb{R}^n$, on obtient le résultat suivant

Théorème 17.4.2 *Si A est une matrice symétrique définie positive, il y a équivalence entre les deux problèmes suivants (voir annexe A) :*

$$(1) \qquad \begin{cases} \textit{Trouver } X \in I\!R^N \quad \textit{tel que} \\ \\ A\,X = B \end{cases} \tag{17.3}$$

$$(2) \qquad \begin{cases} \textit{Trouver } X \in I\!R^N \quad \textit{tel que} \\ \\ J(X) = \dfrac{1}{2}(A\,X, X) - (B, X) \quad \textit{soit minimal} \end{cases} \tag{17.4}$$

17.4.4 Application aux moindres carrés

Considérons le problème général d'un système linéaire sur-déterminé, c'est à dire dans lequel il y a plus d'équations que d'inconnues. C'est en particulier le cas dans le calcul de la droite des moindres carrés ou plus généralement de polynômes d'approximation au sens des moindres carrés. On ne peut pas obtenir exactement l'égalité

$$AX = B$$

car A est une matrice rectangulaire de N lignes et m colonnes avec $N >> m$. On essaie alors de minimiser l'écart entre les vecteurs AX et B de $I\!R^N$ en minimisant la norme euclidienne de leur différence, ou ce qui revient au même le carré de cette norme.

$$Minimiser\, J(X) = \|AX - B\|^2 = (AX - B, AX - B)$$

On utilise les propriétés classiques du produit scalaire $(AU, V) = (U, A^T V)$ pour obtenir :

$$J(X) = (A^T AX, X) - 2(A^T B, X) + (B, B)$$

la matrice $A^T A$ est une matrice carrée $m \times m$ symétrique définie positive (Les m vecteurs colonnes de A sont supposés indépendants). Le théorème (17.4.2) nous donne l'équivalence de ce problème de moindres carrés avec la résolution du système linéaire

$$A^T AX = A^T B$$

On retrouve ainsi le système carré de m équations à m inconnues, dit "système des équations normales".

17.4.5 Exemples de problèmes d'optimisation sans contraintes

Nous allons présenter deux exemples simples de problèmes de minimisation sans contraintes.

Moulage par thermo-formage

On considère le domaine Ω représentant une section plane d'un moule chauffant pour thermo-formage de plastique (Figure 17.1) dont on cherche à calculer la répartition de température.

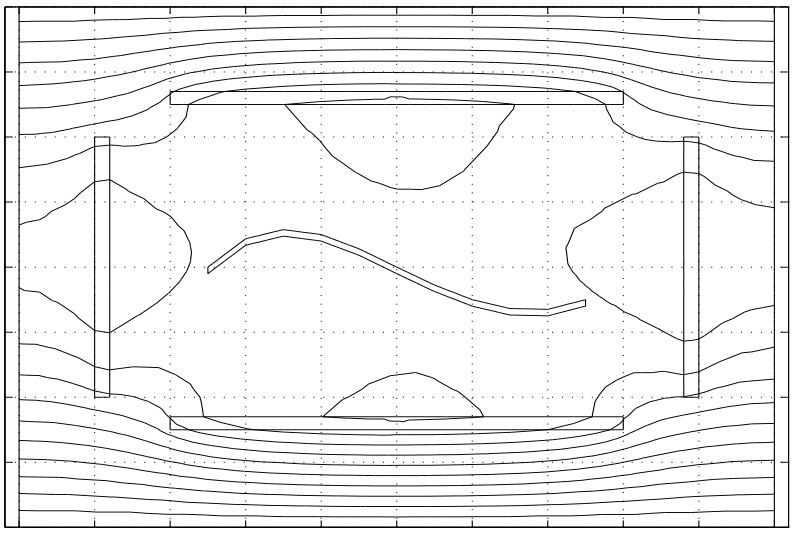


FIGURE 17.1 – Moule thermique.

On suppose le moule chauffé par 4 résistances électriques et maintenu à une température de 20 degrés sur son bord inférieur Γ_1, de 50 degrés sur son bord supérieur Γ_2 et isolée thermiquement sur ses bords latéraux Γ_0.
Soit u la température dans Ω. La modélisation de ce problème s'écrit

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = 20 & \text{sur } \Gamma_1 \\ u = 50 & \text{sur } \Gamma_2 \\ \dfrac{\partial u}{\partial n} = 0 & \text{sur } \Gamma_0 \end{cases} \tag{17.5}$$

Question : Par une méthode de moindres carrés déterminer les valeurs à donner aux puissances f dans les 4 résistances symbolisées ici par les rectangles D_2, D_3, D_4, D_5 pour obtenir la température la plus proche possible de 150 degrés sur le profil γ.

Il s'agit d'un exemple typique de problème inverse simple, car linéaire. En effet, en utilisant le principe de superposition, on peut décomposer la solution inconnue

en la combinaison de 5 solutions.

$$u = u_1 + f_2 u_2 + f_3 u_3 + f_4 u_4 + f_5 u_5.$$

Ainsi, suivant les notations du problème de minimisation ci-dessus, les paramètres de contrôle sont ici $x = \{f_2, f_3, f_4, f_5\}$.

La solution de base u_1 s'obtient par résolution du problème 17.5 dans le cas $f = 0$, donc en l'absence de chauffage du moule. Les quatre solutions élémentaires u_2, u_3, u_4, u_5 s'obtiennent respectivement en mettant à un la résistance associée au rectangle d'indice correspondant et en annulant les trois résistances restantes ainsi que les conditions aux limites. Il est facile de vérifier, par linéarité, que la solution globale u obtenue dans le cas des puissances de chauffe f_2, f_3, f_4, f_5, est bien donnée par la combinaison ci-dessus.

Le problème se ramène au problème de minimisation de la somme des carrés des écarts à 150 degrés des températures sur le profil γ.

$$\min_{x=f_2,f_3,f_4,f_5} \sum_{X_j \in \gamma} (150 - u(X_j))^2. \tag{17.6}$$

La stratégie de calcul est donc simple. Une fois déterminées la solution de base u_1 et les solutions élémentaires u_2, u_3, u_4, u_5, les coefficients f_2, f_3, f_4, f_5 de la solution optimale sont solutions d'un problème de moindres carrés. Le vecteur des valeurs de $f_2u_2 + f_3u_3 + f_4u_4 + f_5u_5$ sur le profil γ est la projection du vecteur $150 - u_1$ correspondant. On obtient donc en définitive les coefficients f_2, f_3, f_4, f_5 en résolvant un système linéaire de 4 équations à 4 inconnues.

$$Ax = G$$

avec $A_{i,j} = \sum_{X_k \in \gamma} u_i(X_k)\, u_j(X_k)$ et $G_i = \sum_{X_k \in \gamma} (150 - u_1(X_k)\, u_i(X_k)$.

Localisation GPS et suivi de sources

Nous rencontrons ce problème de minimisation lors du positionnement d'un téléphone mobile en utilisant les temps de parcours des signaux entre le mobile et les stations de base. La méthode adoptée ici est basée sur la minimisation des sommes des écarts de temps de parcours entre stations et mobile. Le problème est non-linéaire mais relativement simple. La figure ci-dessous illustre la convergence des itérations pour un mobile se déplaçant en ligne droite dans un maillage triangulaire du domaine où les stations de base se trouvent aux noeuds du maillage. La position initiale du mobile définit le triangle du départ. La fonctionnelle est donnée par :

$$J(x) = \sum_{i=1}^{3} (t_i(x) - t_i^d)^2, \quad t_i = d(x, x_i)/c,$$

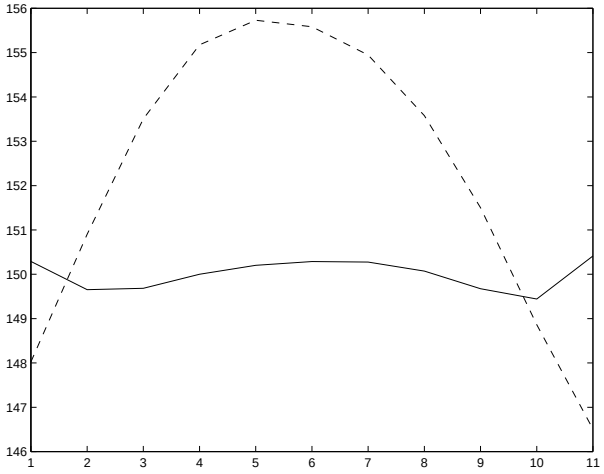


FIGURE 17.2 – Profil de température avant et après optimisation

où t_i est le temps de parcours du signal entre le mobile se trouvant en x et la base i située en x_i, d désigne la distance et c est la vitesse du signal. On cherche donc à trouver x de sorte que les temps de parcours soient les plus proches des trois temps t_i^d mesurés.

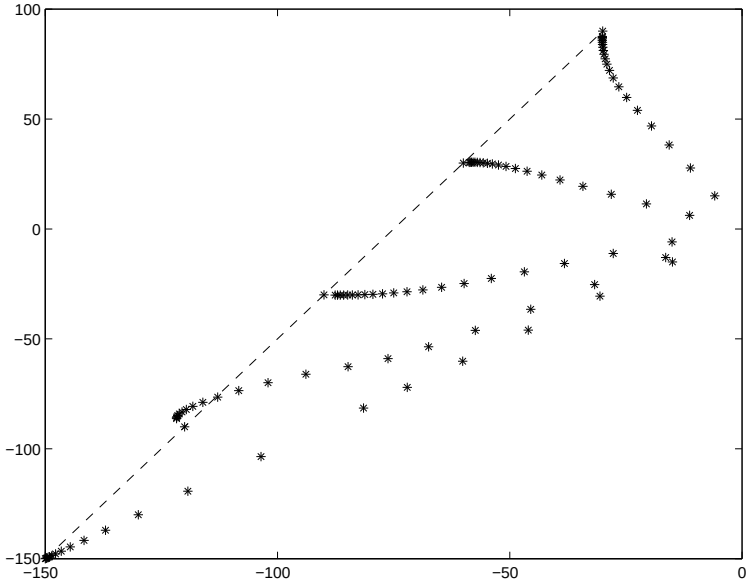


FIGURE 17.3 – Positionnement d'un téléphone mobile

17.5 Optimisation avec contraintes

17.5.1 Théorème général

Le problème classique consiste à trouver le minimum d'une fonctionnelle strictement convexe dans un sous-ensemble convexe K d'un Hilbert H (en dimension finie, un espace euclidien). Le sous-ensemble K est le domaine admissible. On a le théorème

Théorème 17.5.1 *Soit J une fonctionnelle donnée sur un espace de Hilbert H. Sous les hypothèses :*

1. *K est un sous-ensemble convexe fermé de H*
2. *J est strictement convexe et dérivable*
3. *Si K est non borné, on suppose de plus que J est coercive sur H : $(\lim_{\|x\|\to\infty} J(x) = +\infty)$*

alors J admet un minimum unique dans K.

17.5.2 Lagrangien

Considérons maintenant un problème d'optimisation dans un Hilbert H pour lequel le convexe K est défini par un ensemble de m contraintes égalités ou inégalités $E_i(x) \leq 0$ pour $i = 1, ..m$. Le problème d'optimisation s'écrit donc :

$$\begin{cases} \text{Trouver le minimum de la fonctionnelle } J \\ \text{sous les contraintes } E_i(x) \leq 0 \text{ pour } i = 1, ..m \end{cases} \tag{17.7}$$

Définition 17.5.1 (Lagrangien) *On appelle Lagrangien L associé au problème d'optimisation 17.7, la fonction*

$$L : (x, p) \in H \times I\!R^m_+ \longrightarrow L(x, p) = J(x) + \sum_{i=1}^{m} p_i \, E_i(x) = J(x) + p^T E(x) \tag{17.8}$$

Les nombres p_i s'appellent les multiplicateurs de Lagrange. Dans le cas de contraintes égalités $E_i(x) = 0$, le Lagrangien est défini sur $H \times I\!R^m$, les multiplicateurs de Lagrange ne sont pas obligatoirement ≥ 0.

Dans le cas d'un problème d'optimisation avec contraintes, la solution n'est plus, sauf cas particulier où le minimum global de J vérifie les contraintes, un zéro de la dérivée de J. Cette propriété, extrêmement commode, était à la base de la plupart des algorithmes de calcul de la solution des problèmes

d'optimisation sans contraintes. L'intérêt crucial du Lagrangien dans le cas de problèmes d'optimisation sous contraintes et de ramener la recherche du minimum à celle d'un point stationnaire du Lagrangien donc à nouveau à une recherche de zéro de dérivée.

Définition 17.5.2 (Point-selle) *Un couple $(x^*, p^*) \in H \times I\!R_+^m$ est point-selle du Lagrangien L si quels que soient $x \in H$, $p \in I\!R_+^m$ on a :*

$$L(x^*, p) \leq L(x^*, p^*) \leq L(x, p^*) \tag{17.9}$$

Autrement dit

$$\sup_{p \in I\!R_+^m} \inf_{x \in H} L(x, p) \leq L(x^*, p^*) \leq \inf_{x \in H} \sup_{p \in I\!R_+^m} L(x, p) \tag{17.10}$$

Le résultat fondamental est donné par le théorème suivant :

Théorème 17.5.2 (Kuhn et Tucker) *Soit J une fonctionnelle convexe, on suppose que les contraintes E_i sont également convexes et qu'il existe m nombres p_i^* tels que les relations suivantes, dites relations de Kuhn et Tucker soient vérifiées :*

$$\begin{cases} J'(x^*) + \displaystyle\sum_{i=1,m} p_i^* E_i'(x^*) = 0, \quad p_i^* \geq 0, \quad \forall i = 1, m \\ \text{et } \displaystyle\sum_{i=1,m} p_i^* E_i(x^*) = 0 \end{cases} \tag{17.11}$$

Alors x^* est un minimum de la fonctionnelle J vérifiant les contraintes E_i.

Le couple (x^*, p^*) est point-selle du Lagrangien

$$L(x, p) = J(x) + \sum_{i=1}^{m} p_i \, E_i(x) = J(x) + p^T E(x)$$

.

Algorithme d'Uzawa

La caractérisation du minimum x^* comme premier argument d'un point-selle du Lagrangien est la base de méthodes d'optimisation avec contraintes appelées méthodes *duales*. L'algorithme d'Uzawa utilise

$$\sup_{p^\in I\!R_+^m} \inf_{x \in H} L(x, p) \leq L(x^*, p^*) \leq \inf_{x \in H} \sup_{p \in I\!R_+^m} L(x, p^*) \tag{17.12}$$

en calculant itérativement :

À partir d'un élément initial $p^0 \in I\!R^m_+$,
la suite des x^k, p^k définie par :
x^k minimise $L(x, p^{k-1})$ sur H (optimisation sans contraintes)
p^k maximise $L(x^k, p)$ sur $I\!R^m_+$.

On prendra de façon pratique $p_i^k = max(0, p_i^{k-1} + \rho E_i(x^k))$ avec ρ choisi de manière à assurer la convergence de la méthode.

17.5.3 Exemple de problèmes d'optimisation avec contraintes

On veut trouver le rectangle, de côtés a et b, ayant à surface donnée $ab = S$, le périmètre $2(a+b)$ le plus petit . Le Lagrangien pour ce problème d'optimisation s'écrit : $L = 2(a+b) + p(ab - S)$. Et la recherche de solution telle que $(\partial L/\partial a) = (\partial L/\partial b) = (\partial L/\partial p) = 0$ aboutit à $a = b = \sqrt{S}$. Cet exemple, très simple, est cependant représentatif de nombreuses applications industrielles en conception de formes.

17.6 Un nouvel algorithme récursif de minimisation globale

Au chapitre 2, nous avons vu plusieurs algorithmes pour trouver les zéros d'une fonction. Ces algorithmes s'adaptent au contexte de l'optimisation.

En général, les méthodes de gradient ne peuvent converger que vers des minima locaux. Or, dans de nombreuses applications pratiques, on recherche l'optimum global d'une fonctionnelle.

Considérons le problème d'optimisation sous contraintes généralisé suivant (pour simplifier on considère des contraintes scalaires) :

$$\min_{x \in \mathcal{O}_{ad}} j(x), \quad E(x) = 0, \quad G(x) \leq 0. \tag{17.13}$$

Ce problème peut être reformulé selon :

$$\min_{x \in \mathcal{O}_{ad}, \eta > 0, \zeta > 0} J(x, \eta, \zeta), \quad J = j + \eta |E| + \zeta max_r(0, G). \tag{17.14}$$

où max_r est une régularisation du maximum (voir plus loin). Ainsi, en augmentant la taille de l'espace de contrôle, on peut transformer un problème sous contraintes en un problème sans contrainte. Cette opération introduit souvent des minima

locaux (voir plus loin pour un exemple simple) d'où l'importance de disposer d'une méthode peu coûteuse de minimisation globale.

L'idée fondamentale dans cet algorithme est qu'une méthode de gradient classique ne peut trouver le minimum global que si le point de départ appartient au bassin d'attraction du minimum global. Le but est donc de rechercher la bonne initialisation. Nous allons formuler le problème de minimisation (17.14) comme un problème à valeurs aux limites afin d'obtenir une méthode robuste pour la minimisation de fonctionnelles ayant de multiples minima.

On suppose que le problème admet une solution et que l'on connaît l'optimum global.

$$J_m = \min(J(x), \quad \{x/ \quad dJ/dx(x) = 0\}).$$

De plus on suppose $J \in C^1((\mathcal{O}_{ad}, \eta > 0, \zeta > 0), I\!R)$. Si l'on connaît une valeur initiale pour le paramètre d'optimisation x_0, une suite minimisante peut être construite selon :

$$x^{n+1} = x^n - \tau^n M^n \frac{dJ}{dx}^n, \quad x^0 = x_0, \quad \tau^n > 0, \tag{17.15}$$

- Si $M^n = Id$, cette itération correspond à la méthode de la plus grande pente.
- Si M^n est l'inverse du Hessien (d^2J/dx^2) de la fonctionnelle (resp. une approximation de l'inverse) on retrouve une méthode de Newton (resp. quasi-Newton).

L'itération (17.15) réduit J si :

$$J^{n+1} = J^n + \frac{dJ}{dx}^n (x^{n+1} - x^n) = J^n - \tau (\frac{dJ}{dx}^n)^T M^n (\frac{dJ}{dx}^n) \leq J^n.$$

Remarque 17.6.1 *Ceci est le cas si M^n est définie positive. Cette remarque explique pourquoi les méthodes de type quasi-Newton (comme BFGS), vues au chapitre 2, sont souvent plus efficaces qu'une méthode de Newton qui pourtant ne comporte pas d'approximations. En effet, dans ces méthodes, on utilise des approximations définies positives de l'inverse du Hessien.*

Pour simplifier l'exposé on considère $M^n = Id$. Le pas de descente τ^n peut être fixe ou optimisé par une recherche linéaire.

$$\tau^n = \text{argmin}_{\tau>0}(J(x^n - \tau \frac{dJ}{dx}^n)). \tag{17.16}$$

17.6.1 Problème différentiel du premier ordre

L'itération (17.15) est une discrétisation de $(\dot{x} = -\nabla J, \quad x(0) = x_0)$. Comme le problème a une solution, résoudre le problème de minimisation globale est équivalent à trouver $x(T)$ solution du problème surdéterminé suivant pour T fini :

$$\dot{x} = -\nabla J, \quad x(0) = x_0, \quad J(x((T)) = J_m. \tag{17.17}$$

En pratique, on choisit un nombre fini d'itérations N de (17.15) et $T = \sum_{n=0}^{N-1} \tau^n$ est une quantité positive et finie.

17.6.2 Suppression de la surdétermination

La surdétermination peut être levée en considérant $x_0 = v$ comme une nouvelle variable que l'on déterminera par une méthode de tir adaptée à la minimisation de $h(v) = J(x_v) - J_m$ où x_v est la solution obtenue après N itérations de (17.15) en partant de v.

L' algorithme est alors :

v_1, $v_2 = v_1 + \alpha$, ε, P donnés.

Evaluer $h(v_1) = J(x_{v_1}) - J_m$ et $h(v_2) = J(x_{v_2}) - J_m$,

Pour $p = 2, .., P$ **Faire**

Si $h(v_p) = J(x_{v_p}) - J_m > \varepsilon$ et $||v_p - v_{p-1}|| > \varepsilon$, **Alors**

$$v_{p+1} = v_p - h(v_p)\frac{(v_p - v_{p-1})}{h(v_p) - h(v_{p-1})}, \tag{17.18}$$

Trouver $x_{v_{p+1}}$ après N itérations de (17.15) en partant de v_{p+1} .

Sinon Fin

Fin Si

Fin Pour (p)

α est une variation admissible telle que $h(v_1) \neq h(v_2)$. (17.18) est la méthode de la sécante pour trouver un zéro de $h(v)$ le long de (v_1, v_2).

17.6.3 Interprétation géométrique en dimension un

Une condition nécessaire pour la convergence de la méthode de tir est la continuité de la fonctionnelle à minimiser $(h(v) = J(x_v) - J_m)$ par rapport à v. En une dimension, ceci est le cas si (17.15) est utilisé. En effet, pour tout $x_0 = v$ on tend alors vers un minimum local où $\dot{x} = 0$ (voir Fig. (17.4)). De plus, on tend vers le même minimum en partant des points du même bassin d'attraction. Ainsi, si N est assez grand, $h(v)$ est constante par morceaux avec les plateaux correspondants aux valeurs de $J(x_v) - J_m$ aux minima locaux. $h(v)$ est discontinue aux points

où la fonctionnelle atteint un maximum local. La figure 17.4, montre le graphe de $J(x) = x\sin(20x)\cos(x) + |x|^{0.1}$ (où $J_m = 0$) non-différentiable à l'origine et $h(v)$ construit pour illustration en utilisant un échantillonnage uniforme du domaine de variation du paramètre $[-1, 2]$. Cette construction aboutit donc à une convexification de J.

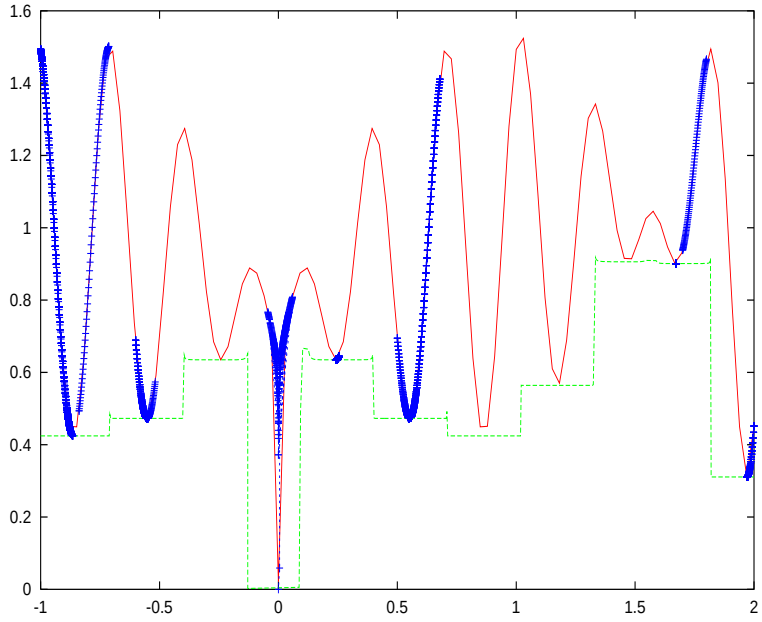


FIGURE 17.4 – Graphe de $J(x) = x\sin(20x)\cos(x) + |x|^{0.1}$ (courbe continue) et de $h(x_0) = J(x_{x_0})$ (pointillé). Les points visités par la méthode de minimisation sont reportés (croix). Le minimum global est atteint avec un algorithme à deux niveaux.

17.6.4 Méthode de tir multi-niveau récursive

Il arrive que l'algorithme ne converge pas vers le minimum global. Ceci est visible par exemple sur la figure (17.4) où les paliers sont tels que la méthode de tir ci-dessus ne suffira pas, la fonctionnelle générée n'étant pas convexe. L'idée est alors de recommencer la méthode de tir pour cette fonctionnelle constante par morceaux en ajoutant une boucle externe :

v_1^2, $v_2^2 = v_1^2 + \beta$, ε_1, K, donnés
Pour $k = 2, .., K$ **Faire**
Evaluer $h^2(v_1^2)$ et $h^2(v_2^2)$,
Si $h^2(v_k^2) > \varepsilon_1$ et $||v_k^2 - v_{k-1}^2|| > \varepsilon_1$, **Alors**

$v_1^1 = v_k^2$, $v_2^1 = v_k^2 + \alpha_k$, ε_2, P donnés,
Evaluer $h^1(v_1^1)$ et $h^1(v_2^1)$,
Pour $p = 2, .., P$ **Faire**

Si $h^1(v_p^1) > \varepsilon_2$ et $||v_p^1 - v_{p-1}^1|| > \varepsilon_2$, **Alors**

$$v_{p+1}^1 = v_p^1 - h^1(v_p^1)\frac{(v_p^1 - v_{p-1}^1)}{h^1(v_p) - h^1(v_{p-1}^1)},$$

Trouver $x_{v_{p+1}^1}$ après N itérations de (17.15) en partant de v_{p+1}^1,

Sinon (p) **Stop**

Fin Si (p)

Fin Pour (p)

$$h^2(v_k^2) = h^1(v_p^1) \quad \textbf{et} \quad v_k^2 = v_p^1,$$

$$v_{k+1}^2 = v_k^2 - h^2(v_k^2)\frac{v_k^2 - v_{k-1}^2}{h_k^2(v_k^2) - h_k^2(v_{k-1}^2)},$$

Sinon (k) **Fin**
Fin Si (k)
Fin Pour (k)

$h^i(v)$ indique les fonctionnelles générées successivement avec $h^1(v) = J(x_v) - J_m$, i est le niveau de la boucle externe. Les α_k et β sont linéairement indépendants. Cette construction peut être poursuivie de façon récursive en ajoutant d'autres boucles externes.

17.6.5 Compléments sur la prise en compte des contraintes

Dans les exemples ci-dessus, les contraintes d'inégalité de "boîte" de la forme $\underline{x} \leq x(t) \leq \overline{x}$ ont été prises en compte par projection. On applique

$$x(t) = \min(\max(x(t), \underline{x}), \overline{x}),$$

à chaque composante de x.

Pour les contraintes d'inégalité, on peut aussi introduire des fonctions barrières utilisant la distance à la frontière du domaine admissible :

$$\tilde{J}(x(t)) = J(x(t)) + \alpha(t)(\frac{1}{||x(t) - \overline{x}||} + \frac{1}{||x(t) - \underline{x}||}), \quad \alpha(t) > 0, \quad \lim_{t \to T} \alpha(t) = 0.$$

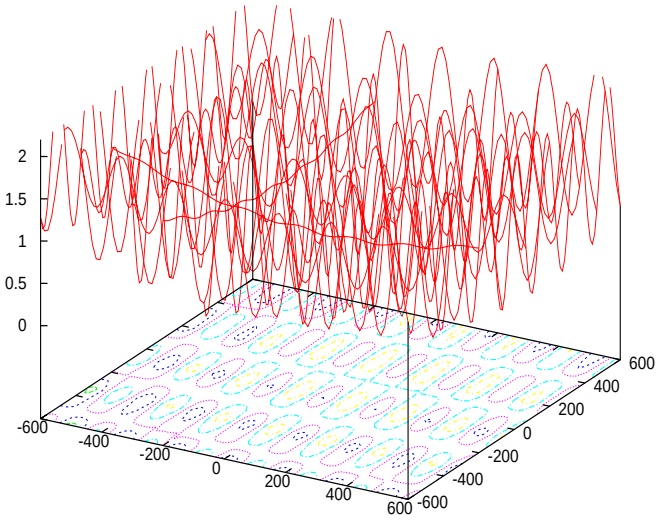


FIGURE 17.5 – Fonction de Griewank : graphe de $J(x) = 1 - \Pi_{i=1}^{I} \cos(x_i - 100) + 10^{-6} \sum_{i=1}^{I} (x_i - 100)^2, x \in [-600, 600]^I$ le long des deux premières dimensions. On considère les configurations de $I = 5, 10$ et 20.

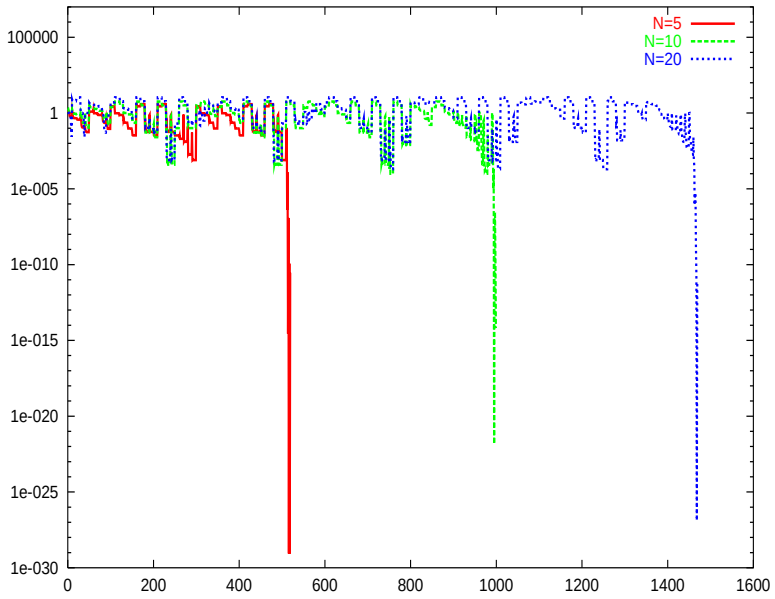


FIGURE 17.6 – Fonction de Griewank : évolution de J en fonction de l'accumulation des itérations d'optimisation pour les initialisations provenant de tir à trois niveaux. Plusieurs minima locaux ont été visités et le nombre total d'itérations croît de façon sous-linéaire par rapport à la dimension de l'espace des paramètres.

$\alpha(t) \sim 1/t$ est destiné à ramener la fonctionnelle sous sa forme initiale en fin d'optimisation.

De même, une contrainte d'inégalité de la forme $H(x(t), U(x(t))) \leq 0$ peut être prise en compte par une pénalisation barrière :

$$\tilde{J}(x(t)) = J(x(t)) + \gamma(H^+)^2, \quad \gamma > 0,$$

avec $H^+ = \max(0, H(x, U(x(t))))$. Pour avoir une fonctionnelle dérivable, on peut régulariser H^+. On choisit $H_\varepsilon^+ = 0$ si $H \leq -\varepsilon$, $H_\varepsilon^+ = 1/\varepsilon$ si $H \geq 0$ et $H_\varepsilon^+ = H(x(t), U(x(t)))/\varepsilon^2 + 1/\varepsilon$ si $-\varepsilon \leq H \leq 0$. ε est un seuil à définir (par exemple $\varepsilon = H(x_0, U(x_0))/2$). D'autres régularisations de la fonction max sont possibles. Par exemple, on peut aussi utiliser la fonction erf comme dans la figure (17.7).

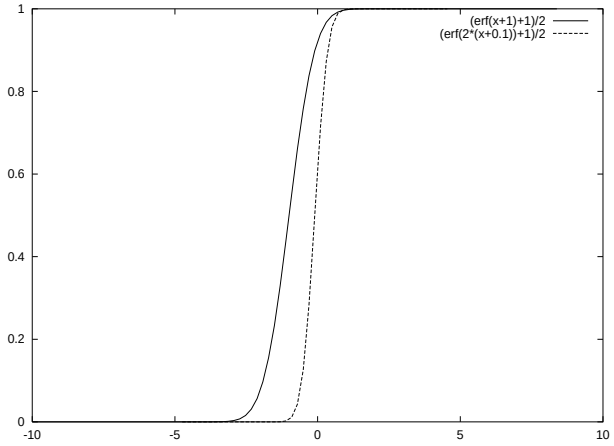


FIGURE 17.7 – Régularisation de la fonction max pour l'utilisation dans une fonction barrière lors de la prise en compte de contraintes d'inégalité dans la fonctionnelle.

Pour la prise en compte des contraintes d'inégalités, il existe aussi des méthodes de "points intérieurs" garantissant l'appartenance à l'espace admissible, de tous les points intermédiaires lors de l'optimisation. Ces algorithmes utilisent les gradients des contraintes. On demande que les angles entre la direction de descente d et les gradients de la fonctionnelle et des contraintes soient toujours aigus : $d.(dJ/dx) \geq 0$ et $d.(dH/dx) \geq 0$. Ceci n'est pas toujours possible. On affaiblira alors certaines contraintes en les relaxant comme avec la pénalisation. Une discussion plus large sur ces méthodes dépasse cependant le cadre de cet ouvrage.

17.7 Evaluation du gradient

La plupart des méthodes déterministes de minimisation sont basées sur l'utilisation du gradient de la fonctionnelle pour trouver des directions de descente. Nous nous intéressons ici à l'évaluation de ce gradient et nous essayons de mettre en évidence la complexité de cette opération.

Le gradient de J contient diverses contributions avec des complexités d'évaluation variables :

$$\frac{dJ}{dx} = \frac{\partial J}{\partial x} + \frac{\partial J}{\partial q}\frac{\partial q}{\partial x} + \frac{\partial J}{\partial U}\frac{\partial U}{\partial x}. \tag{17.19}$$

La dernière contribution est la plus coûteuse à évaluer car elle exige la linéarisation de l'équation d'état pour le calcul de $\frac{\partial U}{\partial x}$. Une façon simple de se rendre compte de la complexité consiste à utiliser les différences finies et à exprimer la complexité des évaluations à travers le nombre de fois où l'état doit être réévalué. Pour un calcul de la fonctionnelle, pour un point donné dans l'espace des contrôles, une évaluation de l'état U est nécessaire, tandis que le calcul de $\frac{\partial U}{\partial x}$ demandera N calculs supplémentaires de U. Ceci explique pourquoi la recherche d'une paramétrisation de faible dimension est très importante.

17.7.1 Différences finies

Comme nous l'avons dit, l'utilisation des différences finies est la méthode la plus utilisée car elle ne nécessite pas la connaissance explicite des opérateurs et convient donc aux outils boîte-noire.

$$\frac{dJ}{dx_i} \approx \frac{1}{\epsilon}[J(\vec{x} + \epsilon\vec{e_i}, U(\vec{x} + \epsilon\vec{e_i})) - J(\vec{x}, U(\vec{x}))].$$

Lors de l'utilisation de cette technique, les difficultés sont :

1. choix difficile pour l'incrément ϵ, surtout si les paramètres ont des dimensions différentes, ce qui implique souvent des increments différents pour chaque variable.
2. erreurs dans la soustraction de nombres réels voisins, comme celles rencontrées au chapitre 1 lors de la manipulation de nombres réels en discret.
3. complexité proportionnelle à la dimension de x.

Les deux premières difficultés peuvent être réduites en utilisant des différences finies centrées, en doublant la complexité :

$$\frac{dJ}{dx_i} \approx \frac{1}{2\epsilon}[J(\vec{x} + \epsilon\vec{e_i}, U(\vec{x} + \epsilon\vec{e_i})) - J(\vec{x} - \epsilon\vec{e_i}, U(\vec{x} - \epsilon\vec{e_i}))].$$

En pratique, on utilisera dans la mesure du possible cette approche pour identifier le bon incrément ϵ, puis on utilisera la première formule (voir figure (17.8).

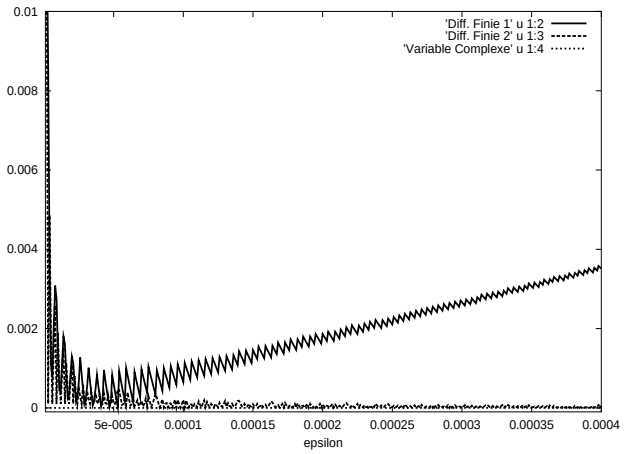


FIGURE 17.8 – Erreurs des différences finies décentrées et centrées en fonction de ϵ par rapport à la méthode des variables complexes pour l'exemple ci-dessous. Cette dernière n'est pas sensible au choix de l'incrément ϵ. Les différences finies centrées sont assez peu sensibles tandis que les différences finies décentrées le sont très fortement.

17.7.2 Travailler en variables complexes

Pour réduire encore l'importance du choix de l'incrément ainsi que l'erreur due à la soustraction de nombres réels voisins, on peut travailler en variables complexes. En effet, si J est une fonctionnelle réelle, on a :

$$J(x_i + i\epsilon, U(x_i + i\epsilon)) = J(x, U(x)) + i\epsilon J'_x - \frac{\epsilon^2}{2} J''_{xx} - i\frac{\epsilon^3}{6} J'''_{xxx} + o(\epsilon^3),$$

où $x_i + i\epsilon$ représente l'ajout au i_{eme} composante de x l'increment $\epsilon\sqrt{-1}$. D'où

$$\frac{dJ}{dx_i} = \frac{Im(J(x_i + i\epsilon, U(x_i + i\epsilon)))}{\epsilon} + o(\epsilon).$$

La difficulté ici est de passer en complexe et pour cela il faut disposer du programme source. Ce qui est rarement le cas en pratique. On peut cependant faire cela de façon partielle, notamment sur des routines utilisateurs que l'on est souvent amené à fournir aux codes industriels.

17.7.3 Linéarisation directe

De même, si l'on dispose de façon partielle du programme source on peut, pour réduire l'influence de ϵ, utiliser le calcul des variations. Notons $\delta x = \epsilon \vec{e^i}$ et

δU la variation de U (i.e. $\delta U = \delta x_i \partial_{x_i} U$). En linéarisant $E(x, U(x)) = 0$, nous avons

$$\frac{\partial E}{\partial U}\delta U = -\frac{\partial E}{\partial x}\delta x \approx -\frac{1}{\epsilon}[E(x + \epsilon\vec{e_i}, U(x)) - E(x, U(x))]. \tag{17.20}$$

En utilisant une méthode quasi-Newton pour la solution de l'équation d'état, nous avons :

$$\frac{\partial E^n}{\partial U}(U^{n+1} - U^n) = -E(x, U^n(x)). \tag{17.21}$$

Ainsi, on peut donc utiliser la même itération, mais en remplaçant N fois le second membre avec les colonnes de $\frac{\partial E}{\partial x}$. On obtient ainsi $\frac{\partial U}{\partial x}$. Les autres contributions au gradient étant facilement calculables.

17.7.4 Méthode de Lagrange

Considérons le Lagrangien $L = J + p^T E$ (voir définition 17.5.1) , et écrivons l'indépendance du Lagrangien par rapport à la variable dépendante U :

$$\frac{\partial L}{\partial U} = \frac{\partial J}{\partial U} + p^T \frac{\partial E}{\partial U} = 0, \quad \Leftrightarrow \quad p^T = -\frac{\partial J}{\partial U}(\frac{\partial E}{\partial U})^{-1}. \tag{17.22}$$

Après substitution, on a

$$\frac{dJ}{dx} = \frac{\partial L}{\partial x} = \frac{\partial J}{\partial x} + p^T \frac{\partial E}{\partial x}.$$

Ainsi, on obtient le gradient par une seule résolution du système (17.22) qui a la même complexité que (17.21). Ce qui veut dire que le coût de cette évaluation est indépendant de N.

On peut aussi voir la méthode de l'adjoint comme un regroupement différent des termes du gradient aboutissant à une complexité minimale. Considérons à nouveau (17.19) :

$$\frac{dJ}{dx} = \frac{\partial J}{\partial x} + \frac{\partial J}{\partial q}\frac{\partial q}{\partial x} + \frac{\partial J}{\partial U}\frac{\partial U}{\partial x}.$$

Nous avons vu que la majeure partie du coût dans cette évaluation provient du terme $\partial U/\partial x$, qu'il faut de plus stocker avant d'en faire le produit avec $\partial J/\partial U$. On peut cependant réécrire formellement le dernier terme comme :

$$\frac{\partial J}{\partial U}\frac{\partial U}{\partial x} = \frac{\partial J}{\partial U}((\frac{\partial E}{\partial U})^{-1}\frac{\partial E}{\partial x}) \quad \text{car} \quad \frac{\partial E}{\partial U}\frac{\partial U}{\partial x} + \frac{\partial E}{\partial x} = 0,$$

Mais on peut regrouper différemment les termes, en déplaçant les parenthèses :

$$\frac{\partial J}{\partial U}\frac{\partial U}{\partial x} = -(\frac{\partial J}{\partial U}(\frac{\partial E}{\partial U})^{-1})\frac{\partial E}{\partial x}.$$

On voit réapparaître la variable intermédiaire p solution de

$$p^T(\frac{\partial E}{\partial U}) = -\frac{\partial J}{\partial U},$$

qui a la complexité d'une évaluation de l'état et qui a la même taille que celui-ci.

17.7.5 Différentiation automatique

Soit f une fonction composée de la forme :

$$x \in R^p \to y = h(x) \in R^n \to z = g(y) \in R^n \to u = f(z) \in R^q.$$

$$u' = f'(z)g'(y)h'(x), \tag{17.23}$$

où $f' \in R^{q\times n}$, $g' \in R^{n\times n}$ et $h' \in R^{n\times p}$. Pour faire cette évaluation, il faut stocker $M = g'(y)h'(x) \in R^{p\times n}$ puis calculer $u' = f'(z)M$. Après transposition, on a

$$u'^T = h'^T(x)g'^T(y)f'^T(z). \tag{17.24}$$

Le stockage nécessaire est maintenant $M = g'^T(y)f'^T(z) \in R^{n\times q}$. La complexité dépend alors de la dimension des espaces de départ et d'arrivée. Ces choix s'appellent modes direct et inverse de la différentiation automatique. Ceci est aussi une autre façon de présenter les complexités des deux approches par linéarisation directe et méthode adjointe que nous avons vu ci-dessus. D'un point de vue pratique, ils existent deux approches majeures pour la mise en oeuvre de la différentiation automatique : la pré-compilation et la surcharge d'opérateurs. Dans le premier cas, l'outil de différentiation automatique produit un code pour la dérivée en partant du code direct. Dans le second cas, par contre, un nouveau code n'est pas généré pour la dérivée. On introduit plutôt de nouvelles classes de variables, incluant la classe initiale, mais aussi ses dérivées. On redéfinit les opérations natives pour prendre en compte les manipulations des dérivées. L'exécution du code direct avec les variables dans ces nouvelles classes permettra alors d'accéder aux dérivées directement. Cette approche nécessite donc un langage objet de type C++, tandis que la pré-compilation est applicable aux langages de bas-niveaux.

17.7.6 Un exemple $R \to R^2 \to R$

Soit $f = x^2 + 3x$ ($f' = 2x + 3$),

```
y_1=x
y_2=x**2+2*y_1
f  =y_1+y_2
```

calculons df/dx.
Mode direct Une dérivation ligne à ligne en fonction de x nous donne :

$$\frac{dy_1}{dx} = 1, \quad \frac{dy_2}{dx} = 2x + 2\frac{dy_1}{dx},$$

$$\frac{df}{dx} = \frac{dy_1}{dx} + \frac{dy_2}{dx} = 1 + 2x + 2.$$

On doit stocker tous les calculs intermédiaires.
Mode inverse Considérons le Lagrangien du programme

$$L = y_1 + y_2 + p_1(y_1 - x) + p_2(y_2 - x^2 - 2y_1).$$

$$\frac{df}{dx} = \frac{\partial L}{\partial x} = -p_1 - 2p_2 x,$$
$$\frac{\partial L}{\partial y_1} = 1 + p_1 - 2p_2 = 0,$$
$$\frac{\partial L}{\partial y_2} = 1 + p_2 = 0.$$

Ce système triangulaire est résolu du bas vers le haut (remontée). Donc x n'intervient qu'à la dernière opération. Ce qui est intéressant si x est un vecteur de grande taille.

17.7.7 Illustration de la différentiation

Pour illustrer notre propos, nous donnons un exemple de calcul de dérivée en utilisant la différentiation automatique en mode directe, les différences finies et la méthodes des variables complexes pour la fonction $f(x) = sin(x^2 + 3x)$ au voisinage de $x = 1$.

```
Function test_gradient

Real x,y,f;
Complex xc,yc,fc,epsc;

!
!  calcul du gradient analytique
!  pour f=sin(x**2+3*x) au voisinage de x=1
!
!  definition de la fonctionnelle
!
```

```
x=1;
y=x**2+3*x;
f=sin(y);
!
!  calcul du gradient par differentiation automatique
!                   en mode direct
dy=2*x+3;
df=cos(y)*dy;
!
!  calcul du gradient par differences finies d'ordre 1
!
eps=0.0001;
x1=1+eps;
y1=x1**2+3*x1;
f1=sin(y1);
dffd=(f1-f)/eps;
!
!  calcul du gradient par differences finies d'ordre 2
!
x2=1.-eps;
y2=x2**2+3.*x2;
f2=sin(y2);
dffd2=(f1-f2)/(2.*eps);
!
!  calcul du gradient par variables complexes
!
epsc=cmplx(0.,eps) ! calcul CVM
xc=(1.,0)+epsc;
yc=xc**2+3.*xc;
fc=sin(yc);
fcvm=real(fc);
dfcvm=imag(fc)/eps;
!
End
```

17.8 Gradient incomplet

Nous avons vu que le coût de la linéarisation provient surtout de celle de l'équation d'état. Nous verrons aux travers d'exemples simples comment réduire ce coût en introduisant des simplifications lors de l'évaluation des gradients. Dans

une grande majorité de cas, la fonctionnelle peut avoir la forme :

$$J(x) = \int_\Gamma f(x)g(u)d\gamma,$$

où Γ désigne le support de la paramétrisation. Pour ces configurations, l'approximation qui consistera à négliger la contribution de l'état dans la linéarisation, aboutit cependant à une bonne estimation du gradient obtenu en (17.19) :

$$\frac{dJ}{dx} \sim \frac{\partial J}{\partial x} + \frac{\partial J}{\partial q}\frac{\partial q}{\partial x}.$$

Plus généralement, pour une fonctionnelle de la forme :

$$J(x) = A(x) \ \ B_{S(x)}(u),$$

où $S(x)$ est le support du contrôle x. Le gradient de J par rapport à x s'écrit :

$$\frac{dJ}{dx} = \frac{\partial A}{\partial x}B + A\frac{\partial B}{\partial x}. \tag{17.25}$$

Le gradient incomplet ne retient que le premier terme de (17.25) :

$$\frac{dJ}{dx} \sim \frac{\partial A}{\partial x}B.$$

Considérons la fonctionnelle $J = a\frac{\partial u}{\partial x}(a)$ où u est solution d'une équation d'advection-diffusion :

$$\frac{\partial u}{\partial x} - Pe^{-1} \ \ \frac{\partial^2 u}{\partial x^2} = 0, \quad \text{on} \ \ ]a,1[, \quad u(a) = 0, \quad u(1) = 1,$$

qui a comme solution :

$$u(x) \ = \ \frac{\exp(Pe^{-1}\ a) - \exp(Pe^{-1}\ x)}{\exp(Pe^{-1}\ a) - \exp(Pe^{-1})} \Rightarrow \frac{\partial u}{\partial x}(a) = \frac{-Pe^{-1}\exp(Pe^{-1}\ a)}{\exp(Pe^{-1}\ a) - \exp(Pe^{-1})} \tag{17.26}$$

Calculons

$$\frac{\partial J}{\partial a}(a) = \frac{\partial u}{\partial x}(a) + a\frac{\partial^2 u}{\partial a \partial x}(a)$$

on obtient :

$$\frac{\partial J}{\partial a}(a) = \frac{\partial u}{\partial x}\ (a)\left(1 + a\frac{Pe^{-1}\ \exp(Pe^{-1}\ a)}{\exp(Pe^{-1}\ a) - \exp(Pe^{-1})}\right)$$

$$= \frac{\partial u}{\partial x}(a)\left(1 - a\frac{\partial u}{\partial x}\ (a)\right).$$

Si $Pe >> 1$, $a\frac{\partial u}{\partial x}(a) << 1$. Ainsi, la contribution de la linéarisation de l'état peut être négligée, tout en conservant le bon signe et une bonne estimation du gradient.

On peut vérifier aussi ce concept pour l'équation non-linéaire de Burgers où le paramètre de contrôle a porte sur la frontière gauche :

$$\frac{\partial u}{\partial t} + 0.5\frac{\partial u^2}{\partial x} = 0.3xu, \quad \text{sur} \quad]a, 1[, \quad u(a) = 1, \quad u(1) = -0.8. \tag{17.27}$$

La fonctionnelle est donnée par $J(a) = a\frac{\partial u}{\partial x}(a)$ et son gradient par

$$\frac{\partial J}{\partial a}(a) = \frac{\partial u}{\partial x}(a) + a\frac{\partial^2 u}{\partial a \partial x}(a).$$

Nous sommes dans le domaine d'application des gradients incomplets. Connaissant la solution de l'équation de Burgers (voir plus loin), nous avons $\frac{\partial u}{\partial x}(a) = 0.3\,a$ dans les régions où elle est régulière. Ainsi le gradient exact ($\frac{\partial J}{\partial a}(a) = 0.3a + a0.3$) peut être comparé au gradient incomplet ($0.3a$). On constate que le signe est correct et qu'il manque un facteur 2. Ce qui n'est pas important lors d'une utilisation avec une méthode de type gradient avec un pas optimal par exemple. Ceci est vrai pour toute fonctionnelle de la forme $J = f(a)g(u, \frac{\partial u}{\partial x})$.

17.8.1 Redéfinition des fonctionnelles

Le but est de se ramener dans le domaine de validité des gradients incomplets, pour les fonctionnelles qui ne sont pas sous la forme donnée ci-dessus. L'exemple ci-dessous montre comment redéfinir la fonctionnelle dans ce but. On considère l'équation de Poiseuille qui décrit l'écoulement dans un canal soumis à un gradient de pression le long de l'axe du canal ($\frac{\partial p}{\partial x}$). Les parois du canal sont en $y = +/-a$:

$$\frac{\partial^2 u}{\partial y^2} = \frac{1}{\nu}\frac{\partial u}{\partial x}, \quad u(-a) = u(a) = 0. \tag{17.28}$$

Cette équation a une solution simple : $u(a, y) = \frac{1}{2\nu}\frac{\partial p}{\partial x}(y^2 - a^2)$. Intéressons nous au débit et à son gradient par rapport à l'épaisseur du canal. Le débit est donné par $J_1(a) = \int_{-a}^{a} u(a, y)dy = \frac{-2a^3}{3\nu}\frac{\partial p}{\partial x}$. Son gradient est

$$\frac{dJ_1}{da} = \int_{-a}^{a} \partial_a U(a, y)dy = \frac{-2a^2}{\nu}\frac{\partial p}{\partial x}.$$

Dans ce cas, la fonctionnelle n'est pas dans le domaine de validité des gradients incomplets. Considérons maintenant plutôt la fonctionnelle suivante obtenue en multipliant le débit par a :

$$J_2(a) = aJ_1(a).$$

Fonctionnelle dont le gradient est :

$$\frac{dJ_2}{da} = J_1(a) + a\frac{dJ_1}{da}.$$

Ici, la première contribution est le gradient incomplet et la deuxième, la contribution provenant de la linéarisation de l'état :

$$\frac{dJ_2}{da} = \frac{-a^3}{\nu}(\frac{2}{3} + 2)\frac{\partial p}{\partial x}.$$

On remarque que les deux contributions ont le même signe. Notons $\widetilde{\frac{dJ_2}{da}}$ le gradient incomplet. On peut accéder à dJ_1/da par :

$$\frac{dJ_1}{da} = \frac{1}{a}\widetilde{\frac{dJ_2}{da}} = \frac{-2a^2}{\nu}\frac{\partial p}{\partial x},$$

qui a le bon signe et diffère d'un facteur 1/3 du gradient exact. Cette redéfinition des fonctionnelles est très utile lors du traitement d'applications complexes.

17.8.2 Utilisation des modèles à complexité réduite

Le but est de remplacer l'équation d'état initiale par une version ayant une complexité réduite. Cette nouvelle équation sera utilisée uniquement lors de la linéarisation. Revenons à notre boucle de simulation :

$$x \to q(x) \to U(q(x)) \to J(x, q(x), U(q(x))),$$

et remplaçons l'équation d'état par :

$$x \to q(x) \to \tilde{U}(q(x))(\frac{U}{\tilde{U}}),$$

où $\tilde{U} \sim U$ est obtenu par un modèle réduit, moins coûteux mais ne pouvant cependant donner la bonne réponse pour la simulation en ce qui concerne l'état. On exprime alors la contribution de l'état dans le gradient à travers ce modèle :

$$\frac{dJ}{dx} \sim \frac{\partial J(U)}{\partial x} + \frac{\partial J(U)}{\partial q}\frac{\partial q}{\partial x} + \frac{\partial J(U)}{\partial U}\frac{\partial \tilde{U}}{\partial q}\frac{\partial q}{\partial x}\frac{U}{\tilde{U}}.$$

Des modèles à complexité réduite sont souvent utilisés en pratique et sont issus de la connaissance du comportement de la solution du modèle complet dans des contextes particuliers, permettant soit une réduction de dimension du problème, soit une simplification de celui-ci en négligeant certains termes. Nous en avons vu quelques exemples au chapitre sur le couplage des modèles.

A titre d'exemple, considérons l'équation de Stokes dans un domaine bidimensionnel et intéressons-nous à la réduction de cette équation dans la direction normale à une paroi (désignée par $y = y_w$, y étant la distance normale à cette paroi) permettant le calcul de la composante tangentielle $u = \vec{U}.\vec{t}$ de la vitesse $(\vec{U})$ parallèle à la paroi. Ceci est similaire au développement des conditions aux limites équivalentes rencontré au chapitre 3.

$$-\nu\Delta\vec{U} + \nabla p = 0 \quad \rightarrow \quad \frac{\partial^2 u}{\partial y^2} = \frac{1}{\nu}\frac{\partial p}{\partial x} = \text{constante}, u(y_w) = 0, u(\delta) = u_\delta,$$

où y désigne la direction normale à la paroi et δ est l'épaisseur sur laquelle le modèle réduit 1D ci-dessus est supposé valide. Ceci peut être aussi vu comme l'interface de couplage entre les approches 1D et 2D. u_δ est le résultat du modèle 2D en δ. Constatant, ou supposant dans le cadre de la modélisation, que les variations en x sont négligeables par rapport à celle en y, on considère localement $\frac{\partial p}{\partial x}$ comme une constante en y au voisinage de la paroi. Ceci permet alors une intégration explicite dans la direction normale et fournit un profil parabolique pour u. On obtient alors le gradient de la composante tangentielle de la vitesse par rapport aux variations normales à la paroi (y_w) : $\partial u/\partial y_w$.

Une autre possibilité est l'utilisation des méthodes modales ou la décomposition orthogonale propre, que nous avons vue au chapitre 11. Dans ces méthodes, on remplace la solution d'un modèle complexe par celle de d'un modèle réduit pour les coefficients de la solution dans une base partielle de l'espace des solutions. Ce système réduit pourra alors être utilisé pour le calcul du gradient de la fonctionnelle par rapport à l'état.

17.8.3 Différences finies et gradients incomplets

La simplification due au gradient incomplet permet l'utilisation des approches élémentaires comme les différences finies ou la méthodes des variables complexes. Bien entendu, la complexité reste proportionnelle au nombre de paramètres de contrôle, mais chaque évaluation est très peu coûteuse :

$$\overline{J(x,U)} = J(x) = (x,U) \rightarrow q(x) \rightarrow J(x,q(x),U).$$

$$\frac{dJ}{dx}(i) = \frac{J(x_i+\varepsilon) - J(x)}{\varepsilon} \sim \frac{\overline{J(x_i+\varepsilon,U)} - \overline{J(x,U)}}{\varepsilon}.$$

De plus, cette approche permet l'utilisation des outils boîte-noire pour l'évaluation des états car l'équation d'état n'a pas été linéarisée. On utilisera cette évaluation ci-dessous pour la solution d'un problème inverse en reconstruction d'état.

17.9 Les problèmes inverses

Les problèmes inverses présentent un grand intérêt applicatif. Ils englobent les problèmes d'optimisation de formes et de contrôle d'état, mais concerne aussi plusieurs autres applications, comme par exemple :

— Les problèmes de conception inverse où, connaissant l'état souhaité, on recherche la distribution des paramètres de contrôle permettant de l'obtenir. Ceci est le cas par exemple en aéronautique, où connaissant la distribution de pression souhaitée sur les ailes, on recherche leur forme. En conception de pièces mécaniques, on pourra chercher celle qui permettra la réalisation d'une carte d'efforts donnée. Ce type de problème existe dans tous les domaines des sciences de l'ingénieur.

— L'assimilation des données par les modèles mathématiques et numériques est un autre exemple de l'utilisation des problèmes inverses. Dans ce cas, connaissant le comportement, probablement partiel, de l'état pour une configuration donnée, et connaissant le modèle mathématique sous-jacent, on souhaitera identifier les coefficients du modèle. Ceci ne peut en général pas être effectué de façon exacte.

— L'identification des sources en utilisant la connaissance partielle de l'équation d'état est un autre domaine d'application des problèmes inverses. Ainsi, ayant observé le comportement de l'état, connaissant partiellement le modèle mathématique sous-jacent et les coefficients associés, on pourra compléter les parties manquantes dans le modèle. Ceci est différent de la problématique précédente où le modèle mathématique était connu et où seuls les coefficients étaient manquants. Une application typique concerne l'identification des sources thermo-mécanique, en utilisant les observations du comportement thermique et mécanique d'un matériau donné. L'observation est en général effectuée via une caméra infra-rouge. Les équations d'états sont l'équation de la chaleur et l'équation d'élasto-dynamique des milieux continus.

— Le couplage de modèles incompatibles par la solution d'un problème inverse est une autre application possible. L'incompatibilité provenant souvent des différences d'échelles de temps et d'espace sur lesquelles les modèles sont basés. En couplage de modèles (voir chapitre 13), il devient alors impossible d'utiliser directement la solution d'un des modèles comme conditions aux limites pour l'autre. On présente au chapitre 19 un exemple d'incompatibilité qui apparaît lors de l'utilisation du filtrage pour la solution des EDP et nous verrons comment formuler une condition de raccord entre la solution de ces modèles à travers la solution d'un problème de minimisation.

Il est à noter que les problèmes précédents concernent la recherche d'un contrôle x minimisant la fonctionnelle $J(x) = \|u(x) - u_{des}\|$, où u est l'état, solution de l'équation d'état $E(u(x)) = 0$ et u_{des} l'état désiré. Bien entendu, x, u et E sont différents dans chaque cas.

Ci-dessous, nous présentons deux problèmes inverses pour les problématiques 1 et 3 ci-dessus.

17.9.1 Reconstruction d'état

Nous allons étudier le problème de reconstruction de la solution de l'équation de Burgers en utilisant un contrôle au second membre. Nous présentons une méthode d'approximation de gradient pour réduire la complexité de l'évaluation de ces derniers. Par ailleurs, nous montrons qu'une redéfinition de la fonctionnelle peut améliorer les caractéristiques du problème d'optimisation.

On considère l'équation de Burgers (17.27) rencontrée plus haut. Nous allons contrôler la position du choc à travers un contrôle placé au second membre de l'équation :

$$u_t + 0.5(u^2)_x = F(x)u, \quad u(t,-1) = 1, u(t,1) = -0.8, \tag{17.29}$$

$F(x) \in \mathcal{O}_{ad}$ est le paramètre de contrôle appartenant à l'espace fonctionnel des contrôles admissibles. Le choix de $\mathcal{O}_{ad}$ s'appelle aussi paramétrisation du problème comme pour les problèmes d'optimisation de formes. On peut choisir une discrétisation uniforme de $[-1, 1]$ en n intervalles avec $n + 1$ points de discrétisation et alors F est un élément de $\mathbb{R}^{n+1}$. Une autre possibilité est, par exemple, le choix de l'espace des polynômes de degré inférieur ou égal à m comme espace de contrôle : $\mathcal{O}_{ad} = \mathcal{P}_m$. Pour simplifier, supposons que les gradients sont évalués par différences finies.

La fonctionnelle est donnée par :

$$J_\alpha(F) = \int_{-1}^{1} |u_F(x) - u_{des}(x)|^\alpha \, dx, \quad \text{avec} \quad \alpha = 2,$$

où u_{des} est la solution cible obtenue pour $F = 0.3x$. La solution est quadratique par morceaux avec un saut en x_s :

$$u(x) = 0.15x^2 + 0.85 \quad \text{pour } x < x_s,$$

$$u(x) = 0.15x^2 - 0.95 \quad \text{pour } x > x_s,$$

la position du choc x_s est telle que :

$$u_s^- = -u_s^+ \quad \text{ce qui implique} \quad x_s = -\sqrt{1/3}.$$

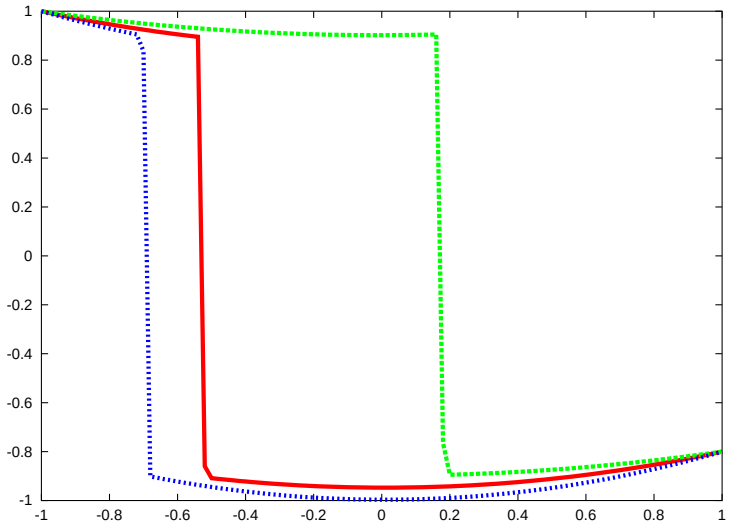


FIGURE 17.9 – Solution de l'équation de Burgers pour (de gauche à droite) $F = 0.2\,x$, $F = 0.3\,x$ et $F = 0.4\,x$, sur un maillage régulier de 100 points.

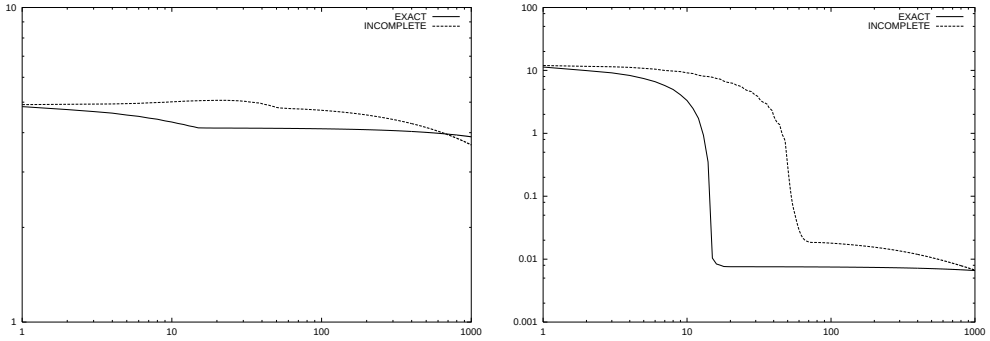


FIGURE 17.10 – Historiques de convergence pour $||F - F_{tar}||$ (gauche) et J_2 (droite) avec le gradient complet (courbe continue) et incomplet (courbe discontinue). Les résultats sont assez similaires, le gradient incomplet semble même être plus robuste.

Il est possible de trouver explicitement le contrôle connaissant la solution $u_{des}(t, x)$. En effet, d'après l'équation d'état, dans les régions où la solution est régulière, on a :

$$F(t, x) = \partial_t(\log(u_{des})) + \partial_x(u_{des}).$$

De plus, comme le contrôle recherché n'est pas fonction du temps, et connaissant $u_{des}(x)$, le contrôle est obtenu comme $\partial_x(u_{des}(t, x)) = 0.3x$. Cependant, cette approche n'est pas générique et on préfère utiliser une technique de minimisation,

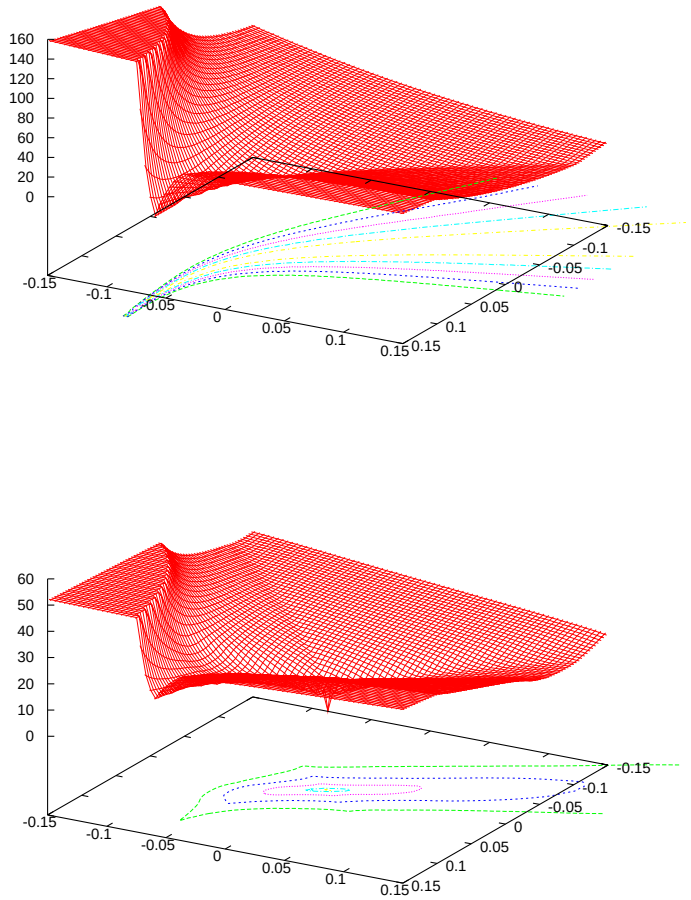


FIGURE 17.11 – Surface et iso-contours de $(a,b) \mapsto J(F_{tar} + ax + b)$ (haut) et de $(a,b) \mapsto J_{0.3}(F_{tar} + ax + b)$ (bas). On remarque que la pente est plus forte et que les iso-contours sont concentriques au voisinage du minimum (Calculs réalisés par A. Cabot à Montpellier).

comme celle de la plus grande pente, pour générer une suite minimisante dans $\mathcal{O}_{ad}$.

Bien que le contexte de ce problème ne corresponde pas au cadre d'applications des gradients incomplets, il est cependant intéressant de voir si on peut tirer avantage d'une résolution incomplète des états intermédiaires nécessaires pour l'évaluation des gradients. Supposant que l'équation de Burgers est résolue par un des schémas explicites stabilisés présentés aux chapitres 2 et 12, on considérera

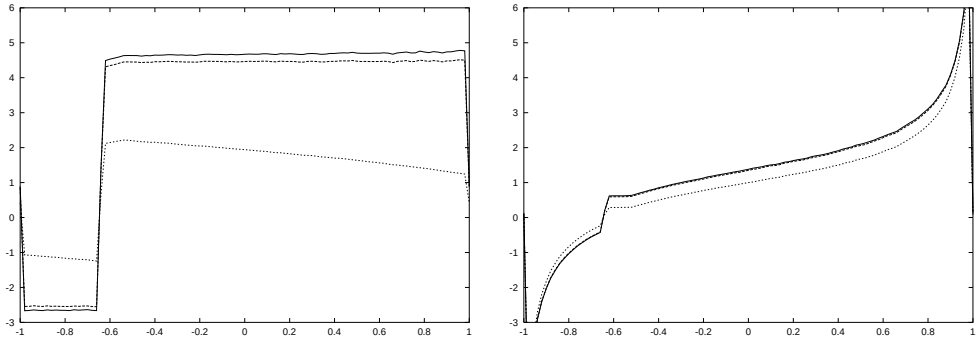


FIGURE 17.12 – Gradients incomplets (courbes pointillées) obtenus avec deux et dix fois moins d'itérations en temps que pour le calcul complet, à comparer avec le gradient exact (courbe continue) pour J_2 (gauche) et pour $J_{0.3}$ (droite). On constate qu'en plus d'une convexification du problème, $J_{0.3}$ est mieux adaptée pour l'utilisation des gradients incomplets.

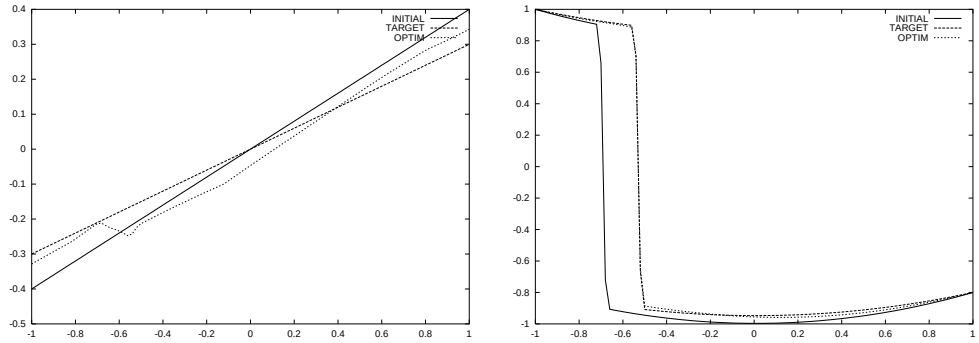


FIGURE 17.13 – (Gauche) : $F_{ini} = 0.4\,x$, $F_{tar} = 0.3\,x$ et F_{opt}. (Droite) : les états correspondants u_{ini}, u_{des} et u_{opt} pour $\mathcal{O}_{ad} = I\!R^n$ avec $n = 100$. Le contrôle initial est défini par une discrétisation de $F_{ini}(x) = 0.4x$. Cette difficulté reste posée avec les méthodes de minimisation plus sophistiquées.

alors pour l'évaluation de ces états un nombre réduit d'itérations en temps, bien inférieur à celui nécessaire pour une convergence complète.

Dans tous les cas, on constate que la convergence n'est pas satisfaisante. Pour comprendre d'ou vient ce défaut de convergence il est utile d'avoir une idée géométrique de J au voisinage du minimum. On se restreint alors à un espace $\mathcal{O}_{ad}$ de dimension inférieure permettant un échantillonnage de l'espace. Considérons $\mathcal{O}_{ad} = \mathcal{P}_1$, l'espace de dimension 2 des fonctions affines sur $[-1, 1]$. Bien entendu, le minimisant global appartient à cet espace. On présente en 17.11 la fonctionnelle

$$\left\{ \begin{array}{rl} [-0.15, 0.15] \times [-0.15, 0.15] & \to I\!R \\ (a, b) & \mapsto J(F_{tar} + ax + b). \end{array} \right.$$

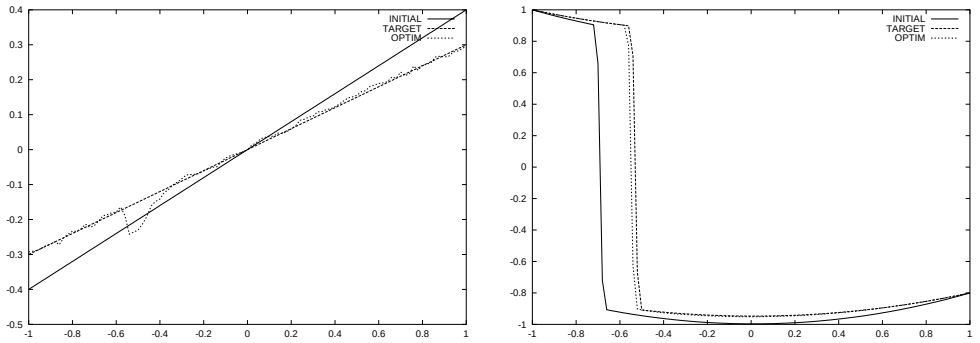


FIGURE 17.14 – (gauche) : contrôle initial $F_{ini} = 0.4\,x$, cible $F_{tar} = 0.3\,x$ et après minimisation F_{opt} pour la nouvelle fonctionnelle. (Droite) : les états correspondants u_{ini}, u_{des} et u_{opt}.

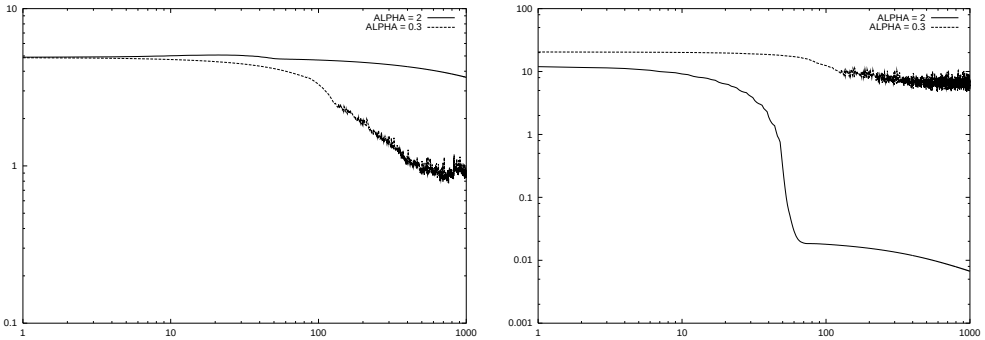


FIGURE 17.15 – Convergence de $||F - F_{tar}||$ (gauche) et J et J_α (droite).

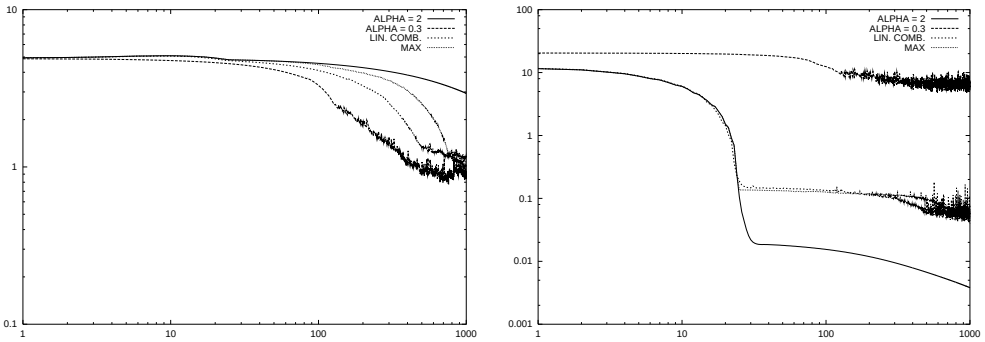


FIGURE 17.16 – Convergence de $||F - F_{tar}||$ (gauche) et pour différentes fonctionnelles : J_2, $J_{0.3}$, $(0.01\,J_{0.3} + 0.99\,J_2)$ et $\max(J_2, 0.01\,J_{0.3})$ (droite).

La figure 17.11 montre que les lignes de niveaux ne sont pas concentriques autour de $a = b = 0$ et J est plate dans certaines directions. Un pré-conditionnement linéaire ne changera rien à la géométrie du problème et n'améliorera pas la

convergence. Par contre, on verra qu'une redéfinition de la fonctionnelle peut grandement changer les caractéristiques du problème de minimisation. Ainsi, on considère α comme un paramètre de contrôle supplémentaire. Le but est alors de définir le meilleur α, rendant la fonctionnelle aussi convexe que possible (voir Figure 17.11 bas), pour un espace de contrôle de faible dimension, par dichotomie par exemple, et ensuite effectuer la minimisation pour J_α dans l'espace de contrôle initial.

Revenons à $\mathcal{O}_{ad} = I\!R^{100}$, et étudions le cas $\alpha = 0.3$. On constate alors une bonne convergence en contrôle, mais une moins bonne convergence en état que dans le cas $\alpha = 2$.

Il est donc clair que $J_{\alpha=2}$ et $J_{\alpha=0.3}$ se comportent différemment. Une bonne minimisation du problème demande la minimisation simultanée des deux fonctionnelles. Il faut donc envisager un problème multi-critère par une approche min-max par exemple :

$$\min(\max(J_{\alpha=2}, J_{\alpha=0.3})). \tag{17.30}$$

D'autres approches sont possibles, comme par exemple la minimisation d'une combinaison linéaire de $J_{\alpha=2}$ et $J_{\alpha=0.3}$:

$$\min(a\, J_{\alpha=2} + b\, J_{\alpha=0.3}),$$

où a et b sont des coefficients positifs. En pratique, on prend

$$a = \frac{(1-\epsilon)}{J^0_{\alpha=2}}, \quad b = \frac{\epsilon}{J^0_{\alpha=0.3}},$$

avec $0 < \epsilon << 1$, $J^0_{\alpha=2}$ et $J^0_{\alpha=0.3}$ désignent les valeurs initiales.

Les deux approches ci-dessus améliorent la convergence quels que soient les choix de a et b (figure 17.16).

Remarque 17.9.1 *Ceci est différent d'une minimisation en norme L^p avec p croissant, pour éviter la non-différentiabilité de la norme L^∞.*

17.9.2 Reconstruction de sources par l'équation d'état et dérivation numérique

Dans cette section, nous présentons certains aspects du traitement de signaux et images mettant en jeu la discrétisation des EDP. Ces problèmes relèvent aussi du domaine des problèmes inverses.

Le but de la dérivation numérique est la reconstruction partielle de l'équation d'état par la connaissance de l'état. Cette reconstruction permettra d'identifier

les sources par exemple. Ainsi, connaissant u et l'équation d'état suivante (où L contient des dérivées partielles en espace) :

$$u_t + L(u) = f, \quad u(0) = u_0,$$

on peut identifier f par assemblage, après discrétisation :

$$f_h = (u_h)_t + L_h(u_h).$$

Cependant, même avec une discrétisation conforme cette tâche n'est pas aisée car le champ u observé n'est pas nécessairement compatible avec la discrétisation.

Par exemple, lorsque les observations $\hat{u}$ sont entachées d'incertitudes (on dira qu'elles sont bruitées), on ne pourra pas reconstruire correctement la source. Il faut débruiter les observations, en les projetant sur l'espace des solutions de l'équation discrète $\hat{u}_h = P_h(\hat{u})$, avant de procéder à la reconstruction : $f_h = (\hat{u}_h)_t + L_h(\hat{u}_h)$.

La définition d'un filtre pertinent n'est pas une chose aisée. On peut utiliser les filtres présentés au chapitre 19 mais rien ne garantit que le champ résultat sera compatible avec l'équation discrète.

Si on dispose d'une information partielle sur le champ non-bruité (par exemple, pour le problème de Cauchy ci-dessus, la connaissance de la condition initiale non-bruitée), on peut utiliser l'équation elle-même comme filtre. Considérons une discrétisation implicite de l'équation ci-dessus. Le pas de temps Δt correspond au temps écoulé entre deux observations. Supposons que u_h^n est non-bruité, alors à partir de u_h^n et $\hat{u}_h^{n+1}$ on construit u_h^{n+1} et f_h^n comme suit :

— Prédiction de la source :

$$f_h^{n+1} = \frac{\hat{u_h}^{n+1} - u_h^n}{\Delta t} + L_h(\hat{u}_h^{n+1}),$$

— Prédiction du champ :

$$u_h^{n+1} = u_h^n + \Delta t(-L_h(\hat{u}_h^{n+1}) + f_h^{n+1}),$$

— Correction de la source :

$$f_h^{n+1} = \frac{u_h{}^{n+1} - u_h^n}{\Delta t} + L_h(u_h^{n+1}),$$

— Correction du champ :

$$u_h^{n+1} = u_h^n + \Delta t(-L_h(u_h^{n+1}) + f_h^{n+1}).$$

Ces quatre étapes permettent de ne pas recourir à un filtrage a priori. Bien entendu, on peut utiliser une discrétisation plus précise en temps de l'équation comme celles présentées au chapitre 10.

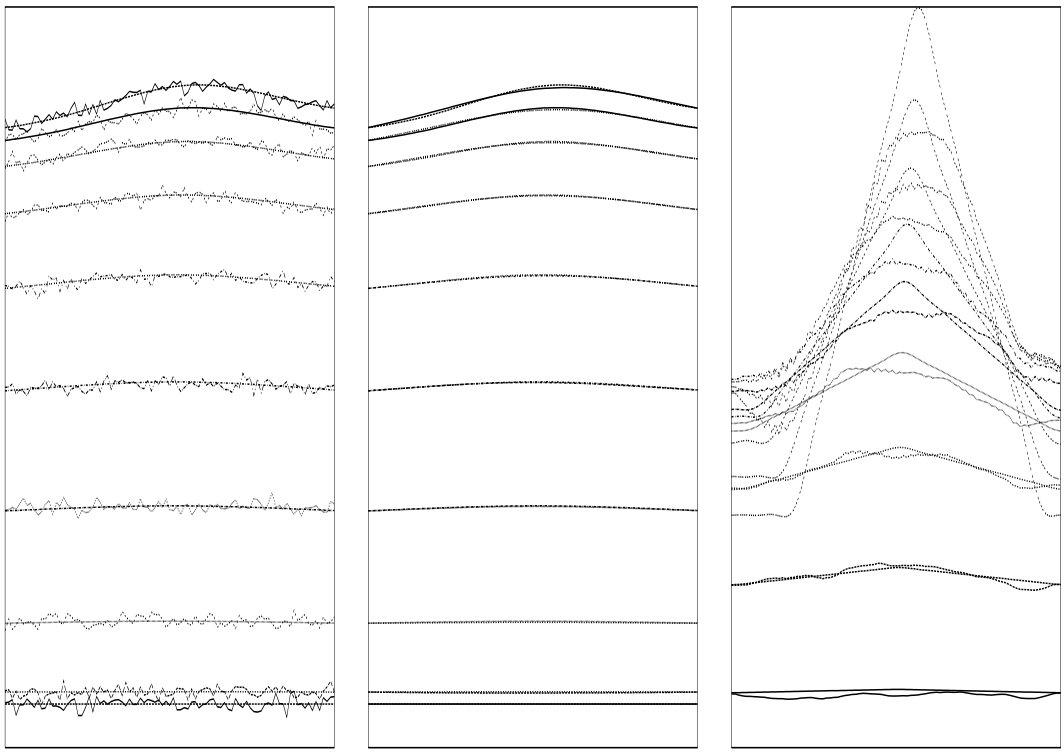


FIGURE 17.17 – Exemple d'identification des sources thermiques. Expériences réalisées par A. Chrysochoos à Montpellier. Chaque ligne horizontale représente une observation. Gauche : observations temporelles de la température (contenant du bruit) et le champ connu (cible). Milieu : température reconstruite comparée au champ cible. Droite : source reconstruite comparée à la source cible. La détection de la source est satisfaisante mais n'est pas parfaite. Aucun filtre n'a été utilisé.

Identification de sources thermiques

On présente l'application de l'algorithme prédicteur-correcteur ci-dessus à la reconstruction d'une source thermique à partir des observations de champs de température. On utilise l'équation de chaleur instationnaire suivante et une discrétisation uniforme cartésienne provenant d'images numériques obtenues pour la température par une caméra infra-rouge.

$$\frac{\partial T}{\partial t} - \nu \frac{\partial^2 T}{\partial x^2} + \alpha T = f(x,t), \quad \text{sur} \quad \Omega =]0,1[\tag{17.31}$$

avec une condition initiale uniforme connue (cette information n'est pas bruitée) et des conditions aux limites de Fourier :

$$T(x, t=0) = T_0, \quad T(\partial\Omega, t) + a \frac{\partial T}{\partial x}(\partial\Omega, t) = b.$$

On présente le résultat en Fig. 17.17. On constate que le filtrage par l'équation est efficace et l'évolution de la source est relativement bien identifiée, mais n'est pas parfaite malgré un champ presque parfaitement reconstruit.

Chapitre 18

Estimation d'erreur et adaptation de maillage

Nous considérons, dans ce chapitre, le problème de l'erreur de discrétisation spatiale dans les modèles numériques et de l'adaptation du maillage afin de minimiser, ou, au moins, de contrôler cette erreur. L'expérience montre que la qualité du maillage est cruciale pour la qualité du résultat.

Nous commençons par des résultats de majoration d'erreur a priori dans le cas de maillages en éléments finis. Nous présentons ensuite les aspects pratiques de l'estimation d'erreur a posteriori et de l'adaptation de maillage. Il est important de remarquer que les estimations a priori supposent la connaissance de la solution et de ses dérivées. Tandis que les estimations a posteriori utilisent les solutions approchées et permettent l'adaptation des maillages en pratique.

La performance des méthodes d'éléments ou de volumes finis est étroitement liée à la qualité du maillage. Bien sur, on peut envisager de raffiner le maillage jusqu'à l'obtention d'une solution indépendante du maillage (ou convergée en maillage). Mais ceci implique un coût de calcul très important. En effet, à chaque raffinement uniforme, on multiplie le nombre de points par 8 en 3D. On recherchera alors des raffinements locaux.

Pour un numéricien il est utile de connaître les étapes essentielles des algorithmes de maillage car dans les applications on utilise de façon croissante des outils de simulation boîte-noire, proposant leur mailleur intégré. Nous nous intéressons à l'algorithme de Delaunay car il s'adapte bien au contexte de simulation et adaptation en maillage nonstructuré. L'adaptation se fait à deux niveaux : définition d'une distribution locale des tailles de mailles, puis utilisation d'un outil de maillage géométrique pour générer le maillage. Nous tenterons de présenter les concepts en dimension un pour plus de simplicité.

18.1 Analyse d'erreur a priori dans les méthodes d'éléments finis

Nous ferons l'étude de l'erreur dans les deux cas d'intégration exacte et de quadrature approchée.

Les résultats d'existence et d'unicité des problèmes continus et discrets de la forme $a(u,v) = l(v)$ ont été donnés au chapitre 7.

18.1.1 Résultat général de majoration d'erreur a priori

Théorème 18.1.1 *Soit M la constante intervenant dans l'hypothèse de continuité de a : $a(u,v) \leq M \|u\| \|v\|$ et m la constante intervenant dans l'hypothèse d'ellipticité, on a la majoration d'erreur suivante :*

$$\|u - u_h\| \leq \frac{M}{m} inf_{v_h \in V_h} \|u - v_h\| \tag{18.1}$$

Si a est symétrique :

$$a(u - u_h, v_h) = 0 \quad \forall v_h \in V_h$$

signifie que u_h est la projection de u dans V_h au sens du produit scalaire a. Dans ce cas on a

$$\|u - u_h\| \leq \sqrt{\frac{M}{m}} inf_{v_h \in V_h} \|u - v_h\| \tag{18.2}$$

Dans la suite, nous considérons des problèmes elliptiques d'ordre 2. L'espace des fonctions tests considérées est l'espace H^1 ou un sous-espace de H^1.

18.1.2 Majoration d'erreur en éléments P_k ou Q_k avec intégration exacte

Nous supposons le domaine de calcul Ω exactement recouvert par l'ensemble des éléments du maillage. Dans le cas de triangles ou quadrangles droits le domaine sera donc supposé polygonal. Nous admettrons alors le résultat suivant :

Théorème 18.1.2 (Général) *Dans le cas d'éléments finis de degré k (P_k ou Q_k) et d'une solution exacte u du problème elliptique suffisamment régulière ($u \in H^{k+1}(\Omega)$), on a la majoration d'erreur suivante en norme H^1 :*

$$\|u - u_h\|_{1,2} \leq C h^k |u|_{k+1,2} \tag{18.3}$$

où h est la longueur du plus grand côté d'élément du maillage et où la constante C est indépendante de h.

Remarque : Si le domaine Ω n'est pas exactement recouvert par le maillage, on doit l'approcher par un maillage Ω_h dont la frontière suit les côtés des éléments droits ou isoparamétriques courbes choisis.

18.1.3 Un premier exemple simple : Erreur pour les éléments P1 en dimension un

Introduisons, comme dans le chapitre 7, une discrétisation de l'intervalle [a,b] en N sous-intervalles ou éléments $T_i = [x_{i-1}, x_i]$. V_h est alors l'espace des fonctions continues affines par morceaux (affines sur les segments T_i). Nous utilisons le résultat

$$\|u - u_h\|_{1,2} \leq \frac{M}{m} \|u - v_h\|_{1,2} \quad \forall v_h \in V_h \tag{18.4}$$

démontré plus haut.

En choisissant pour v_h l'interpolé $P_h u$ de u dans l'espace d'éléments finis V_h, on majore l'erreur d'approximation $\|u - u_h\|$ par une constante fois l'erreur d'interpolation $\|u - P_h u\|$. On en déduit les résultats suivants :

Théorème 18.1.3 *Posons* $h = \max_i |x_i - x_{i-1}|$, *on a* $\forall x \in [a, b]$ *:*

$$|u(x) - P_h u(x)| \leq \frac{h^2}{8} \max_{x \in [a,b]} |u''(x)| \tag{18.5}$$

$$|u'(x) - P_h' u(x)| \leq \frac{h}{2} \max_{x \in [a,b]} |u''(x)| \tag{18.6}$$

d'où

$$\|u - P_h u\|_{1,2} \leq C\, h \max_{x \in [a,b]} |u''(x)| \tag{18.7}$$

et enfin

$$\|u - u_h\|_{1,2} \leq C\, h \max_{x \in [a,b]} |u''(x)| \tag{18.8}$$

Démonstration :
Dans chaque intervalle $[x_{i-1}, x_i]$, notons de façon classique w_{i-1} et w_i les deux fonctions de base de V_h associées respectivement aux points x_{i-1} et x_i. Rappelons que leur restriction dans $[x_{i-1}, x_i]$ s'écrivent :

$$w_{i-1}(x) = \frac{x_i - x}{x_i - x_{i-1}} \qquad w_i(x) = \frac{x - x_{i-1}}{x_i - x_{i-1}} \tag{18.9}$$

On a les relations suivantes quel que soit x dans $[x_{i-1}, x_i]$:

$$w_{i-1}(x) + w_i(x) = 1 \qquad x_{i-1} w_{i-1}(x) + x_i w_i(x) = x$$

et

$$P_h u(x) = u(x_{i-1}) w_{i-1}(x) + u(x_i) w_i(x)$$

donc évidemment

$$(P_h u)'(x) = u(x_{i-1}) w'_{i-1}(x) + u(x_i) w'_i(x)$$

On utilise alors les développements de Taylor :

$$u(x_{i-1}) = u(x) + (x_{i-1} - x)u'(x) + \frac{(x_{i-1} - x)^2}{2} u''(\xi_i)$$

$$u(x_i) = u(x) + (x_i - x)u'(x) + \frac{(x_i - x)^2}{2} u''(\eta_i)$$

et l'on obtient :

$$P_h u(x) = u(x) + \frac{1}{2}[(x_{i-1} - x)^2 w_{i-1}(x) u''(\xi_i) + (x_i - x)^2 w_i(x) u''(\eta_i)]$$

et

$$(P_h u)'(x) = u'(x) + \frac{1}{2}[(x_{i-1} - x)^2 w'_{i-1}(x) u''(\xi_i) + (x_i - x)^2 w'_i(x) u''(\eta_i)]$$

D'où l'on déduira en exercice les majorations d'erreur annoncées en remarquant que $w_{i-1}(x)$ et $w_i(x)$ sont compris entre 0 et 1 et que

$$w'_{i-1}(x) = -w'_i(x) \quad \text{pour } x \in [x_{i-1}, x_i]$$

Remarque : Supposons que u soit un polynôme de degré deux. Alors u'' est une constante. On observe que dans ce cas le maximum de l'erreur d'interpolation de la solution u est atteint au milieu de l'intervalle. La dérivée de u est alors une fonction affine. La dérivée de son interpolée $P'_h u$ est une fonction constante égale à la valeur de u' au milieu de l'intervalle. En conclusion, au point milieu de l'intervalle $[x_{i-1}, x_i]$, on a dans ce cas, à la fois un maximum de l'erreur d'interpolation sur u et une erreur d'interpolation nulle sur u'. Ceci explique pourquoi il est avantageux d'estimer les dérivées de la solution aux points milieux des intervalles.

18.1.4 Les éléments P1 en dimension deux

Supposons le domaine polygonal plan Ω discrétisé par un maillage en triangles T_i. Prenons pour espace d'approximation V_h l'espace des fonctions continues affines par morceaux (affines sur les triangles T_i). Nous utilisons encore le résultat

$$\|u - u_h\|_{1,2} \leq \frac{M}{m} \|u - v_h\|_{1,2} \quad \forall v_h \in V_h \tag{18.10}$$

En choisissant à nouveau pour v_h l'interpolé $P_h u$ de u dans l'espace d'éléments finis V_h, on majore l'erreur d'approximation $\|u - u_h\|$ par une constante fois l'erreur d'interpolation $\|u - P_h u\|$.

On en déduit les résultats suivants :

Théorème 18.1.4 *Notons h la longueur du plus grand côté et θ_0 le plus petit angle au sommet de tous les triangles du maillage et notons D^2v la matrice Hessienne d'une fonction deux fois continûment différentiable v.*

$$D^2v(x,y) = \begin{pmatrix} \dfrac{\partial^2 v}{\partial x^2} & \dfrac{\partial^2 v}{\partial x \partial y} \\ \dfrac{\partial^2 v}{\partial x \partial y} & \dfrac{\partial^2 v}{\partial y^2} \end{pmatrix} \tag{18.11}$$

et $|D^2v(x,y)|$ sa norme spectrale.
On a alors $\forall x, y \in \Omega$:

$$|u(x,y) - P_h u(x,y)| \leq \frac{h^2}{2}\, sup_{x,y\in\Omega} |D^2 v(x,y)| \tag{18.12}$$

$$|gradu(x,y) - gradP_h u(x,y)| \leq 3\frac{h}{sin(\theta_0)}\, sup_{x,y\in\Omega} |D^2 v(x,y)| \tag{18.13}$$

d'où

$$\|u - P_h u\|_{1,2} \leq C\, h\, sup_{x,y\in\Omega} |D^2 v(x,y)| \tag{18.14}$$

et enfin

$$\|u - u_h\|_{1,2} \leq C\, h\, sup_{x,y\in\Omega} |D^2 v(x,y)| \tag{18.15}$$

Démonstration :
Nous suivons la démonstration de R. Glowinski (voir bibliographie).

Dans chaque triangle T_i, notons de façon classique λ_1, λ_2 et λ_3 les trois coordonnées barycentriques, restrictions dans le triangle des fonctions de base de V_h associées aux trois sommets du triangle T_i. On a les relations suivantes quel que soit x, y dans T_i :

$$\lambda_1(x,y) + \lambda_2(x,y) + \lambda_3(x,y) = 1$$

$$x_1\lambda_1(x,y) + x_2\lambda_2(x,y) + x_3\lambda_3(x,y) = x$$

$$y_1\lambda_1(x,y) + y_2\lambda_2(x,y) + y_3\lambda_3(x,y) = y$$

et

$$P_h u(x,y) = u(x_1,y_1)\lambda_1(x,y) + u(x_2,y_2)\lambda_2(x,y) + u(x_3,y_3)\lambda_3(x,y)$$

donc évidemment

$$grad(P_h u)(x,y) = u(x_1,y_1)grad\lambda_1 + u(x_2,y_2)grad\lambda_2 + u(x_3,y_3)grad\lambda_3$$

On utilise alors les développements de Taylor :

$$u(x_i,y_i) = u(x,y) + \overrightarrow{MA_i}.\overrightarrow{gradu}(x,y) + \frac{1}{2}\overrightarrow{MA_i}.D^2u(\xi,\eta)\overrightarrow{MA_i}$$

et l'on obtient :

$$P_h u(x,y) = u(x,y) + \frac{1}{2}\sum_{i=1}^{i=3} \lambda_i(x,y)\overrightarrow{MA_i}.D^2u(\xi,\eta)\overrightarrow{MA_i}$$

et

$$\overrightarrow{grad}P_h u(x,y) = \overrightarrow{grad}u(x,y) + \frac{1}{2}\sum_{i=1}^{i=3} \overrightarrow{grad}\lambda_i(x,y).\overrightarrow{MA_i}.D^2u(\xi,\eta)\overrightarrow{MA_i}$$

D'où l'on déduira en exercice les majorations d'erreur annoncées.

Remarques :

1) Le terme dominant de l'erreur est celui portant sur les gradients. Il sera minimal pour $\dfrac{h}{sin(\theta_0)}$ minimal. Parmi tous les triangles de plus grand côté de longueur h, celui pour lequel ce terme sera minimal est celui pour lequel $sin(\theta_0)$ sera maximal. On trouve ainsi comme forme optimale le triangle équilatéral. C'est aussi le triangle d'aire maximale inscrit dans un cercle de rayon donné.

2) Supposons que u soit un polynôme de degré deux. Alors D^2u est une matrice constante. Dans le cas particulier de problème homogène et isotrope, la matrice Hessienne D^2u est de plus scalaire. Dans chaque élément triangulaire le maximum de l'erreur d'interpolation de la solution u est alors atteint au centre du cercle circonscrit au triangle. Le gradient de u est alors une fonction affine. Le gradient de son interpolée $\mathbf{grad}P_h u$ est une fonction constante égale à la valeur de $\mathbf{grad}u$ au centre du cercle circonscrit au triangle. En conclusion, en ce point, on a dans ce cas, à la fois un maximum de l'erreur d'interpolation sur u et une erreur d'interpolation nulle sur $\mathbf{grad}u$. Ceci explique pourquoi il est avantageux d'estimer les dérivées de la solution aux points centres des cercles circonscrits aux éléments.

Par contre dans le cas de problèmes non-isotropes où la solution varie beaucoup plus fortement dans une direction, l'analyse précédente est insuffisante. Une analyse plus fine conduit à rechercher les valeurs propres et vecteurs propres des matrices Hessiennes dans chaque triangle et à en déduire comme triangle optimal celui d'aire maximale inscrit dans une ellipse. Dans le cas de fortes variations dans une direction donnée (phénomènes de couches limites ou de chocs) des triangles très aplatis donnent de bons résultats.

18.2 Analyse de l'erreur en cas d'intégration numérique

Dans le cas d'intégration approchée (voir les différentes formules d'intégration numérique, en dimension un, deux ou trois dans le chapitre 2), on passe du problème P_h

$$(P_h)\begin{cases} \text{Trouver la fonction } u_h \text{ appartenant à l'espace de Hilbert } V_h \text{ telle que :} \\ a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h \end{cases}$$

à un nouveau problème P_h^*

$$(P_h^*)\begin{cases} \text{Trouver la fonction } u_h^* \text{ appartenant à l'espace de Hilbert } V_h \text{ telle que :} \\ a_h(u_h^*, v_h) = l_h(v_h) \quad \forall v_h \in V_h \end{cases}$$

a_h et l_h sont les approximations par intégration numérique des formes exactes a et l, et u_h^* est la solution approchée du nouveau problème P_h^*.

Deux questions se posent alors :

1) Ce nouveau problème admet-il une solution unique ?

2) Quelle est l'erreur commise ? C'est à dire : quelle majoration obtient-on pour $\|u - u_h^*\|_{1,2}$ ou, ce qui revient au même, quelle précision doit-on imposer à l'intégration approchée pour que cette nouvelle erreur soit du même ordre que l'erreur d'interpolation ? Ceci afin de conserver, dans le cas de quadrature approchée, une erreur du même ordre que dans le cas d'intégration exacte.

La réponse à la première question est donnée par le théorème de Lax-Milgram. Le problème P_h^* admet une solution unique dans V_h si les formes a_h et l_h sont continues et si a_h est elliptique dans V_h c'est à dire s'il existe une constante $m > 0$ telle que :

$$a_h(v_h, v_h) \geq m\|v_h\|_{1,2}^2 \quad \forall v_h \in V_h$$

Comme V_h est un espace de dimension finie, les formes linéaires a_h et l_h sont continues. Il reste donc à s'assurer de l'ellipticité de la forme approchée a_h afin d' avoir une solution unique au problème.

18.2.1 Condition d'ellipticité

Les conditions à imposer à la formule d'intégration numérique pour assurer l'ellipticité de la forme bilinéaire approchée sont les suivantes

1) La formule doit être à coefficients positifs.

2) Elle doit comporter un nombre de points suffisants.

Plus précisément, dans le cas d'éléments P_k en dimension un, deux, ou trois, le nombre de points d'intégration par éléments doit être tel qu'il suffise à définir de façon unique un polynôme P_{k-1}

Dans le cas d'éléments Q_k, le nombre de points d'intégration par éléments doit être tel qu'il suffise à définir de façon unique une fonction de Q_k

Nous préciserons plus loin, dans des cas pratiques, les conditions à imposer aux formules d'intégration numérique pour que cette condition d'ellipticité soit vérifiée.

18.2.2 Majoration d'erreur avec intégration numérique

La réponse à la deuxième question est donnée par le théorème suivant :

Théorème 18.2.1 *Dans le cas d'éléments finis P_k, une formule d'intégration exacte sur l'espace des polynômes P_{2k-2} assure une erreur d'intégration en norme H^1 en $O(h^k)$ donc du même ordre que l'erreur d'interpolation. L'erreur globale reste alors d'ordre k.*

Dans le cas d'éléments Q_k, le même résultat est obtenu si la formule d'intégration est exacte sur Q_{2k-1}.

18.3 Conséquences pratiques

18.3.1 En dimension un

Pour le problème modèle :

$$\begin{cases} \text{Trouver la fonction } u \text{ appartenant à l'espace } H_0^1(a,b) \text{ telle que :} \\ a(u,v) = l(v) \quad \forall v \in H_0^1(a,b) \end{cases} \tag{18.16}$$

avec

$$a(u,v) = \int_a^b u'v'dx + \int_a^b u\,v\,dx$$

et

$$l(v) = \int_a^b f\,v\,dx$$

nous obtenons les choix suivants :

Éléments P_1

Une formule d'ordre 0 à un point suffirait. On choisit cependant habituellement la formule des trapèzes qui donne une erreur en $O(h^2)$ en norme L^2 et conduit à des matrices de masse diagonales.

Éléments P_2

Une formule d'ordre 2 à deux points suffit. On peut choisir la formule de Gauss-Legendre ou celle de Simpson (voir chapitre 2, paragraphe 2.4).

Éléments P_k

Une formule d'ordre $2k-2$ à k points suffit.

18.3.2 En dimension deux

Pour le problème modèle en dimension deux :

$$\begin{cases} \text{Trouver la fonction } u \text{ appartenant à l'espace } H_0^1(\Omega) \text{ telle que :} \\ a(u,v) = l(v) \quad \forall v \in H_0^1(\Omega) \end{cases} \tag{18.17}$$

avec

$$a(u,v) = \iint_\Omega \mathbf{grad}u\ \mathbf{grad}v\ dxdy + \iint_\Omega u\, v\, dxdy$$

et

$$l(v) = \iint_\Omega f\, v\, dxdy$$

nous obtenons les choix suivants :

Éléments P_1

Une formule d'ordre 0 à un point suffirait. On choisit cependant habituellement la formule suivante construite sur les sommets :

$$\iint_T F(x,y)\, dxdy = \frac{Aire(T)}{3}[F(A_1) + F(A_2) + F(A_3)] \tag{18.18}$$

qui donne une erreur en $O(h^2)$ en norme L^2 et conduit à des matrices de masse diagonales.

Éléments P_2

Une formule d'ordre 2 à trois points suffit. On peut choisir la formule suivante construite sur les points milieux des côtés :

$$\iint_T F(x,y)\, dxdy = \frac{Aire(T)}{3}[F(A_{12}) + F(A_{23}) + F(A_{31})] \tag{18.19}$$

Éléments P_k

Une formule d'ordre $2k-2$ à $\dfrac{k(k+1)}{2}$ points suffit.

Éléments Q_1

On doit utiliser une formule exacte sur Q_1 à 4 points. On choisit habituellement la formule suivante construite sur les sommets du carré unité :

$$\iint_C F(x,y)dxdy \approx \frac{1}{4}[F(A_1) + F(A_2) + F(A_3) + F(A_4)] \tag{18.20}$$

qui donne une erreur en $O(h^2)$ en norme L^2 et conduit à des matrices de masse diagonales.

Éléments Q_2

On doit utiliser une formule exacte sur Q_3 à 9 points au moins. On peut choisir la formule déduite de Simpson suivante :

$$\begin{aligned}\iint_C F(x,y)dxdy \approx \frac{1}{36}[F(A_1)+F(A_2)+F(A_3)+F(A_4)]\\ +\frac{1}{9}[F(A_{12})+F(A_{23})+F(A_{34})+F(A_{41})]+\frac{4}{9}F(G)\end{aligned} \tag{18.21}$$

où $A_{12}, A_{23}, A_{34}, A_{41}$ sont les milieux et G le barycentre du carré C.

18.3.3 Contre-exemples : formes approchées non elliptiques

Reprenons le problème modèle en dimension deux du paragraphe précédent.

Avec des éléments Q_1 : la formule suivante à un point (G est le barycentre du carré unité C) est également exacte sur Q_1.

$$\iint_C F(x,y)dxdy \approx F(G)$$

Cependant elle n'assure pas l'ellipticité de la forme a_h et conduit à des schémas numériques instables

En effet la fonction

$$v(x,y) = (x-\frac{1}{2})(y-\frac{1}{2})$$

vérifie d'une part

$$a(v,v) > 0 \quad \text{et} \quad \|v\|_{1,2} \neq 0$$

D'autre part cette fonction et son gradient s'annulent au point G et donc on a $a_h(v,v) = 0$ si l'on calcule l'intégrale selon la formule à un point ci dessus. En conséquence il n'y aura pas ellipticité, car on ne peut trouver $m > 0$ tel que $a_h(v,v) \geq m\|v\|_{1,2} \quad \forall v \in H_0^1$

Avec des éléments Q_2 : on vérifiera que la formule déduite de Gauss-Legendre ne permet pas non plus d'assurer l'ellipticité. [Prendre la fonction $v(x,y) = (x-\xi_1)(x-\xi_2)(y-\xi_1)(y-\xi_2)$ où ξ_1 et ξ_2 sont les abscisses et ordonnées des points de Gauss-Legendre dans le carré unité (voir chapitre2, paragraphe 2.4.5).]

18.4 Estimation d'erreur a posteriori

Il existe plusieurs sources d'erreurs : erreurs dues à la modélisation, erreurs dues à la discrétisation du modèle continu et à la qualité de la résolution de ce dernier. Enfin, une fois le résultat obtenu, celles introduites lors de l'analyse des résultats, car les outils de post-traitement ne travaillent pas forcement sur les mêmes espaces que les solveurs. Pensez à un calcul effectué avec des éléments finis d'ordre élevé et un résultat visualisé avec un outil graphique standard ne disposant que d'une représentation linéaire par morceaux. Cette difficulté d'interprétation des résultats est une des difficultés majeures lors de l'utilisation des méthodes d'ordre élevé, permettant d'accéder à la solution et ses dérivées de façon précise, sur un nombre réduit de points.

18.4.1 Critère du gradient

Une première idée lors en adaptation de maillage est de concentrer les points dans les régions où la solution a de fortes variations. Un élément T est subdivisé si le gradient discret

$$\|\nabla_h u_h|_T\| \geq \text{Tolérance.} \tag{18.22}$$

L'algorithme de raffinement s'écrit :

1− Définir une distribution $h(x)$ initiale.
2− Calculer u_h solution de $\mathcal{L}_h u_h = f_h$.
3− Calculer $\nabla_h u_h$.
4− Découper chaque élément T, si (18.22) est vérifié, et retour en 2, 5− Sinon, Fin.

18.4.2 Contrôle local de métrique

Dans cette approche, on modifie le choix du produit scalaire euclidien qui sert à définir les distances dans le maillage pour pouvoir générer les éléments dans une nouvelle métrique suivant certains critères. On cherchera la métrique qui équirépartira l'erreur d'interpolation. Présentons cette technique dans le cadre de la dimension 1. On dispose de la majoration suivante pour l'erreur d'interpolation en élément $P1$ (voir ci-dessus) :

$$\|u - \Pi_h u\|_0 \leq c\, h^2\, |u''|, \tag{18.23}$$

où $\Pi_h u$ est l'interpolation linéaire de u, h la taille des éléments. La métrique locale en chaque nœud x du maillage est définie par :

$$\lambda_i = \min\left(\max\left(\frac{1}{\epsilon}|u''|\,, \frac{1}{h_{\min}^2}\right), \frac{1}{h_{\max}^2}\right), \tag{18.24}$$

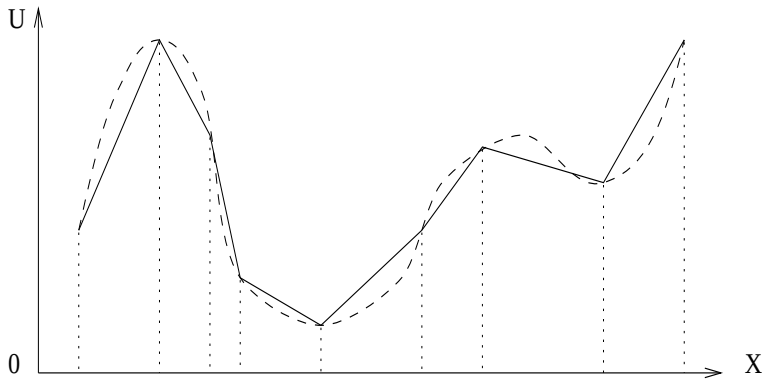


FIGURE 18.1 – Approximation $P1$ Lagrange de u.

où ϵ, $h_{\min}$ et $h_{\max}$ sont respectivement l'erreur et les longueurs minimale et maximale désirées des mailles. La distance entre deux points sera alors définie par :

$$|x-y|_\lambda(\eta) = \sqrt{\lambda(\eta)}|x-y|, \quad \eta \in [x,y].$$

On peut ainsi choisir d'équirépartir l'erreur d'interpolation en générant des segments de même longueur h dans cette nouvelle métrique. En notant a_i la distance entre le i^{eme} et le $(i+1)^{eme}$ nœud de la discrétisation, on a :

$$a_i = \frac{h}{\sqrt{\lambda_i}}\,. \tag{18.25}$$

De la même manière, en dimension supérieure, on définit la distance locale par :

$$\langle x\,,\,y\rangle_\lambda = x^t\,\lambda\,y \iff \|x\,-\,y\|_\lambda = \sqrt{(x\,-\,y)^t\,\lambda\,(x\,-\,y)}\,. \tag{18.26}$$

18.4.3 Problème adjoint et adaptation de maillage

Il arrive que l'on souhaite uniquement optimiser le maillage pour le calcul d'une fonctionnelle locale basée sur la variable du problème, et non pas de façon précise pour le calcul de la variable elle-même. Par exemple, on peut souhaiter calculer la traînée d'un avion avec une certaine précision, sans s'intéresser aux détails de l'écoulement loin de l'avion. On utilise les outils étudiés au chapitre 17 pour l'optimisation en adaptation de maillage. En particulier, on utilisera la variable adjointe, que l'on réintroduit brièvement. Considérons le problème de minimisation suivant :

$$\min_x J(x,u_h), \quad F_h(x,u_h(x)) = 0, \tag{18.27}$$

où x désigne la paramétrisation du problème, J une fonctionnelle coût positive, u_h l'état, solution de l'équation d'état discrète $F_h(u_h) = 0$. On a vu que le calcul du gradient dJ/dx implique le calcul de $\partial u_h/\partial x$ qui est coûteux :

$$\frac{dJ}{dx} = \frac{\partial J}{\partial x} + \frac{\partial J}{\partial u_h}\frac{\partial u_h}{\partial x},$$

mais qui peut être vu formellement comme :

$$\frac{dJ}{dx} = \frac{\partial J}{\partial x} + \frac{\partial J}{\partial u_h}(\frac{\partial u_h}{\partial F_h})^{-1}\frac{\partial F_h}{\partial x}.$$

La technique adjointe consiste alors à introduire le système intermédiaire suivant :

$$v_h^T(\frac{\partial F_h}{\partial u_h}) = \frac{\partial J}{\partial u_h},$$

ce qui permet d'accéder au gradient dJ/dx par :

$$\frac{dJ}{dx} = \frac{\partial J}{\partial x} + v_h^T\frac{\partial F_h}{\partial x}.$$

On peut utiliser la variable adjointe pour l'adaptation de maillage de la manière suivante : considérant une fonctionnelle J, peut-on trouver un maillage optimal pour J, (i.e. contrôlant l'erreur de troncature sur J : $\varepsilon_J = |J(u) - J(u_h)|$) ? u_h est une solution approchée ($F_h(u_h) = 0$).

Pour simplifier on a supposé que le calcul de la fonctionnelle était exact. Dans le cadre d'intégration approchée, il faudrait introduire une fonctionnelle J_h discrète.

L'erreur de troncature sur J peut être liée à l'erreur δu_h (i.e. $u = u_h + \delta u_h$) sur l'état par :

$$J(u) - J(u_h) = \frac{\partial J}{\partial u_h}\delta u_h = \frac{\partial J}{\partial u_h}(\frac{\partial F_h}{\partial u_h})^{-1}F_h(u),$$

car

$$F_h(u) = F_h(u_h) + \frac{\partial F_h}{\partial u_h}\delta u_h = \frac{\partial F_h}{\partial u_h}\delta u_h,$$

car $F_h(u_h) = 0$ par définition. Ce qui donne, en utilisant la définition de la variable adjointe v_h :

$$\varepsilon_J = |J(u) - J(u_h)| = v_h^T F_h(u).$$

Ainsi, si l'on connaît une estimation de l'erreur de troncature $F_h(u)$, on en obtient une pour J.

18.5 Génération automatique de maillage

Nous détaillons brièvement un algorithme de génération automatique de maillage triangulaire utilisant le critère de Delaunay. Cet algorithme sera utilisé pour intégrer l'adaptation de maillage par contrôle de métrique présenté ci-dessus.

18.5.1 Le critère de Delaunay

Il s'agit d'éviter l'apparition de triangles avec des angles petits ou grands, ainsi que la création de triangles voisins de taille très différente.

Dans un maillage formé de triangles, à chaque segment interne du maillage on peut associer un quadrangle formé de deux triangles partageant ce segment.

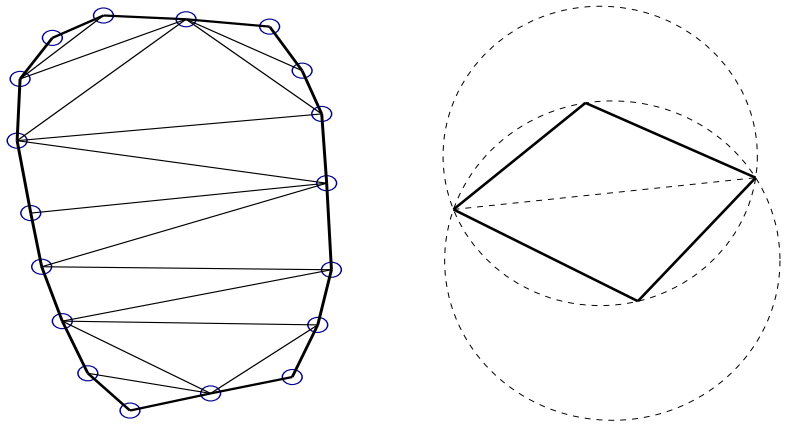


FIGURE 18.2 – A gauche, la discrétisation frontière et le maillage initial liant les points frontières. A droite : deux triangles partageant un segment interne. Le critère de Delaunay n'est pas vérifié, il faut retourner ce segment. On constate que l'angle le plus petit sera augmenté.

Une triangulation est dite Delaunay si pour chaque arête du maillage, aucun des deux cercles définis par les sommets du segment et chacun des deux autres sommets ne contient les quatre sommets (voir figure 18.2).

Retournement ou swap de segment : si les quatre sommets d'un quadrangle associé à un segment interne ne sont pas cocycliques, alors des deux configurations obtenues en retournant ce segment, l'une est Delaunay et l'autre non.

De plus on remarque que lorsque, par retournement de segment, une configuration devient Delaunay, l'angle minimum dans les deux triangles augmente. Ceci est l'intérêt fondamental de cette manipulation.

Algorithme d'enforcement du critère de Delaunay

— Faire jusqu'à convergence :
— Boucle sur les segments internes $E = (q^1, q^2)$:
— Trouver les deux triangles $T_k = (q^1, q^2, q^3)$ et $T_l = (q^2, q^1, q^4)$;
— Vérifier le critère de Delaunay. S'il n'est pas vérifié, remplacer T_k,T_l par les triangles (q^3, q^4, q^1) et (q^4, q^3, q^2).

Cet algorithme converge car, à chaque itération, l'angle minimum augmente, jusqu'à être plus grand qu'un autre angle du maillage. Alors, c'est cet angle qui augmentera. Le nombre de configurations étant fini, l'algorithme convergera. Cet algorithme a une complexité linéaire.

En maximisant l'angle minimum dans les éléments, cet algorithme produit un maillage formé de triangles (tétrahédres en 3D) le plus équilatéraux possible.

Ces opérations peuvent s'effectuer dans une métrique locale et permettent l'introduction de l'anisotropie dans le maillage.

18.5.2 Algorithme de génération de maillage

Un mailleur automatique utilisant le critère de Delaunay fonctionne suivant les étapes suivantes :

— 1. Discrétiser de façon régulière la frontière de Ω_h (régulière pour la métrique euclidienne). La répartition des points sur la frontière définit le raffinement du maillage dans le domaine.
— 2. Lier les points frontières entre eux sans introduire de points intérieurs.
— 3. Répartir régulièrement les points internes sur les segments déjà crées (en utilisant la métrique euclidienne).
— 4. Connecter les points ci-dessus entre eux en utilisant le critère de Delaunay et la métrique euclidienne.
— 5. Eliminer les éléments extérieurs à Ω_h.
— 6. Retour en 3 jusqu'à l'obtention de la qualité requise (dans la métrique euclidienne).

Maillage d'un rectangle par Delaunay

Soit $N + 4$ points, dont les quatre derniers désignant un rectangle contenant les autres points. Les points sont numérotés $(q^0, ..., q^{N-1}, q^N, q^{N+1}, q^{N+2}, q^{N+3})$.

Nous commençons avec les deux triangles (q^N, q^{N+1}, q^{N+2}), (q^{N+2}, q^{N+3}, q^N).

On effectue une boucle rétrograde sur les points tout en stockant la liste des triangles déjà créés identifiés par leurs sommets.

Pour $i = N - 1$ à 0 **Faire**

Si Il existe un triangle déjà créé qui contient strictement q^i. Alors, on remplace ce triangle par 3 sous-triangles ayant q^i comme sommet.

Sinon, Si q^i est sur la frontière du rectangle. Alors on remplace le triangle contenant strictement q^i par deux sous-triangles ayant q^i comme sommet.

Sinon q^i appartient à deux triangles (i.e. q^i est sur le segment commun). Chaque triangle est remplacé par deux sous-triangles ayant q^i comme sommet.

Fin

Fin de la boucle

18.6 Adaptation de maillage

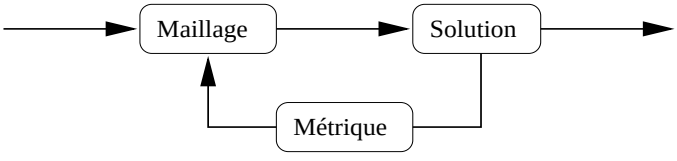
Il est évident qu'un raffinement uniforme du maillage conduit à des temps de calculs trop lourds. Toujours, en utilisant la métrique locale définie plus haut, deux solutions sont possibles : adapter le maillage initial en raffinant-déraffinant localement ou bien regénérer complétement le maillage.

18.6.1 Raffinement-déraffinement

La première technique d'adaptation de maillage à la solution est le raffinement-déraffinement. Le raffinement est relativement facile à mettre en oeuvre. Il consiste à couper les éléments, de façon conforme, en faisant apparaître de nouveaux éléments. Par exemple, en 2D en découpant les triangles en 4,3 ou 2, on raffine localement un maillage formé de triangles. Le problème alors est la perte de la qualité des éléments (définie par le rapport hauteur-base ou des angles min et max par exemple). En coupant uniquement par 4 chaque triangle, on conserve la qualité initiale du maillage mais on doit alors raffiner tout le maillage. Le déraffinement est plus difficile car lorsque l'on supprime un point, il faut établir une nouvelle connectivité localement suivant certains critères, comme par exemple celui de Delaunay.

18.6.2 Remaillage par contrôle de métrique

La structure de la boucle adaptative est :



Maillage de fond A chaque itération d'adaptation, le maillage, où la solution et la métrique sont disponibles, sert de maillage de fond pour la définition du nouveau maillage. Pour générer le nouveau maillage, connaissant au moins un

nœud de celui-ci (en général un point limite) et la métrique qui lui est associée, on place le nœud suivant. Étant donné que l'on ne connaît pas la métrique associée à ce nouveau nœud, on interpole à partir du maillage de fond. Puis on place le nouveau nœud et on réitère le procédé. Il faut insister sur le fait que, dans le cas où il y a unicité de solution, le maillage initial ne doit pas être nécessairement fin. Nous présentons un exemple ci-dessous où le maillage de départ ne comporte que 3 sommets et où on aboutit cependant à la solution avec un maillage final d'environ 200 sommets.

Exemple d'une équation d'advection-diffusion

On considère :

$$0.5\frac{\partial u}{\partial x} - 0.01\frac{\partial^2 u}{\partial x^2} = f(x), \quad \text{sur} \quad]0,1[,$$

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

$$f(x) = \begin{cases} 0 & \text{si } 0.2 \leq x < 0,5 \\ 20\,\sin(100\,x)\,\cos(50\,x) & \text{si } 0,5 \leq x < 0,8 \\ -40\,\sin(200\,x)\,\cos(50\,x) & \text{si } x \geq 0,8 \end{cases} \tag{18.28}$$

Pour comparer plusieurs maillages, on doit calculer l'erreur commise lors du calcul de la solution sur chaque maillage par rapport à la solution exacte. Cependant, dans le cas général, on ne connaît pas la solution exacte du problème, on calcule cette solution de référence sur un maillage très fin. Le pas uniforme du maillage de référence est $h = 0,003$, ce maillage a donc 266 nœuds. Les maillages adaptés successifs sont construits avec les paramètres suivants :

$$h_{\min} = 0,003 \qquad h_{\max} = 0,1 \qquad err = 0,001.$$

On part d'un maillage uniforme de 3 nœuds. Après 6 adaptations l'erreur $\|u - u_h\|_0$ est d'environ 2 %.

18.7 Adaptation de maillages en instationnaire

L'utilisation de l'adaptation lors d'un calcul instationnaire introduit, à chaque changement de discrétisation et de connectivité, une source d'erreur importante par l'interpolation de la solution de l'ancien maillage sur le nouveau qui modifie les dérivées temporelles de la solution. Cette erreur existe aussi en stationnaire mais n'affecte pas la solution finale car les dérivées temporelles s'annulent. Nous avons évoqué l'approche ALE au chapitre 13 permettant la prise en compte de mouvements de maillage à nombre de points et connectivité fixe, mais ici

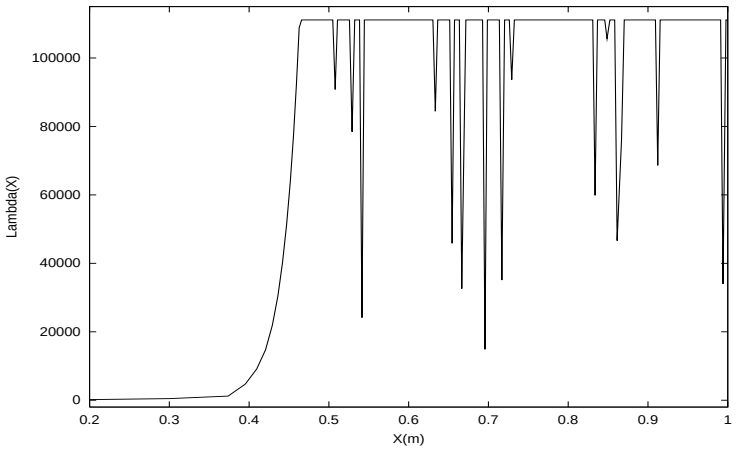


FIGURE 18.3 – Métrique associée au maillage adapté.

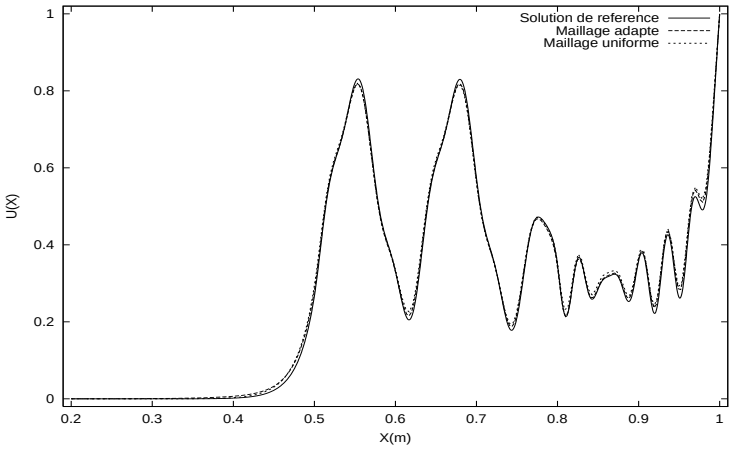


FIGURE 18.4 – Comparaison de la solution du maillage adapté de 189 nœuds et de la solution du maillage uniforme de 189 nœuds par rapport à la solution de référence.

la connectivité et le nombre de points varient. Nous allons présenter une autre façon de réduire l'effet de cette erreur sur les calculs. Considérons le problème d'advection-diffusion ci-dessus, mais avec un terme source dépendant du temps.

18.7.1 Un algorithme de point fixe

A l'itération i d'adaptation, on note le maillage, la solution discrète de l'équation et la métrique par $\mathcal{H}_i$, $\mathcal{S}_i$, $\mathcal{M}_i$. $\Delta t = \frac{\tau}{N_{adap}}$, où τ est le temps caractéristique le plus petit que l'on voudrait capturer. On pourra alors, en faisant deux simulations adaptatives avec N_{adap} et $2N_{adap}$, vérifier si le résultat est indépendant du maillage. La boucle de simulation adaptative est définie par :

initialisation : $i = 0, \quad t = 0$ $\qquad \mathcal{H}_0, \mathcal{S}_0, \Delta t$ donné,
Tant que $(t < t_{max})$**, Faire**
  Calculer la métrique : $(\mathcal{H}_i, \mathcal{S}_i) \rightarrow \mathcal{M}_i$,
  Générer le nouveau maillage : $(\mathcal{H}_i, \mathcal{M}_i) \rightarrow \mathcal{H}_{i+1}$,
  Interpoler l'ancienne solution : $(\mathcal{H}_i, \mathcal{S}_i, \mathcal{H}_{i+1}) \rightarrow \overline{\mathcal{S}}_{i+1}$,
  Avancer l'état en temps par Δt : $(\mathcal{H}_{i+1}, \overline{\mathcal{S}}_{i+1}) \rightarrow \mathcal{S}_{i+1}$,
  $i++$,
$t = t + \Delta t$,
Fin

Cet algorithme introduit des perturbations lors de son utilisation pour la résolution d'un problème dépendant du temps, notamment du fait du décalage entre la solution et la métrique qui sera d'autant plus grand que $\tau << \Delta t$. En effet, l'avance en temps de Δt se fera toujours par plusieurs itérations avec des pas de temps $\delta t < \tau$. On peut éviter cela en adaptant plus souvent le maillage, mais alors l'erreur introduite par l'étape d'interpolation devient dominante et le temps de calcul sera prohibitif. Nous allons introduire deux corrections qui amélioreront grandement l'application aux configurations instationnaires :

— Remplacer la métrique basée sur la dernière itération par (18.29) ci-dessous, prenant en compte les changements de la solution entre deux adaptations, et non pas uniquement la solution à la dernière itération avant adaptation. Ainsi, on utilise une combinaison des métriques intermédiaires. Plus exactement, si $u^p, p = n, ..., m$ sont des itérés successifs entre une adaptation à l'iteration n et la suivante à m, on définit la métrique combinée par :

$$\hat{\mathcal{M}} = \frac{1}{m-n} \sum_{p=n}^{m} M^p(u^p), \tag{18.29}$$

où chaque M^p est définie en utilisant (18.24).

— Introduire une boucle de point-fixe supplémentaire à chaque itération d'adaptation en faisant $NFIX$ iterations internes comme ci-dessous.

Le but est de trouver un point fixe pour l'ensemble (métrique, maillage, solution) :

$\mathcal{H}_0, \mathcal{S}_0, \Delta t, NFIX$ donné, $i = 0$, $t = 0$
Tant que $(t < t_{max})$**, Faire**

$$j = 0 \quad \hat{\mathcal{M}}_i^j = (\mathcal{I}(\frac{1}{h_{max}^2})),$$

Tant que $(j < NFIX$ ou $||(\frac{\partial \mathcal{S}}{\partial t})_i^{j+1} - (\frac{\partial \mathcal{S}}{\partial t})_i^j|| > TOL)$, **Faire**

— Générer le nouveau maillage basé sur l'intersection des métriques :

$$(\mathcal{H}_i^j, \hat{\mathcal{M}}_i^j) \rightarrow \mathcal{H}_i^{j+1},$$

— Interpoler l'ancienne solution et sa dérivée en temps sur le nouveau maillage :

$$(\mathcal{H}_i^{j+1}, \mathcal{S}_i^0, (\frac{\partial \mathcal{S}}{\partial t})_i^0, \mathcal{H}_i^j) \rightarrow (\overline{\mathcal{S}}_i^{j+1}, \overline{(\frac{\partial \mathcal{S}}{\partial t})}_i^{j+1}).$$

On insiste sur le fait que l'interpolation doit être effectuée à partir de la solution en début $(\mathcal{S}_i^0)$ de la boucle point fixe et non pas à partir de la solution courante $(\mathcal{S}_i^j)$.

— Avancer l'état en temps par Δt et définir la métrique pour le nouvel état :

$$(\mathcal{H}_i^{j+1}, \overline{\mathcal{S}}_i^{j+1}, \overline{(\frac{\partial \mathcal{S}}{\partial t})}_i^{j+1}) \rightarrow (\mathcal{S}_i^{j+1}, (\frac{\partial \mathcal{S}}{\partial t})_i^{j+1}, \mathcal{M}_i^{j+1}),$$

— Intersecter les métriques :

$$(\hat{\mathcal{M}}_i^j, \mathcal{M}_i^{j+1}) \rightarrow \hat{\mathcal{M}}_i^{j+1}.$$

$j++$ **Fin**.

$$(\mathcal{H}_{i+1}^0, \mathcal{S}_{i+1}^0, (\frac{\partial \mathcal{S}}{\partial t})_{i+1}^0) \leftarrow (\mathcal{H}_i^j, \mathcal{S}_i^j, (\frac{\partial \mathcal{S}}{\partial t})_i^j),$$

$i++$ **Fin**.

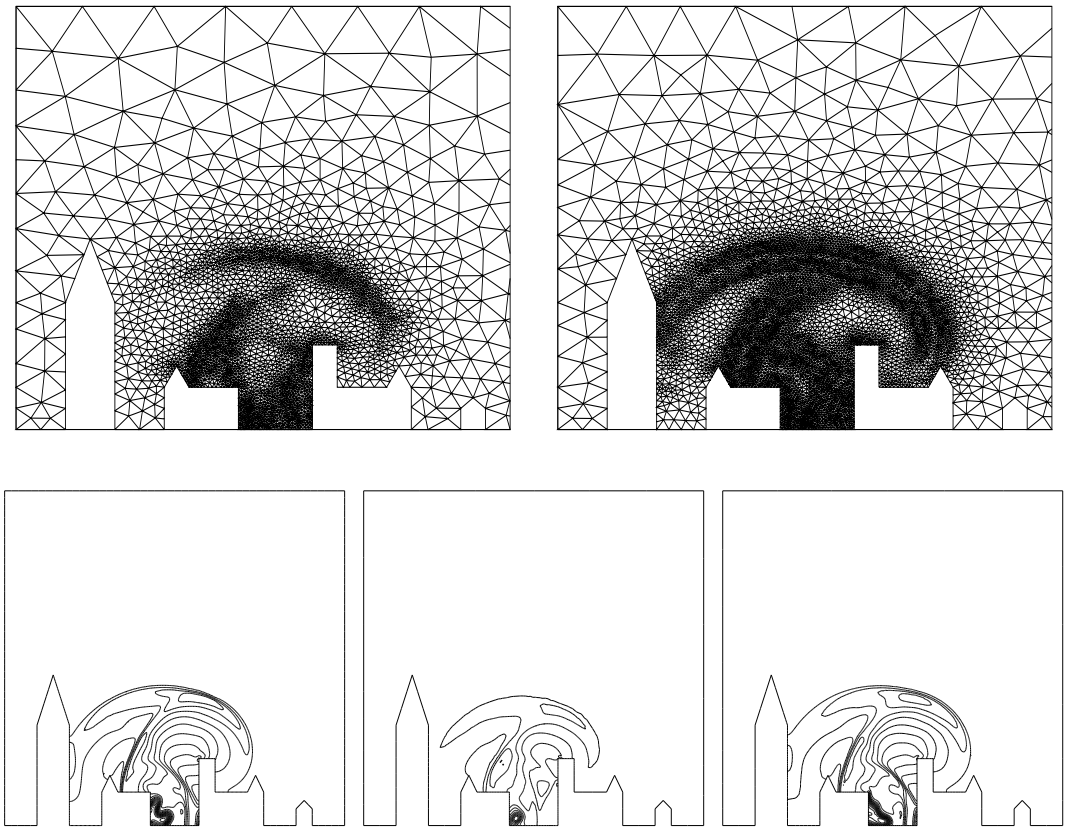


FIGURE 18.5 – *Propagation d'une onde de choc sur une géométrie complexe. Maillage par l'approche classique (haut-gauche) et par l'algorithme d'adaptation par point fixe (haut-droite). Solution de référence sur un maillage fin (bas-gauche). Solution calculée avec l'algorithme classique d'adaptation (bas-milieu). Solution calculée avec l'algorithme d'adaptation par point fixe (bas-droite). On constate l'importance de la recherche d'un point fixe pour l'ensemble (maillage-métrique-solution).*

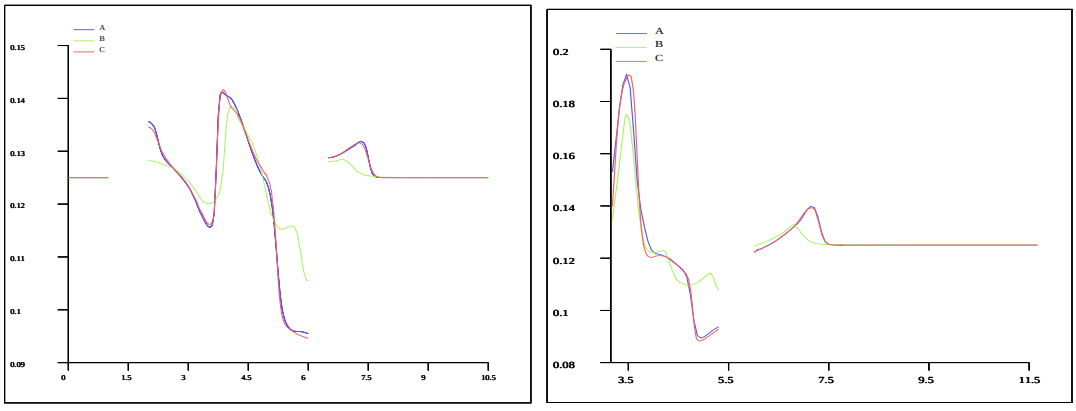


FIGURE 18.6 – *Deux coupes de la solution pour les trois simulations ci-dessus au même instant. On constate que l'algorithme de point fixe permet de retrouver la solution de référence. A : référence, B : algorithme adaptatif stationnaire, C : point fixe instationnaire (Calcul réalisé par F. Alauzet à l'INRIA).*

Chapitre 19

Filtres et EDP

19.1 Introduction

Dans ce chapitre, nous nous intéressons aux phénomènes multi-échelle et à leur capture. C'est un point essentiel pour la simulation dans la mesure où il est en général impossible d'utiliser la finesse de discrétisation nécessaire à la capture des effets de petite échelle. Le raffinement doit se faire à la fois sur les discrétisations temporelle et spatiale, car les différences d'échelles peuvent exister dans les deux domaines. Pour les grandes échelles, un maillage grossier suffit, tandis que pour la capture des effets de petite échelle et de haute-fréquence, il est nécessaire de disposer d'une discrétisation fine. De plus, du fait du conditionnement du système discret, il est difficile de représenter les grandes échelles sur un maillage fin imposé par les petites structures. Ainsi, on choisira de modéliser les petites structures, plutôt que de les capturer. Par modélisation on entend une description, non pas des petites échelles directement, mais plutôt du comportement de quantités moyennes associées ainsi que la prise en compte des corrélations entre petites et grandes échelles. Ce chapitre nous sert par ailleurs d'introduction aux diverses techniques de filtrage et à leurs propriétés.

19.2 Un problème modèle

Considérons l'EDP suivante de type Burgers :

$$\begin{cases} u_t + 0.5(u^2)_x - \nu u_{xx} = w(t,x,H), \\ u(t,0) = u_l, \quad u(t,L) = u_r, \quad u(0,x) = u_0(x), \end{cases} \tag{19.1}$$

La viscosité ν caractérise la capacité de l'EDP à lisser les perturbations introduites par w en espace et en temps. Ces perturbations sont difficiles à capturer numériquement car ceci exige un maillage très fin. H est la plus petite échelle

en espace des fluctuations. Une simulation avec une discrétisation plus fine est appelée une simulation directe et ne nécessite donc aucune modélisation, tandis qu'une discrétisation plus grossière nécessite la modélisation des échelles non représentées. On écrit la solution du modèle dans les variables adimensionnées (voir chapitre 3) $u^+ = u/u_f$ en fonction de $x^+ = xu_f/\nu$ avec $u_f = \sqrt{\nu u_r/L}$.

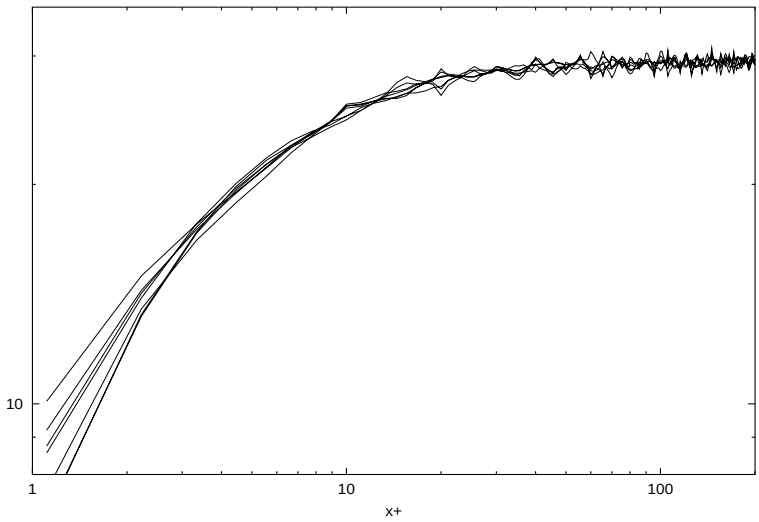


FIGURE 19.1 – Vues instantanées de la solution $u^+ = u/u_f$ en fonction de $x^+ = xu_f/\nu$ avec $u_f = \sqrt{\nu u_r/L}$ en échelle log-log pour une simulation directe sur un maillage uniforme.

19.3 Méthodes Monte Carlo

Les méthodes de Monte Carlo calculent des quantités moyennes en utilisant des corrélations sur un grand nombre de simulations ou observations. Considérons le problème de l'estimation du nombre π par tirage aléatoire. En tirant les points aléatoirement dans un carré de coté 2, centré en $(0, 0)$, la probabilité pour le point de se trouver dans le cercle de rayon 1 et centré en $(0, 0)$ est égale à $\pi/4$ (rapport entre les surfaces du cercle et du carré). On rapporte donc le nombre de fois où le point tiré se trouve à l'intérieur du cercle au nombre total de tirages.

Cette simulation met en évidence trois particularités de ces méthodes :

- Une convergence très lente. En effet, on constate que des milliers de tirage sont nécessaires pour une estimation satisfaisante.

- L'oscillation du résultat pendant et surtout en fin de convergence.

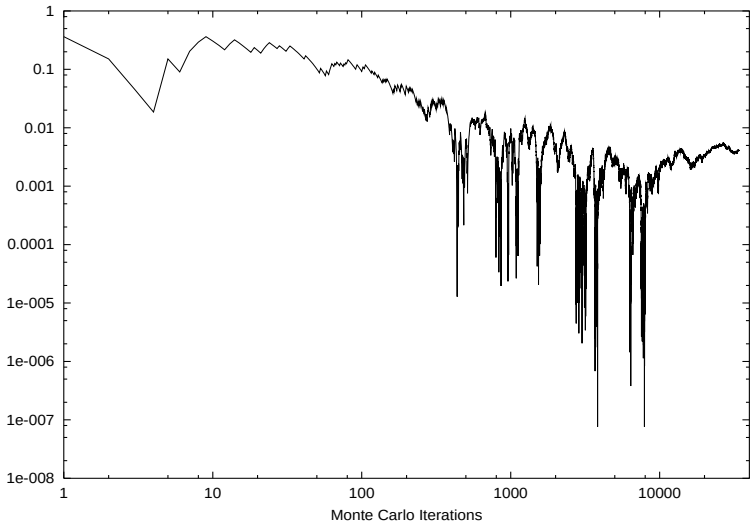


FIGURE 19.2 – Application de la méthode de Monte Carlo à l'estimation de π. Evolution de l'erreur entre la prédiction et 3.1415926.

- La possibilité cependant d'avoir une première estimation rapide mais grossière. L'erreur est réduite de deux ordres de grandeur très rapidement. Ce qui permet une estimation de π entre 3.11 et 3.18 en une dizaine de tirages.

- L'accumulation des erreurs numériques dégrade les résultats (observer l'augmentation de l'erreur pour un grand nombre d'itérations).

En général, les méthodes Monte Carlo sont utilisées si on ne peut ou on ne veut pas, par exemple lorsqu'aucun modèle filtré pertinent n'existe, introduire un filtre au niveau des modèles (voir ci-dessous).

19.4 Filtrage

Très généralement, un filtre est un opérateur qui, appliqué à une fonction, produit une autre fonction pour laquelle certaines caractéristiques ont été éliminées. Nous présentons plusieurs filtres et leurs avantages et inconvénients. Nous ne donnerons pas une vue exhaustive du filtrage.

Le but de l'introduction des filtres lors de la résolution des EDP est l'utilisation d'une discrétisation plus grossière que pour une simulation directe (sans filtrage). De plus, parfois on souhaite obtenir des quantités moyennes sans faire nécessairement appel à une moyenne d'ensemble obtenue par une simulation de type Monte Carlo où, comme nous l'avons vu, on utilise la loi des grands nombres pour obtenir, après un grand nombre de simulations, les quantités moyennes.

Considérons la décomposition suivante d'un champ u en une partie observable $< u >$ et une partie non-observable ou non calculable sur la discrétisation utilisée u' :

$$u =< u > +u'.$$

19.4.1 Moyenne d'ensemble

Une simulation directe de l'équation (19.1) n'est possible que si l'on peut disposer de discrétisations assez fines. Les quantités moyennes, qui sont souvent les quantités que l'on cherche à connaître dans les applications, sont alors accessibles par moyenne d'ensemble. La moyenne d'ensemble est définie par :

$$< u > (x,t) = \frac{1}{N} \sum_{i=1}^{N} u_i(x,t).$$

Par exemple, pour connaître l'évolution moyenne en temps et espace d'une rivière, on peut filmer plusieurs fois sur des périodes différentes l'écoulement. En moyennant, en chaque point et chaque instant l'ensemble des observations, on aboutit à un comportement moyen en temps et en espace de la rivière. Ceci requiert un grand nombre d'évaluations pour converger si les fluctuations sont importantes. On présente figure (19.2) un exemple d'une telle simulation pour le calcul du nombre π par une méthode de Monte Carlo.

19.4.2 Convolution

Le filtre le plus répandu est basé sur la convolution en temps et/ou en espace avec une fonction porte :

$$< u > (x,t) = \frac{1}{TV} \int_{t-T/2}^{t+T/2} \int_{B(x,r)} u(s,\tau) dx d\tau,$$

où T est la fenêtre temporelle du filtrage et $B(x,r)$ est une boule de centre x, de rayon r et de volume V. La difficulté de mise en oeuvre concerne les points dont la distance aux bords est inférieure à la fenêtre du filtre. Une solution consiste alors à réduire la fenêtre de filtrage en fonction de la distance à la frontière.

19.4.3 Filtre en fréquence

On peut aussi utiliser un filtre de Fourier et tenter d'identifier la partie basse-fréquence de u. Ainsi si $F(u)_k$ est la transformée de Fourier de u :

$$F(u)_k(t) = (\frac{1}{2\pi})^3 \int_{R^3} u(x,t) e^{-ik.x} dx,$$

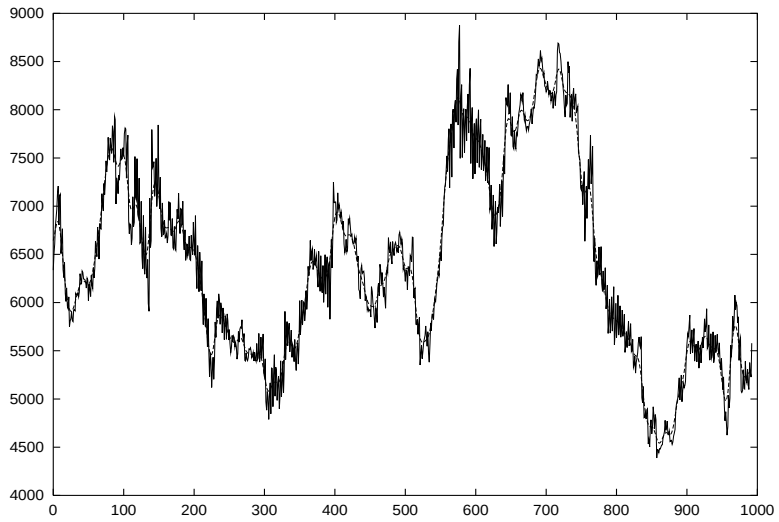


FIGURE 19.3 – Application d'un filtre de convolution adaptatif avec une fonction porte étroite à un signal stochastique comportant à la fois de grandes et de très petites échelles. Sur ce signal, une double application de la convolution laissera le signal presque invariant (on utilisera en pratique le terme de moyenne glissante).

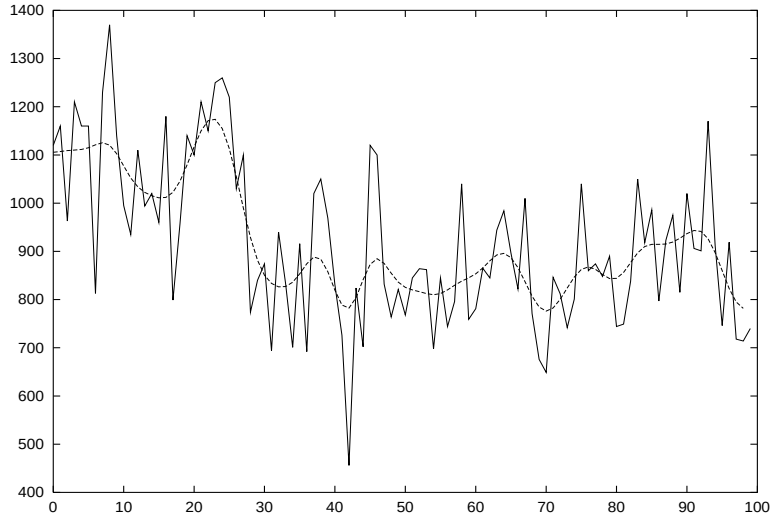


FIGURE 19.4 – Application du même filtre de convolution adaptatif avec une fonction porte large à un signal ayant une moindre dispersion en fréquence. Une double application de la convolution ne laissera pas le signal filtré invariant car l'écart fréquentiel entre le signal filtré et le bruit n'est pas assez important.

et π_N l'opérateur de troncature qui remplace $\sum_{k_i=0,1,...}$ par $\sum_{k_i=0,1,...,N}$ alors

$$< u >_F = F^{-1}\pi_N F(u) = \sum_{|k|\leq N} F(u)_k e^{ik.x}$$

est un filtre de Fourier passe-bas.

19.4.4 Quelques propriétés des filtres

Linéarité :
Les filtres sont en général linéaires.

$$< u + \lambda v >=< u > +\lambda < v >$$

Commutation avec la dérivation :
Parfois les opérateurs de dérivation et filtre commutent. Par exemple, le filtre de convolution temporelle commute avec les dérivées temporelle et spatiale (sous réserve de régularité suffisante pour que toutes les opérations soient définies) :

$$< \partial_x u >= \frac{1}{T}\int_{t-T}^{t} \partial_x u(x,\tau)d\tau = \partial_x \frac{1}{T}\int_{t-T}^{t} u(x,\tau)d\tau$$

$$< \partial_t u >= \frac{1}{T}\int_{t-T}^{t} \partial_t u(x,\tau)d\tau = \frac{1}{T}[u(x,t) - u(x,t-T)]$$

$$\partial_t < u >= \partial_t \frac{1}{T}\int_{t-T}^{t} u(x,\tau)d\tau = \frac{1}{T}[u(x,t) - u(x,t-T)].$$

Invariance :
On souhaite qu'un champ filtré soit invariant par le même filtre :

$$<< u >>=< u > .$$

Ceci est bien sur le cas pour le filtre de Fourier :

$$<< u >_F >_F = F^{-1}\pi_N F F^{-1}\pi_N F(u) = F^{-1}\pi_N F(u) =< u >_F .$$

Ceci est un des avantage des filtres en fréquence. De même la moyenne d'ensemble est invariante.

Il est à noter que le filtre de convolution ne satisfait pas cette propriété :

$$<< u >>= \frac{1}{|B|^2}\int_{B(x,r)}\int_{B(z,r)} u(y,t)dydz \neq \frac{1}{|B|}\int_{B(x,r)} u(y)dy.$$

En fait, cette propriété est presque satisfaite par la convolution si un écart fréquentiel important existe entre les fluctuations et la quantité moyennée (comme dans la figure (19.3) mais pas pour la figure (19.4)).

La raison pour laquelle l'invariance du filtre est recherchée est la suivante. Si un filtre n'est pas invariant (i.e. $<< u >>\neq< u >$), alors $u =< u > +u'$ n'implique pas $< u' >= 0$. On verra que cette propriété est importante lors de la dérivation d'équations pour les quantités moyennées en partant de l'équation initiale. On veut que les équations des quantités moyennées gardent une forme proche de celle des équations initiales, en particulier pour utiliser les outils numériques déjà développés.

Filtrage d'un produit : Une propriété importante des filtres concerne leur comportement pour les produits de variables fluctuantes. On souhaite en général la propriété suivante :

$$< v < u >>=< v >< u > .$$

Le seul filtre vérifiant cette propriété est la moyenne d'ensemble. En effet, la convolution ne laissant pas le champ invariant après filtrage, ne peut vérifier cette propriété (prendre $v = 1$). De même, le filtre de Fourier n'est pas satisfaisant. En effet, considérons $u = e^{i(\omega_1+\omega_2)t}$, avec $\omega_2 = 2\omega_1$ et un filtre de Fourier passe-bas pour les fréquences $\omega < 1.5\omega_1$. On a,

$$< u < u >>=< e^{i(2\omega_1+\omega_2)t} >= 0 \neq< u >< u >= e^{2i\omega_1 t}.$$

19.4.5 Le modèle filtré

On sépare les variables en partie déterministe calculable et partie stochastique ou non calculable par l'approche courante (dans la suite on note $< u >$ par $\overline{u}$).

$$u = \overline{u} + u', \quad \overline{u'} = 0.$$

Si le filtre n'est pas invariant, nous aurons des termes supplémentaires à modéliser :

$$\overline{u'}_t\,, \quad \nu\overline{u'}_{xx}\,, \quad \overline{w'}.$$

L'équation décrivant $\overline{u}$ est obtenue à partir de (19.1) :

$$\overline{u}_t + 0.5(\overline{u}^2)_x - \nu\overline{u}_{xx} = \overline{w} - 0.5(\overline{u'^2})_x, \tag{19.2}$$

$$\overline{u}(t,0) = u_l, \quad \overline{u}(t,L) = u_r, \quad \overline{u}(0,x) = u_0,$$

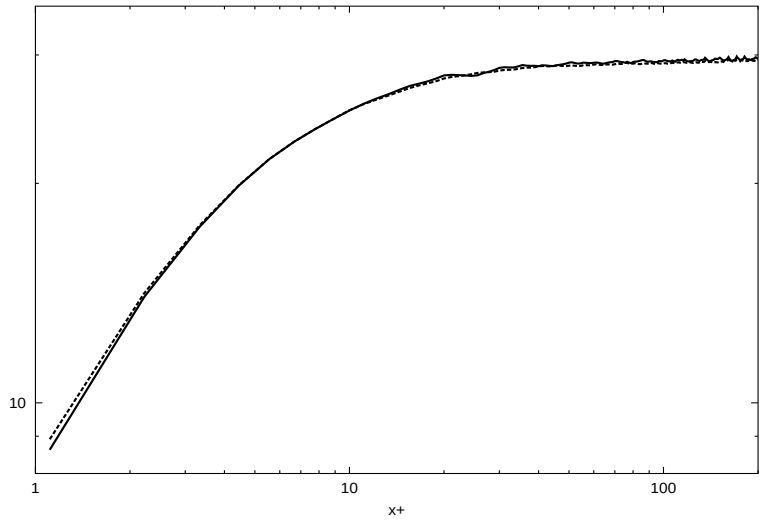


FIGURE 19.5 – Moyenne d'ensemble et filtre de convolution aboutissent presque à la même solution. Cette propriété s'appelle l'ergodicité.

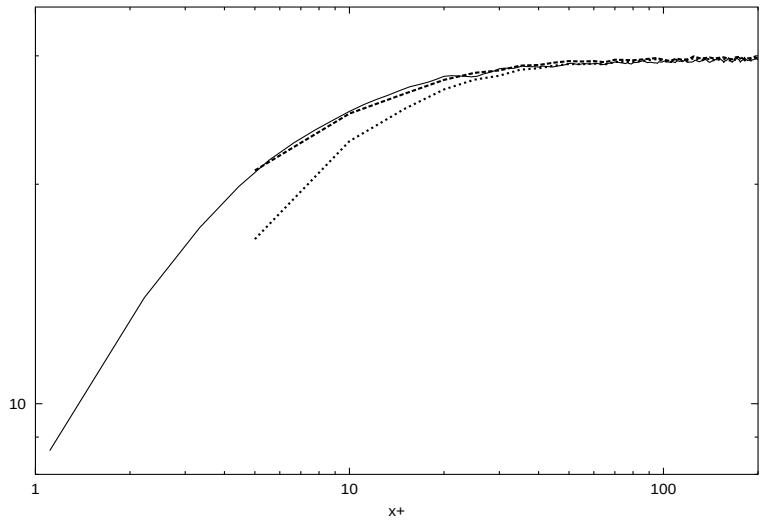


FIGURE 19.6 – Simulation directe sur deux maillages fin et grossier et comparaison avec une simulation sur le maillage grossier utilisant l'équation filtrée et fermée (19.3).

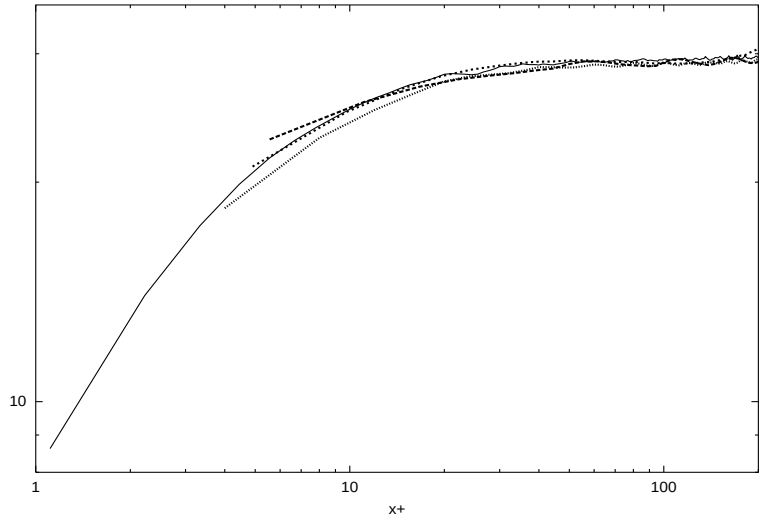


FIGURE 19.7 – Simulation avec le modèle filtré et fermé pour trois discrétisations différentes.

19.4.6 Hypothèse de clôture

Pour clore le modèle (19.2), il faut exprimer les nouvelles quantités $\overline{u'^2}$ et $\overline{w}$ par rapport aux quantités connues (i.e. déterministes). Nous supposons que l'advection par un champ fluctuant est similaire à l'ajout d'une nouvelle viscosité ν_t à la viscosité ν. L'expression de cette nouvelle viscosité est déduite des dimensions physiques des quantités impliquées $\nu_t \sim UL$ où U a la même dimension que $\overline{u}$ et L est une échelle de longueur. Constatant que ces modifications se produisent principalement aux endroits où $\overline{u}_x$ est grand, nous remarquons, qu'une bonne fermeture est obtenue avec le modèle suivant :

$$\overline{u'^2} = -\nu_t \overline{u}_{xx}, \quad \nu_t = -c^+ h^2 |\overline{u}_x| \quad \text{avec} \quad c^+ = \frac{1}{x^+}. \tag{19.3}$$

où $x^+ = x u_f/\nu$ est la variable adimensionnée.
Une amélioration de la modélisation est nécessaire au voisinage de $x^+ = 0$ c'est à dire dans la couche limite. On constate que les gradients du champ moyen sont plus forts au voisinage de ce point. Ceci suggère une modélisation différente, à travers c^+ par exemple, pour cette région, où une anisotropie manifeste est présente. La figure suivante montre l'effet de c^+ sur les résultats. On aboutit à l'équation suivante pour $\overline{u}$:

$$\overline{u}_t + 0.5(\overline{u}^2)_x - ((\nu + \nu_t)_+ \overline{u}_x)_x = \overline{w}, \tag{19.4}$$

avec

$$\overline{u}(t,0) = u_l, \quad \overline{u}(t,L) = u_r, \quad \overline{u}(0,x) = u_0,$$

et où $(\nu + \nu_t)_+ = \max(0, \nu + \nu_t)$. Si les fluctuations sont totalement filtrées, on a $\overline{w} = 0$.

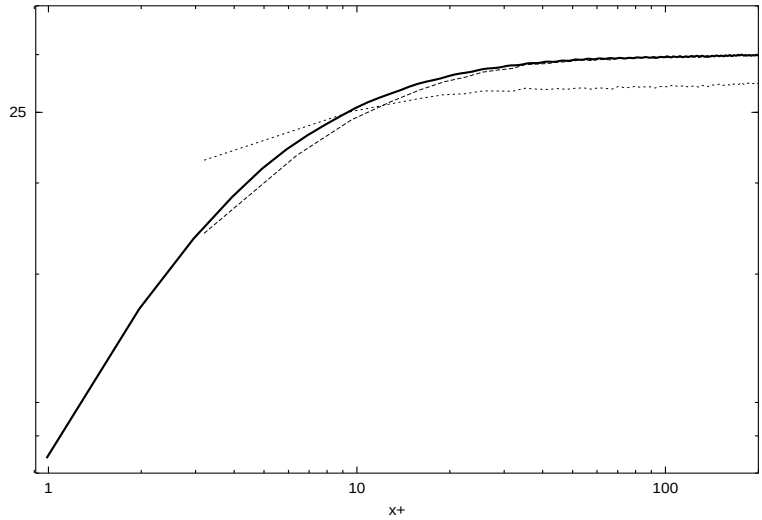


FIGURE 19.8 – Simulation avec le modèle filtré et fermé avec $c^+ = 1/x^+$ et $c^+ = 1$ et comparaison avec la simulation directe en utilisant la moyenne d'ensemble. Ceci montre l'importance d'une modélisation différente dans les couches limites.

Remarque 19.4.1 *La modélisation a introduit une viscosité supplémentaire ν_t négative. La prise en compte des fluctuations sur le champ moyen par un changement de viscosité s'interprète physique en observant la diffusion d'une goutte d'encre dans un fluide au repos, puis agité. Nous avons déjà rencontré cette modélisation dans l'équation de Black et Scholes, vu au chapitre 12, où l'effet des fluctuations est pris en compte dans la viscosité (volatilité) dans l'équation.*

19.4.7 Application en mécanique des fluides

En simulation d'écoulements turbulents, on peut faire appel à des simulations directes (Direct Numerical Simulation) qui consistent à faire le calcul de toutes les échelles de l'écoulement. Ces calculs nécessitent des discrétisations précises en temps et en espace et sont coûteux. On utilise alors la moyenne d'ensemble pour obtenir le comportement moyen de l'écoulement. On peut aussi tenter de calculer directement les variables moyennées en introduisant le filtrage dans les équations de l'écoulement. Ceci est à la base de la simulation des grandes structures (Large Eddy Simulation) où uniquement une partie des échelles de l'écoulement est calculé et où les effets de la partie manquante sur la partie calculée sont modélisés. Suivant le degré de filtrage utilisé, la discrétisation doit être plus ou moins fine.

19.4.8 Conditions aux limites équivalentes

Nous avons vu que les gradients sont concentrés au voisinage de $x = 0$ et ceci même pour le modèle filtré. Le calcul de cette région exige donc un maillage plus fin qu'à l'intérieur du domaine. De plus, l'épaisseur de cette zone se réduit quand ν diminue. Le maillage doit donc être adapté et plus fin proche de $x = 0$. Le but des conditions aux limites équivalentes est de pouvoir s'affranchir de cette difficulté et ne pas trop raffiner les maillages. On applique ces conditions en $x = \delta$ et le domaine de calcul est réduit à (δ, L).

Un développement de Taylor pour u nous donne :

$$\overline{u}(0) = \overline{u}(\delta) - \overline{u}_x(\delta)\delta + \overline{u}_{xx}(\delta)\frac{\delta^2}{2} + \delta^2 o(1),$$

qui à l'ordre 1 se réduit à :

$$\overline{u}(\delta) = u_l + \overline{u}_x(\delta)\delta. \tag{19.5}$$

Ceci est une condition aux limites de Fourier pour $\overline{u}$ en $x = \delta$. Connaissant $\overline{u}_x(\delta)$ on peut calculer la solution en tous points. Nous avons donc besoin d'un modèle à complexité réduite sur $(0, \delta)$ pour exprimer $\overline{u}_x(\delta)$ connaissant la solution sur (δ, L). Ce modèle doit être compatible avec le filtrage utilisé.

Considérons l'équation différentielle suivante pour $A = \overline{u}_x(\delta)$, obtenue à partir de (19.5) et de l'équation filtrée (19.4) :

$$\delta A_t = ((\nu + \nu_t)_+ A)_x - (u_l + A)A + \overline{w}, \quad \text{avec} \quad \nu_t = -c^+ h^2 |A|. \tag{19.6}$$

Le couplage entre cette équation différentielle et le modèle défini sur (δ, L) est effectué par :

$$((\nu + \nu_t)A)_x = ((\nu + \nu_t(\delta + h))u_x(\delta + h) - (\nu + \nu_t(\delta))A)/h$$

ainsi que par la condition (19.5) en $x = \delta$. On constate que le modèle à complexité réduite (19.6) dépend du filtrage par l'intermédiaire de $\overline{w}$. En pratique, les fluctuations sont inconnues sur $(0, \delta)$, on considère un modèle réduit après application d'un filtre vérifiant $< \overline{w} >= 0$, correspondant à un filtrage total des fluctuations. La compatibilité entre les solutions sur $(0, \delta)$ et (δ, L) exige que :

$$< \overline{u} > (\delta) = u_l + A\delta. \tag{19.7}$$

Ceci s'accompagne d'une redéfinition de ν_t.

Pour faciliter le couplage des modèles, on préfère disposer d'une condition au limite en $x = 0$, plutôt qu'en $x = \delta$. Ceci permettra en effet de considérer le modèle global directement sur $(0, L)$ plutôt que sur (δ, L).

Nouvelle modélisation au voisinage de $x = 0$ et couplage de modèles

L'approche précédente n'est valide que si δ reste petit, pour que la condition de Fourier reste valide. Pour s'affranchir de cette contrainte on introduit un nouveau modèle complètement filtré valide sur un voisinage plus grand autour de $x = 0$. On notera δ l'épaisseur de cette région qui n'est plus nécessairement petite. On souhaite pouvoir résoudre ce modèle plus facilement que le modèle initial. Cette modélisation s'effectue en général en modifiant l'expression de la viscosité (19.3) :

$$\overline{u'^2} = -\nu_t \overline{u}_x, \quad \frac{\nu_t}{\nu} = -Cx^+(1 - \exp(-x^+/D))^2, \tag{19.8}$$

où les constantes C et D doivent être calées numériquement pour assurer la continuité entre les modèles.

En retenant le terme dominant dans l'équation (19.4) au voisinage de $x = 0$ (la dérivée seconde) on obtient le modèle partiel ci-dessous sur $[0, \delta]$:

$$-\left((\nu + \nu_t)_+ \overline{u}_x\right)_x = 0, \quad \text{sur} \quad [0, \delta], \tag{19.9}$$

qui permet de calculer $A_0 = \nu \overline{u}_x(0) = (\nu + \nu_t)_+ \overline{u}_x(\delta)$ en utilisant $u_{(0,L)}(\delta)$ fourni par le modèle complet évalué sur $[0, L]$:

$$A_0 = \nu \overline{u}_x(0) = \frac{u_{(0,L)}(\delta)}{\int_0^\delta \frac{dx}{(\nu+\nu_t)_+}}.$$

La modification de la viscosité est destiné à permettre un calcul aisé de l'intégrale ci-dessus. La résolution de ce nouveau modèle est très peu coûteuse.

On utilisera alors un des algorithmes de couplage entre modèles vus au chapitre 13.

Enfin, on souhaite que les deux modèles utilisés soient compatibles au niveau du filtrage comme pour (19.7), non seulement en $x = 0$, mais sur $(0, \delta)$. Cette condition peut être formulée à travers la solution d'un problème de contrôle optimal.

19.5 Utilisation du contrôle optimal

Considérons le couplage de deux modèles filtrés définis respectivement sur $(0, \delta)$ et $(0, L)$. Comme nous l'avons dit, ici δ n'est pas nécessairement petit et désigne l'épaisseur de recouvrement entre les deux modèles. Cependant, la solution du modèle partiellement filtré doit correspondre en moyenne à celle du modèle totalement filtré sur le recouvrement $(0, \delta)$. Ceci doit être pris en compte dans la condition aux limites $A_0(t) = \nu \overline{u}_x(0)(t)$ fournie au modèle partiellement filtré.

L'observation faite au chapitre 17 (paragraphe 17.9.1) sur le choix de la fonctionnelle dans un problème inverse suggère un problème de minimisation avec une fonctionnelle J de la forme :

$$J(A_0(t)) = \int_0^\delta (\overline{u_{(0,L)}(x,t)} - u_{(0,\delta)}(x))^2 dx + \int_0^\delta |\overline{u_{(0,L)}(x,t)} - u_{(0,\delta)}(x)|^q dx. \quad (19.10)$$

avec $0 < q < 1$.

Ce problème de contrôle peut être résolu par différentes approches comme par exemple :

— En utilisant un algorithme de contrôle optimal utilisant le gradient :

$$A_0(0) = \text{donné}, \quad \delta A_0^{n+1} = A_0^n - \rho \frac{dJ}{dA_0}(t^n), \quad \rho > 0. \quad (19.11)$$

Dans ce cas, le contrôle étant scalaire, on utilise la méthode des variables complexes, vue au chapitre 17 :

$$\frac{dJ}{dA_0}(t) = \frac{Im(J(A_0 + i\varepsilon))}{\varepsilon},$$

pour le calcul de J et dJ/dA_0 en même temps.

— En utilisant un algorithme de contrôle sub-optimal, applicable aux configurations générales, où le nombre de contrôle est grand et où l'approche ci-dessus n'est pas applicable.
Par exemple, en définissant un modèle à complexité réduite $\tilde{J}$ pour $J(A_0(t))$ basé sur une relation a priori entre la fonctionnelle et les paramètres de contrôle. On peut ainsi définir le contrôle en utilisant le gradient de ce modèle réduit :

$$\frac{dJ}{dA_0}(t) \sim \frac{d\tilde{J}}{dA_0}(t).$$

Ceci est similaire à ce qui a été fait en optimisation avec les gradients incomplets (voir chapitre 17). (19.9).

Chapitre 20

Calcul parallèle et simulation

20.1 Introduction

L'idée du calcul parallèle est simple. Elle consiste à répartir sur plusieurs processeurs d'une même machine ou sur plusieurs ordinateurs en réseau (grappe ou cluster) le calcul ou la simulation que l'on veut faire. Le calcul parallèle a eu une grande évolution au cours des deux dernières décennies. Évolution fortement liée à celle des machines et surtout à celle des mémoires vives. En règle générale, on observe une évolution vers une parallélisation à grain de plus en plus gros (i.e. on répartit des paquets de plus en plus importants d'instructions entre processeurs ou plutôt entre ordinateurs en réseau).

On verra au travers des étapes effectuées par le calcul parallèle que la recherche de la simplicité a toujours été l'élément moteur, notamment pour permettre une plus large utilisation de l'approche, et aussi assurer la portabilité des codes.

La parallélisation peut même servir d'outil d'optimisation, permettant l'utilisation de méthodes de minimisation robustes mais simples exigeant un grand nombre d'évaluations indépendantes de fonctionnelles. On présentera un exemple d'une telle optimisation dans ce chapitre.

Le but de ce chapitre est de donner une vue générale de la parallélisation plutôt que des détails techniques.

20.2 Parallélisation des instructions

Au début des années 80, la mode était plutôt à la parallélisation des instructions simples. Ce mode s'appelle aussi "data parallel". Ce type de parallélisation était une suite naturelle à la vectorisation. Le modèle SIMD : Single Instruction Multiple Data concerne donc l'éxécution d'une instruction simple sur un nombre

important de processeurs ayant chacun sa mémoire. Des machines comme l'Hypercube ou la Connection Machine 200 fonctionnent sur ce schéma. Sur ces machines la mémoire est distribuée au niveau des processeurs. Ce modèle est actuellement abandonné au profit d'une parallélisation "à grain" de plus en plus gros et la parallélisation des instructions simples est traitée au niveau des processeurs de base modernes de façon transparente pour l'utilisateur. C'est ce qui se passe dans un processeur de PC à 2 Ghz aujourd'hui. Ainsi, la gestion de la parallélisation de type petit grain, concernant un petit nombre d'instructions simples, revient de plus en plus au compilateur.

20.3 Parallélisation des séquences

L'arrivée des machines disposant de moins de processeurs, mais chacun avec plus de mémoires dédiées a permis la parallélisation de séquences d'instruction. Les paquets d'instructions étant effectués sur les données distribuées (`CM5`) ou partagées (`CRAY YMP` multi-processeur), voire mixte (comme sur la `KSR`). Cela marque le début d'une certaine portabilité pour la parallélisation. En particulier, avec l'utilisation des bibliothèques de communication comme `MPI`.

Prenons un exemple très simple en dimension 1. Considérons la parallélisation du problème d'advection-diffusion suivant :

$$u_t + au_x - \nu u_{xx} = 0, \quad \text{sur} \quad [0,1] \quad \text{munie des CL et CI appropriées,} \qquad (20.1)$$

sur deux sous-domaines :

$$S_1 = [0, 0.5], \quad S_2 = [0.5, 1].$$

Supposons que les équations soient discrétisées par une approche différences finies explicite (en temps) et centrée (en espace) sur un maillage uniforme :

$$u_i{}^{n+1} = u_i^n + \Delta t(-a(u_x)_i^n + \nu(u_{xx})_i^n). \qquad (20.2)$$

Ainsi, à chaque itération, nous devons produire une approximation des dérivées première et seconde. Ceci est facile en mono-domaine, mais introduit des communications lorsque les informations ne sont pas disponibles. Si le noeud i appartient à l'interface entre les sous-domaines S_1 et S_2, pour le calcul de

$$u_{x_i}{}^n = \frac{u_{i+1}^n - u_{i-1}^n}{2h}$$

dans le sous-domaine S_1, on aura besoin de u_{i+1} du noeud $i+1$ du sous-domaine S_2 et inversement, pour le calcul de la dérivée approchée sur S_2, de u_{i-1} du noeud $i-1$ du sous-domaine S_1. Le calcul de la dérivée approchée centrée aux interfaces

nécessite donc deux communications. Les mêmes informations seront utilisées pour la dérivée seconde. Une remarque importante est que cela rend les calculs nécessairement synchrones : tous les processeurs doivent être à la même itération en temps. Ce qui introduit une nouvelle difficulté concernant l'équilibrage des charges des processeurs, comme on le verra plus loin.

L'opération qui consiste à rassembler les informations nécessaires pour le calcul s'appelle `Gather`. Une fois le calcul effectué, il faut parfois redistribuer le résultat aux noeuds. C'est l'opération inverse qui s'appelle `Scatter`.

Considérons une itération en temps de (20.2) avec une écriture élément par élément (pour simplifier on considère des conditions aux limites de Dirichlet ou Neumann homogènes) :

$$u_i{}^{n+1} = u_i^n - \frac{\Delta t}{h}\left(\sum_{T_j \in \tau_i} \int_{T_j} (a(u_x)^n w_i + \nu(u_x)^n (w_i)_x) dx\right), \tag{20.3}$$

où $\tau_i = \{T_j, j = 1, ..., nel \text{ tel que } i \in T_j\}$ regroupe les éléments ayant le noeud i en commun, w_i est la fonction de base associé à i sur l'élément T_j. Nous avons utilisé une formule d'Euler explicite pour la dérivée temporelle. L'évaluation parallèle de la boucle sur les éléments ne nécessite aucune communication car chaque élément T_j n'appartient qu'à un sous-domaine et dispose localement des informations nécessaires. Nous n'avons donc pas d'étape `Gather`. Mais, une fois l'évaluation effectuée dans chaque sous-domaine, il faut communiquer ce résultat à l'autre sous-domaine pour l'assemblage pour compléter la boucle sur les éléments ci-dessus. Remarquons qu'en dimension supérieure, un noeud peut être partagé par un grand nombre de sous-domaines.

20.3.1 Recouvrement de domaines

La solution la plus simple pour éviter les communications est d'introduire un recouvrement des domaines adjacents. Ainsi, en dédoublant les points nécessaires à la construction des approximations des dérivées, l'information sera disponible en local. Le problème vient alors des calculs redondants pour les points dédoublés. Bien entendu, la redondance augmente avec la précision demandée, car alors le nombre de points intervenant dans les approximations augmente. On préférera donc des schémas numériques compacts, ne mettant en jeu que des informations locales et des structures de données simples, avec en contre-partie une plus grande exigence sur la qualité des maillages (voir chapitre 18).

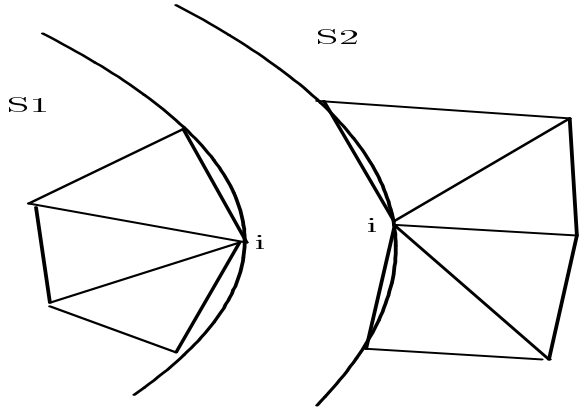


FIGURE 20.1 – Interface entre deux sous-domaines dans un maillage triangulaire. Le noeud i est à l'interface de S_1 et S_2. En éléments finis, après l'évaluation des contributions des éléments dans chaque sous-domaine, il faut communiquer ce résultat à l'autre sous-domaine pour l'assemblage.

20.3.2 Numérotation locale-globale

Une autre difficulté provient de la gestion de la correspondance par numérotation locale (à l'intérieur du sous-domaine) - globale (domaine entier). Ces éléments font que les codes parallèles deviennent rapidement plus difficiles à gérer et à maintenir.

20.3.3 Simplification des structures de données

Minimiser les communications est la clé principale pour réaliser la scalabilité dans un solveur parallèle. La scalabilité, que nous avons définie au chapitre 1, mesure le rapport entre la progression en nombre de processeurs et l'accélération qui en découle. Une progression linéaire avec une pente de 1 serait bien entendu optimale, mais celle-ci n'est jamais obtenue à cause des communications et opérations redondantes.

Si une parallélisation automatique du code est souhaitée, la simplification des structures de données devient inévitable. Ceci est une des raisons pour lesquelles les techniques de différences finies sont plus facilement parallélisables, car elles ne nécessitent qu'une seule structure de données : le noeud. Tandis qu'en éléments ou volumes finis on a besoin de structures de données relationnelles pour les noeuds, arêtes ou faces et éléments. Ceci rend impossible une parallélisation automatique sans la définition par l'utilisateur des sous-domaines et de leurs inter-connections.

20.3.4 Partition de domaines sans recouvrement et adaptation de maillage

L'efficacité d'une exécution en parallèle dépend, par ailleurs, de l'équilibrage de la charge entre les processeurs et de la minimisation des communications. Cela demanderait un découpage du domaine tel que les interfaces entre les sous-domaines soient minimales et que les sous-domaines aient environ le même nombre de degrés de liberté chacun. Actuellement, aucun algorithme n'est disponible pour une gestion efficace des équilibrages de charges en fonction des changements de discrétisation dans les sous-domaines.

Voici les étapes d'un algorithme de calcul parallèle avec maillage adaptatif :
— génération d'un maillage grossier ayant une description précise de la géométrie,
— découpage en sous-domaines de ce maillage grossier (faible complexité),
— boucle d'adaptation de maillage en parallèle :
  1. adaptation de maillage parallèle, conforme aux sous-domaines (i.e. ayant la même discrétisation des deux côtés de l'interface), en utilisant les algorithmes décrits au chapitre 18,
  2. solution parallèle, sur chaque sous-domaine, du modèle avec transfert d'information aux interfaces entre sous-domaines.

Cet algorithme risque d'aboutir assez rapidement à un déséquilibre important entre sous-domaines, ce qui rendra le calcul parallèle inutile. En effet, les sous-domaines les plus rapides auront, de toute façon, à attendre le résultat des sous-domaines les plus chargés. C'est une des raisons pour lesquelles l'adaptation de maillage est rarement utilisée en calcul parallèle. En règle générale, en calcul parallèle, on préférera utiliser des méthodes plus simples, demandant peut être plus de points, souvent mal répartis, plutôt que de relancer les calculs après un redécoupage du domaine après adaptation de maillage. Voila un bon exemple de la non-unicité de la démarche possible en simulation numérique. Les deux démarches étant justifiables, le choix se fera d'après des critères propres à chaque application. On préfère cependant de plus en plus réserver le calcul parallèle au traitement de problèmes indépendants, dont la solution intervient dans l'analyse d'un problème global. On en verra un exemple plus bas lors de l'utilisation du calcul parallèle pour la résolution de plusieurs milliers de problèmes de taille moyenne en optimisation.

20.3.5 Parallélisation en temps

Nous avons vu comment paralléliser les boucles d'intégration spatiale lors de la résolution d'une EDP par une des méthodes différences, volumes ou éléments

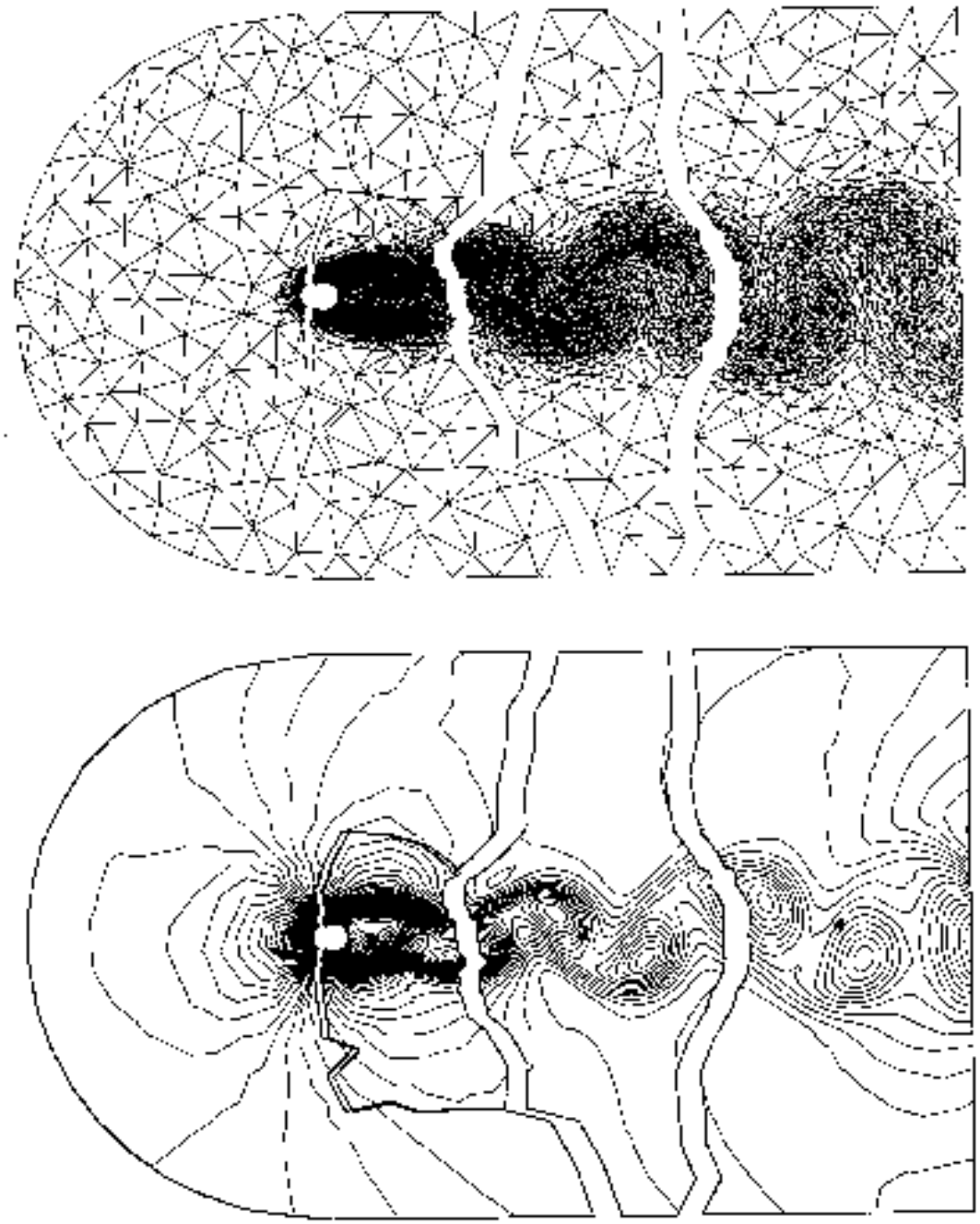


FIGURE 20.2 – Calcul adaptatif parallèle en mécanique des fluides numérique sur un cluster de stations en utilisant le passage de messages par `MPI` (4 sous-domaines). On utilise l'algorithme décrit au chapitre 18. On représente le maillage adapté et les lignes iso-densité au même instant. Le découpage du domaine indique les sous-domaines. On constate qu'après adaptation, le sous-domaine de gauche se trouve avec une densité de points bien inférieure à celle des autres sous-domaines. (Calcul réalisé par E. Saltel à l'INRIA.)

finis. La parallélisation est alors basée sur une décomposition du domaine en sous-domaines avec ou sans recouvrement entre les sous-domaines et communication d'informations entre les sous-domaines aux interfaces dans le second cas. Cette approche ne convient pas dans le cas d'une parallélisation de la résolution temporelle car on ne dispose pas, initialement, des solutions intermédiaires nécessaires au raccord à l'interface des sous-domaines en temps.

Considérons, par exemple, l'équation

$$u'(t) = f(u(t)), \quad u(0) = u_0, \tag{20.4}$$

que l'on voudrait intégrer entre $t = 0$ et $t = T$. On parallélise cette intégration en temps sur deux sous-domaines temporels $[0, T/2]$ et $[T/2, T]$. On propose une méthode de point-fixe parallèle utilisant un algorithme de tir présenté aux chapitres 2 et 17 pour trouver une condition initiale v pour la deuxième partie de l'intégration. Cette condition doit permettre de réaliser le raccord entre les deux parties de l'intégration. Notons $u^{(1)}$ la solution sur le premier sous-domaine temporel $[0, T/2]$ et $u^{(2)}$ la solution sur $[T/2, T]$.

Considérons la fonctionnelle $J(v) = |u^{(1)}(T/2) - v| + |u^{(2)'}(T/2) - u^{(1)'}(T/2)|$, qui mesure la qualité du raccord lors de l'intégration en temps parallèle ci-dessous :

$$u^{(1)'}(t) = f(u^{(1)}(t)), \quad u^{(1)}(0) = u_0, \qquad u^{(2)'}(t) = f(u^{(2)}), \quad u^{(2)}(T/2) = v, \tag{20.5}$$

Si l'intégration en temps est effectuée avec un schéma Runge-Kutta, la fonctionnelle se réduit à $J(v) = |u^{(1)}(T/2) - v|$. Ces deux équations sont résolues de façons indépendantes et en parallèle. En annulant J, la solution globale des deux problèmes parallèles sera celle de l'équation initiale. Bien entendu, on peut étendre cette décomposition à plusieurs intervalles en temps.

20.4 Parallélisation des actions

Comme nous avons dit, l'augmentation de la puissance des processeurs permet de nouvelles approches pour la simulation et l'optimisation. Nous avons vu les exemples de parallélisation de la résolution spatiale et temporelle des équations. Un autre exemple est le couplage distribué de plusieurs modèles physiques (voir chapitre 13) où l'algorithme de parallélisation temporelle ci-dessus peut être utilisé. Une autre utilisation du calcul parallèle concerne la simulation en asynchrone pour un nombre important de configurations définies par un échantillonnage des paramètres de la simulation. Ceci est très utile en optimisation par exemple, où l'on peut obtenir ainsi une bonne définition de la fonctionnelle à optimiser et trouver son minimum sans utilisation d'aucune méthode de minimisation. Cela

peut également servir à trouver une bonne initialisation pour une méthode de descente.

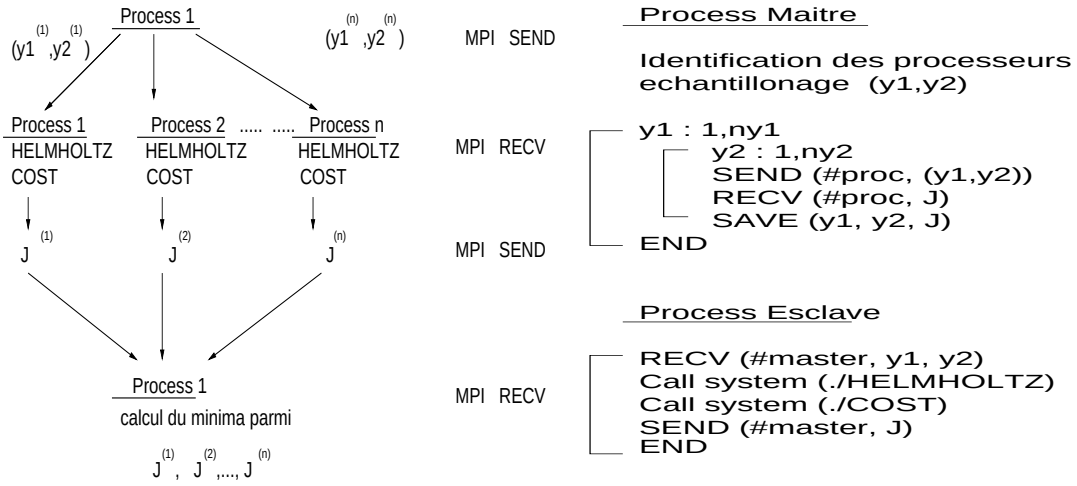


FIGURE 20.3 – A gauche, schéma de l'algorithme parallèle et à droite exemples de programmes types exécutés sur les processeurs maître et esclaves. Les appels aux programmes de résolution pour l'évaluation de l'état et de la fonctionnelle sont effectués par l'utilisation de commandes systèmes permettant l'appel d'exécutables à partir d'un exécutable.

20.4.1 Calcul parallèle et optimisation

On présente un exemple de l'utilisation du calcul parallèle en optimisation. L'objectif est de positionner des sources, ou modifier une partie de la frontière pour contrôler le champ, solution d'une EDP, dans une région particulière donnée. On s'intéresse ici à un problème vibratoire, où l'état est décrit par l'équation des ondes. Ce problème est générique et concerne par exemple le contrôle du bruit dans une enceinte, le contrôle de l'efficacité d'un four à micro-ondes (notre exemple), la conception d'un port artificiel, le contrôle des vibrations dans une pièce mécanique...

Considérons donc l'équation des ondes :

$$\frac{\partial^2 v}{\partial t^2} - c^2 \triangle v = f$$

La linéarité de cette équation rend possible une optimisation multi-point pour plusieurs ondes mono-chromatiques : $v(x,t) = u(x)exp(ikt)$, où u est solution de

l'équation de Helmholtz :

$$u + \alpha \triangle u = f, \quad \alpha = \frac{c^2}{k^2}, \tag{20.6}$$

$$u|_{\Gamma_1} = u_1, \quad u|_{\Gamma_2} = u_2, \quad (\partial_n u)|_{\Gamma_3} = 0.$$

Le domaine de calcul est un rectangle où l'onde provient de deux parties Γ_1 et Γ_2, situées sur les frontières latérales. Le problème d'optimisation concerne le positionnement de ces sources dont la dimension est pré-définie. La constante

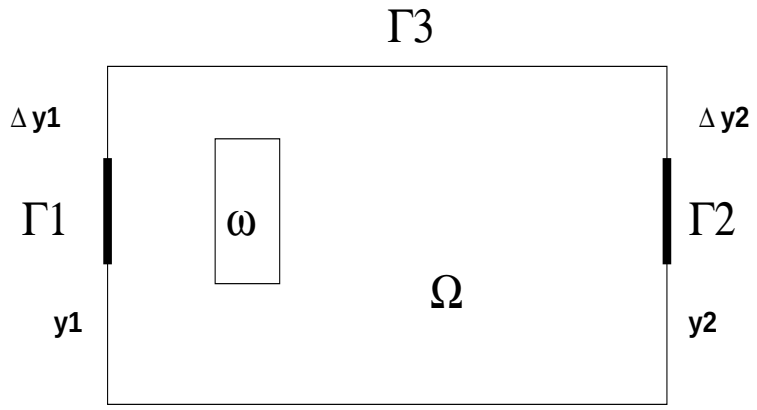


FIGURE 20.4 – Domaine de calcul et variables de contrôle.

α prend deux valeurs différentes dans ω et Ω, prenant en compte la présence du corps à réchauffer ω. La fonctionnelle à minimiser est donnée par

$$J = \frac{1}{\int_\omega u^2 dx},$$

car la température est solution de

$$-\Delta T = u^2|_\omega \quad \text{sur} \quad \Omega.$$

Et donc minimiser J aura pour conséquence une augmentation de T.

Pour minimiser cette fonctionnelle, on peut utiliser une méthode de gradient (cf chapitre 17), ou bien des méthodes nécessitant uniquement des évaluations de la fonctionnelle (par exemple une méthode basée sur la dichotomie). Ces méthodes sont plus robustes mais aussi plus coûteuses. Elles sont cependant très intéressantes si l'on dispose de moyens de calculs suffisants. Encore plus

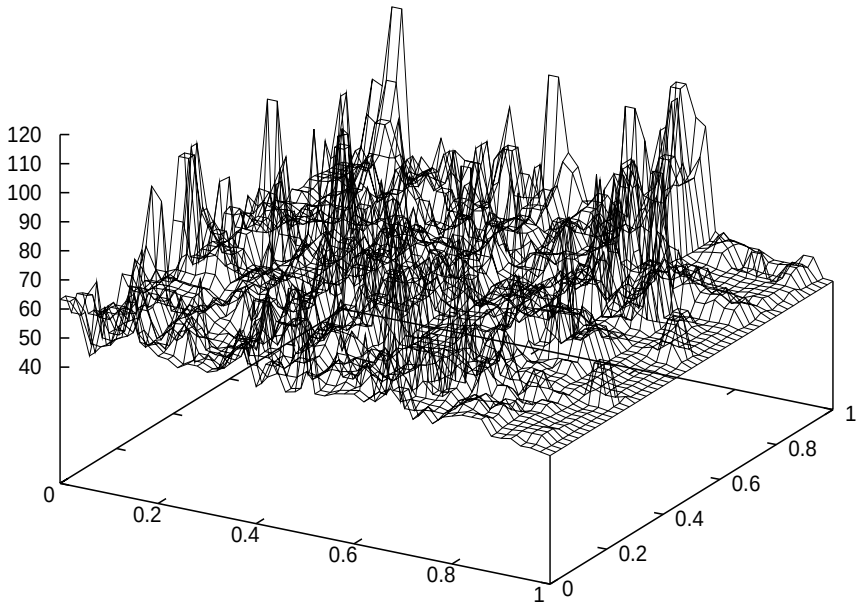


FIGURE 20.5 – Fonctionnelle obtenue sur un échantillonnage de 2500 points de l'espace de contrôle. La forme de la fonctionnelle suggère qu'une approche par gradient seul aurait probablement du mal à trouver l'optimum global, tandis qu'une formulation par problème à valeurs aux limites peut être efficace (voir chapitre 17).

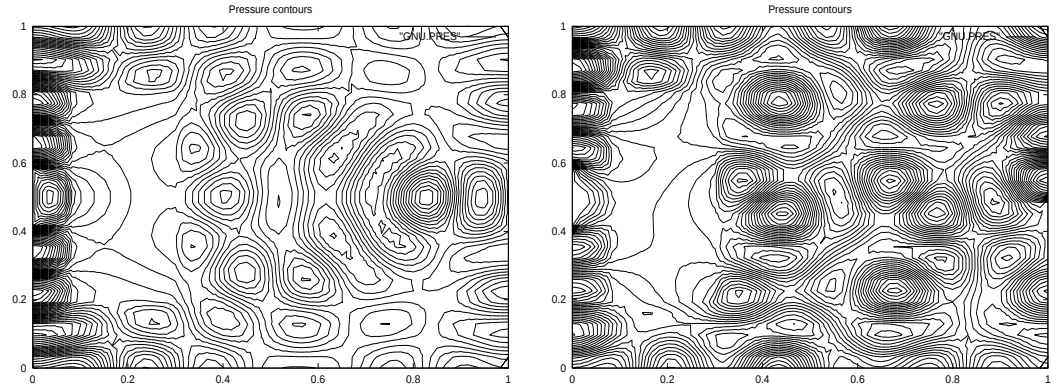


FIGURE 20.6 – Iso-valeurs de u pour les points de fonctionnement initial et optimal (Calculs réalisés par S. Galera à Montpellier).

simplement, on peut échantillonner l'espace de contrôle admissible $(y_1, y_2) \in \Pi\{y_1^1, ..., y_1^{n1d}\}\{y_2^1, ..., y_2^{n2d}\}$ où y_1 et y_2 désignent les extrémités basses des sources. Le minimum de la fonctionnelle désignera alors le couple le plus intéressant. Ainsi, en utilisant une bibliothèque de passage de messages, comme MPI, on peut résoudre (20.6) de façon indépendante, pour chaque couple (y_1^i, y_2^j). L'algorithme général est présenté figure (20.3).

Annexe A

Rappels d'algèbre linéaire

Nous nous limitons pour la suite au cas d'espaces vectoriels sur le corps des réels.

A.1 Espaces vectoriels

Définition A.1.1 (Espace vectoriel) *On appelle espace vectoriel réel E, un ensemble muni de deux lois de composition : une loi de composition interne, l'addition et une loi de composition externe, la multiplication par un scalaire réel.*

- *l'addition donne à E une structure de groupe commutatif.*
- *la multiplication associe à tout réel λ et tout vecteur (élément de E) x, un vecteur noté λx et vérifie les propriétés suivantes :*

$$\lambda(x+y) = \lambda x + \lambda y \quad (\lambda+\mu)x = \lambda x + \mu x$$

$$\lambda(\mu)x = (\lambda\mu)x \quad et \quad 1x = x$$

pour tous λ et μ réels et tous x, y de E.

A.2 Exemple fondamental : $\mathbb{R}^n$

Soit $\mathbb{R}$ le corps des réels, un élément x de $\mathbb{R}^n$ est une collection de n réels. On note

$$x = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix} \quad \text{vecteur colonne}$$

$$x^T \text{ (transposé de } x) = [\, x_1, x_2, ., ., x_n] \quad \text{vecteur ligne}$$

Les propriétés de l'addition vectorielle et de la multiplication par un scalaire réel confèrent à $\mathbb{R}^n$ une structure d'espace vectoriel réel.

A.3 Sous-espaces

Définition A.3.1 *Un sous-espace vectoriel est un sous-ensemble d'un espace vectoriel, stable pour les deux lois. C'est à dire que F est un sous-espace de E si*

$$\begin{cases} \forall x, y \in F \Longrightarrow x + y \in F \\ \\ \forall \lambda \in \mathbb{R}, \forall x \in F \Longrightarrow \lambda x \in F \end{cases}$$

A.3.1 Somme directe. Sous-espaces supplémentaires

Définition A.3.2 *On dit que deux sous-espaces vectoriels F_1 et F_2 d' un espace E sont supplémentaires, ou, ce qui est synonyme, que E est somme directe de F_1 et F_2, et l'on note :*

$$E = F_1 \oplus F_2$$

si tout élément x de E s'écrit de manière unique comme somme $x = x_1 + x_2$ d'un élément $x_1 \in F_1$ et d'un élément $x_2 \in F_2$.

Cette notion s'étend à la somme directe de plusieurs sous espaces :

$$E = F_1 \oplus F_2 \oplus F_3 \oplus ... \oplus F_{p-1} \oplus F_p$$

Exemple : Dans l'espace géométrique, on considère un plan P et une droite D n'appartenant pas au plan P. Tout vecteur V de l'espace admet une décomposition unique $V = V_1 + V_2$ avec $V_1 \in P$ et $V_2 \in D$

A.4 Dépendance et indépendance linéaire

Définition A.4.1 *p vecteurs $x_1, x_2, ...x_p$ d'un espace E sont linéairement indépendants si on a l'implication :*

$$\sum \lambda_i x_i = 0 \Longrightarrow \lambda_i = 0 \quad \forall i = 1, ..p$$

p vecteurs sont dépendants s'ils ne sont pas indépendants, donc si l'un d'entre eux peut s'écrire comme une combinaison linéaire des autres.

A.4.1 Famille génératrice

Définition A.4.2 *Une famille de vecteurs d'un espace E est dite génératrice si tout vecteur $x \in E$ peut s'obtenir comme combinaison linéaire de vecteurs de la famille. On dit aussi que la famille engendre l'espace E.*

A.4.2 Bases

Définition A.4.3 *On appelle base d'un espace E une famille génératrice formée de vecteurs linéairement indépendants.*

Soit $\{e_1, e_2, e_3, .., e_n\}$ une base de E, on a donc une décomposition unique de tout vecteur x de E sous la forme

$$x = \sum x_i e_i$$

Les scalaires réels x_i s'appellent les composantes du vecteur x dans la base $\{E_i\}$

Exemples
1) La base canonique de $\mathbb{R}^n$ est constituée des n vecteurs e_i de composantes $(e_i)_j = \delta_{ij}$ (δ_{ij} est le symbole de Kronecker : $\delta_{ii} = 1, \quad \delta_{ij} = 0$ si $i \neq j$).

$$e_i = \begin{bmatrix} 0 \\ 0 \\ : \\ 1 \\ : \\ 0 \end{bmatrix}$$

2) La base canonique de l'espace des polynômes de degré inférieur ou égal à n est constituée des $n+1$ polynômes $1, x, x^2, ...x^n$

A.4.3 Dimension

Définition A.4.4 *On appelle dimension d'un espace E, le plus petit nombre de vecteurs non nuls qui peuvent l'engendrer. Un espace est de dimension finie s'il peut être engendré par un nombre fini de vecteurs. La dimension est aussi le nombre maximum de vecteurs linéairement indépendants. C'est le nombre de vecteurs de base.*

A.5 Applications linéaires

Définition A.5.1 *Une application ϕ d'un espace vectoriel E dans un espace vectoriel F est linéaire si elle vérifie les deux propriétés de compatibilité suivantes*

avec les opérations qui définissent un espace vectoriel.

$$\forall x, y \in E \quad \phi(x+y) = \phi(x) + \phi(y)$$

$$\forall x \in E, \forall \lambda \in I\!R \quad \phi(\lambda x) = \lambda\ \phi(x)$$

A.5.1 Espace image d'une application linéaire

Définition A.5.2 *Soient E et F deux espaces vectoriels. On appelle espace image d'une application linéaire ϕ de E dans F, on note $Im(\phi)$, l'ensemble des éléments de F image par ϕ d'un vecteur de E. Il est facile de montrer que $Im(\phi)$ est un sous-espace de F.*

A.5.2 Noyau d'une application linéaire

Définition A.5.3 *Soient E et F deux espaces vectoriels. On appelle noyau d'une application linéaire ϕ de E dans F, on note $Ker(\phi)$, l'ensemble des éléments de E dont l'image par ϕ est le vecteur nul de F. Il est facile de montrer que $Ker(\phi)$ est un sous-espace de E.*

A.5.3 Rang d'une application linéaire

Définition A.5.4 *Soient E et F deux espaces vectoriels. On appelle rang d'une application linéaire ϕ de E dans F, on note $rg(\phi)$, la dimension de l'espace image $Im(\phi)$.*

Théorème A.5.1 *Soit ϕ une application linéaire de E, espace vectoriel de dimension n, dans un espace vectoriel F. On a alors l'égalité suivante :*

$$dim(Ker(\phi)) + dim(Im(\phi)) = n$$

Conséquence : Soit E un espace de dimension n. Si une application linéaire ϕ de E dans E est injective ($dim(Ker(\phi)) = 0$) elle est aussi surjective ($dim(Im(\phi)) = n$), donc bijective, et inversement, si elle est surjective , elle est injective.

A.6 Matrices

Considérons un espace E de dimension finie n et un espace F de dimension finie p. Soit ϕ une application linéaire de E dans F. Supposons E muni d'une base $\{e_1, e_2, ...e_n\}$ et F d'une base $\{f_1, f_2, ...f_p\}$. Un vecteur x de E sera représenté,

dans la base $\{e_i\}$ par un vecteur X de $\mathbb{R}^n$. De même un vecteur y de F sera représenté, dans la base $\{f_i\}$ par un vecteur Y de $\mathbb{R}^p$. On représente alors, conventionnellement, l'application linéaire ϕ par le tableau rectangulaire à p lignes et n colonnes des composantes, dans la base des $\{f_i\}$, des images par ϕ des vecteurs $\{e_i\}$ de la base de E, rangés colonnes par colonnes.

$$\begin{pmatrix} a_{11}, & a_{12}, & ..., & ... & a_{1n} \\ a_{21}, & a_{22}, & ..., & ... & a_{2n} \\ ..., & ..., & ... & ... & ... \\ ..., & ..., & ... & ... & ... \\ ..., & ..., & ... & ... & ... \\ a_{p1}, & a_{p2}, & ..., & ... & a_{pn} \end{pmatrix}$$

Les composantes y_i du vecteur image dans la base $\{f_i\}$ s'obtiennent par produit scalaire du vecteur ligne i de la matrice avec le vecteur colonne X des composantes x_i dans la base $\{e_i\}$ du vecteur objet x.

$$y_i = \sum_{j=1,n} a_{ij} x_j$$

Ce que l'on représente conventionnellement sous la forme

$$Y = AX$$

Ceci conduit à la définition de l'opération produit matrice vecteur et par extension à celle du produit matriciel représentant la composition d'applications linéaires.

Un cas particulier très important est celui des endomorphismes, applications linéaires d'un espace E dans lui-même, qui sont donc représentées par des matrices carrées. Si elles représentent une application bijective, elles sont inversibles et sont dites régulières. Dans le cas où elles ne sont pas inversibles, on les dit singulières.

A.7 Valeurs et vecteurs propres

Définition A.7.1 *Soit une application linéaire ϕ d'un espace vectoriel E dans lui-même, on appelle valeur propre λ et vecteur propre associé v, le nombre réel ou complexe et tout vecteur de $E \neq 0$ tels que*

$$\phi(v) = \lambda v$$

Appliquée à un vecteur propre, une application linéaire se réduit donc à une homothétie. Si E est de dimension finie N et soit A la matrice qui représente ϕ dans une base de E, on définit de même les valeurs propres et les vecteurs propres de la matrice A, comme les scalaires λ et les vecteurs V non-nuls tels que

$$AV = \lambda V$$

Remarquons que le terme vecteur propre est un peu ambigu. En effet, si V est vecteur propre associé à λ, tout multiple kV de V, l'est aussi. En réalité, à une valeur propre λ est associée une (ou plusieurs) directions propres. Dans le cas de plusieurs directions propres, on parle de sous-espace propre.

Pour qu'il existe une valeur propre λ, et un vecteur propre V associé, il faut (prenons le cas des matrices) que la matrice $A - \lambda I$ soit singulière. En effet, il faut qu'il existe $V \neq 0$ tel que

$$AV = \lambda V \Rightarrow (A - \lambda I)V = 0$$

C'est d'ailleurs une technique classique de calcul des valeurs propres à la main que de rechercher les zéros du determinant de la matrice $A - \lambda I$ (ce n'est pas, en général, la bonne approche pour leur calcul numérique, voir chapitre 2).

A.8 Formes linéaires et bilinéaires

A.8.1 Formes linéaires

Définition A.8.1 *Une forme linéaire l sur un espace vectoriel réel E est une application linéaire de E dans $\mathbb{R}$.*

Si E est de dimension n elle sera donc représentée par un vecteur ligne L de n composantes. Chacune des composantes est l' image par la forme linéaire l d'un vecteur de base de E. L'image d'un vecteur x quelconque de E s'obtient alors par produit scalaire de L par le vecteur X des composantes de x

$$l(x) = (L, X)$$

Retenons qu'en dimension finie, toute forme linéaire se représente par un produit scalaire.

A.8.2 Formes bilinéaires

Définition A.8.2 *Une forme bilinéaire a sur un espace vectoriel réel E est une application de $E \times E$ dans $\mathbb{R}$, linéaire par rapport à chacun de ses deux arguments.*

Soit a la forme bilinéaire, on a donc :

$$a(\lambda_1 u_1 + \lambda_2 u_2, v) = \lambda_1 a(u_1, v) + \lambda_2 a(u_2, v)$$
$$a(u, \mu_1 v_1 + \mu_2 v_2) = \mu_1 a(u, v_1) + \mu_2 a(u, v_2)$$

Représentation matricielle

Toute forme bilinéaire sur un espace E de dimension finie n se représente, dans une base $\{e_i\}$ par une matrice carrée d'ordre n. Les coefficients A_{ij} de la matrice A représentant l'application a sont donnés par

$$A_{ij} = a(e_i, e_j)$$

On a

$$a(u, v) = (AU, V) = (U, A^T V)$$

si $(.,.)$ représente le produit scalaire usuel de $\mathbb{R}^n$ et A^T est la matrice transposée de A définie par

$$A^T_{ij} = A_{ji} \quad \forall i, j = 1...N$$

A.8.3 Formes bilinéaires symétriques définies positives

Définition A.8.3 *Une forme bilinéaire a sur un espace vectoriel réel E est symétrique si :*

$$\forall u, v \in E \quad a(u, v) = a(v, u)$$

Définition A.8.4 *Une forme bilinéaire a sur un espace vectoriel réel E est définie positive si :*

$$\forall u \in E \quad a(u, u) \geq 0$$

et

$$a(u, u) = 0 \Longleftrightarrow u = 0$$

Les formes bilinéaires symétriques sont représentées par des matrices symétriques $A_{ij} = A_{ji}$. Les formes bilinéaires symétriques définies positives sont représentées par des matrices symétriques définies positives, qui vérifient donc :

$$(AU, U) \geq 0 \quad \forall U \in \mathbb{R}^N \quad \text{et} \quad (AU, U) = 0 \Rightarrow U = 0$$

Théorème A.8.1 (de Schur) *Les matrices symétriques réelles ont des valeurs propres réelles, sont diagonalisables et admettent une base de vecteurs propres orthonormés. Les matrices symétriques définies positives ont des valeurs propres strictement positives et donc sont inversibles.*

A.9 Équivalence entre résolution d'un système linéaire et minimisation quadratique

Théorème A.9.1 *Si A est une matrice symétrique définie positive, il y a équivalence entre les trois problèmes suivants :*

$$(1) \quad \begin{cases} \text{Trouver } X \in I\!R^N \quad \text{tel que} \\ \\ A\,X = B \end{cases}$$

$$(2) \quad \begin{cases} \text{Trouver } X \in I\!R^N \quad \text{tel que} \\ \\ (A\,X, Y) = (B, Y) \quad \forall\, Y \in I\!R^N \end{cases}$$

$$(3) \quad \begin{cases} \text{Trouver } X \in I\!R^N \quad \text{tel que} \\ \\ J(X) = \dfrac{1}{2}(A\,X, X) - (B, X) \quad \text{soit minimal} \end{cases}$$

Démonstration
$1 \Longrightarrow 2$ est évident.
$2 \Longrightarrow 1$ en prenant pour Y les vecteurs de base e_i de $I\!R^N$.
$2 \Longrightarrow 3$: On calcule $J(X + \lambda Y)$ pour tout λ réel et tout $Y \in I\!R^N$, on obtient .

$$J(X + \lambda Y) = J(X) + \lambda[(AX, Y) - (B, Y)] + \frac{\lambda^2}{2}(AY, Y)$$

en utilisant la symétrie de la matrice A.

On en déduit, si $(AX, Y) - (B, Y) = 0$ que $J(X + \lambda Y) = J(X) + \frac{\lambda^2}{2}(AY, Y)$
d'où en utilisant le fait que A est définie positive :

$$J(X + \lambda Y) > J(X)$$

si λ et Y sont non nuls. Donc on a montré que si X vérifie (2), X minimise J.
Inversement $3 \Longrightarrow 2$, car si X minimise J, on a

$$\lambda[(AX, Y) - (B, Y)] + \frac{\lambda^2}{2}(AY, Y) \geq 0 \quad \forall \lambda, \forall y$$

Le trinôme en λ ci-dessus doit être toujours positif. Ceci entraîne que son discriminant soit toujours négatif ou nul. Or ce discriminant est

$$\Delta = [(AX, Y) - (B, Y)]^2$$

Ceci implique (2). On a donc démontré les équivalences $1 \Longleftrightarrow 2$ et $2 \Longleftrightarrow 3$ et donc l'équivalence des 3 problèmes.

A.10 Application aux moindres carrés

Considérons le problème général d'un système linéaire sur-déterminé, c'est à dire dans lequel il y a plus d'équations que d'inconnues. C'est en particulier le cas dans le calcul de la droite des moindres carrés ou plus généralement de polynômes d'approximation au sens des moindres carrés. On ne peut pas obtenir exactement l'égalité

$$AX = B$$

car A est une matrice rectangulaire de N lignes et m colonnes avec $N >> m$. On essaie alors de minimiser l'écart entre les vecteurs AX et B de $\mathbb{R}^N$ en minimisant la norme euclidienne de leur différence, ou ce qui revient au même le carré de cette norme.

$$Minimiser\, J(X) = \|AX - B\|^2 = (AX - B, AX - B)$$

On utilise les propriétés classiques du produit scalaire $(AU, V) = (U, A^T V)$ pour obtenir :

$$J(X) = (A^T AX, X) - 2(A^T B, X) + (B, B)$$

la matrice $A^T A$ est une matrice carrée $m \times m$ symétrique définie positive. Le théorème (1,3) nous donne l'équivalence de ce problème de moindres carrés avec la résolution du système linéaire

$$A^T AX = A^T B$$

On retrouve ainsi le système carré de m équations à m inconnues, dit "système des équations normales".

Annexe B

Rappels d'analyse fonctionnelle

Nous nous limitons pour la suite au cas d'espaces vectoriels sur le corps des réels.

B.1 Produit scalaire

Définition B.1.1 (Produit scalaire) *On appelle produit scalaire dans un espace vectoriel réel E, une forme bilinéaire symétrique définie positive sur E. On note le produit scalaire de 2 éléments x, y de E sous la forme (x, y).*

On a donc $\forall\, x, y, z \in E$ et $\forall\, a, b \in I\!R$, les propriétés suivantes :

$$(x, y) \in I\!R$$

$$(x, ay + bz) = a(x, y) + b(x, z) \quad \text{et} \quad (ax + by, z) = a(x, z) + b(y, z)$$

$$(x, y) = (y, x)$$

$$(x, x) \geq 0 \quad \text{et} \quad (x, x) = 0 \Longleftrightarrow x = 0$$

B.1.1 Norme déduite du produit scalaire

Nous allons montrer que l'application de E dans $I\!R$ qui à $x \in E$ associe $\| x \| = (x, x)^{\frac{1}{2}}$ est une norme sur E.

Les 2 premiers axiomes de définition d'une norme :

$$x = 0 \quad \Longleftrightarrow \quad \|x\| = 0$$

$$\|\lambda x\| = |\lambda| \, \|x\|$$

se déduisent immédiatement de la définition du produit scalaire.

La démonstration de l'inégalité triangulaire

$$\|x+y\| \leq \|x\| \, + \, \|y\|$$

nécessite l'inégalité fondamentale suivante

B.1.2 Inégalité de Schwarz

Théorème B.1.1 *$\forall \; x, y \in E$, on a l'inégalité :*

$$|(x,y)| \leq \|x\| \, \|y\|$$

Démonstration.

Pour $x, y \in E$ et $\lambda \in I\!R$ on a :

$$(x+\lambda y, x+\lambda y) = \|x\|^2 \, + \, 2\,\lambda\,(x,y) \, + \, \lambda^2 \, \|y\|^2 \geq 0$$

Le trinôme en λ doit donc être toujours ≥ 0. Ceci entraîne que son discriminant soit ≤ 0 d'où

$$(x,y)^2 \leq \|x\|^2 \, \|y\|^2$$

et le résultat en prenant la racine carrée des 2 membres.

On en déduit simplement l'inégalité triangulaire

$$\|x \, + \, y\| \leq \, \|x\| \, + \, \|y\|$$

B.2 Espace de Hilbert

Définition B.2.1 *Un espace de Hilbert est un espace vectoriel muni d'un produit scalaire (x,y) et qui est complet lorsqu'il est normé par la norme associée à ce produit scalaire. En dimension finie, un espace vectoriel muni d'un produit scalaire est un espace euclidien*

B.2.1 Exemples d'espaces de Hilbert

a) Les espaces $I\!R^n$ sont pour tout n des espaces euclidiens pour le produit scalaire euclidien classique

b) L'espace l^2 des suites x_n de carré sommable, c'est à dire telles que la série $\sum_{n \in N} x_n^2$ converge, muni du produit scalaire

$$(x,y) = \sum_{n \in N} x_n y_n$$

est un espace de Hilbert

c) L'espace $L^2(I)$ des fonctions de carré sommable sur I, où I est un intervalle ouvert $]a, b[$ de $\mathbb{R}$, c'est à dire telles que l'intégrale

$$\int_a^b (f(x))^2 \, dx$$

existe, muni du produit scalaire

$$(f, g) = \int_a^b f(x)g(x) \, dx$$

est un espace de Hilbert.

d) L'espace $L^2(\Omega)$ des fonctions de carré sommable sur Ω, où Ω est un domaine ouvert de $\mathbb{R}^2$ ou $\mathbb{R}^3$, c'est à dire telles que l'intégrale

$$\int_\Omega (f(x))^2 \, dx$$

existe, muni du produit scalaire

$$(f, g) = \int_\Omega f(x)g(x) \, dx$$

est un espace de Hilbert.

B.2.2 Orthogonalité

Deux vecteurs x et y d'un espace de Hilbert H sont orthogonaux si leur produit scalaire $(x, y) = 0$.

Soit E un sous-espace de H, l'ensemble des éléments de H orthogonaux à E est un sous-espace de H appelé l'orthogonal de E et noté $E^\perp$.
En effet $E^\perp$ est évidemment un sous espace vectoriel de H que l'on peut munir du produit scalaire de H. De plus, la continuité de l'application $x \longrightarrow (x, y)$ entraîne que $E^\perp$ est fermé dans H, donc complet.

B.2.3 Représentation des applications linéaires continues

L'application $u \longrightarrow (u, v)$ pour tout u et v de H est une application linéaire continue de norme $\|v\|$. Réciproquement, on admettra le théorème fondamental suivant :

Théorème B.2.1 (de Riesz) *Si l est une forme linéaire continue sur H, il existe un élément L unique de H tel que*

$$l(u) = (L, u) \quad \forall \, u \in H$$

B.3 Projection

B.3.1 Projection sur un convexe fermé

Définition B.3.1 (convexe) *Un sous-ensemble K d'un espace H est dit convexe si $\forall x, y \in K$ et $\forall t \in [0,1]$, on a $tx + (1-t)y \in K$.*

On admettra le théorème suivant :

Théorème B.3.1 *Soit K un sous-ensemble convexe fermé non vide d'un Hilbert H, pour tout $u \in H$ il existe un élément unique $\bar{u} \in K$ tel que*

$$\|u - \bar{u}\| = \min_{v \in K} \|u - v\|$$

On note ce projeté $\bar{u}$ de u dans K : $\Pi_K u$.

Propriété caractéristique

Le projeté $\Pi_K u$ de u sur K est caractérisé par la propriété :

$$(u - \Pi_K u, w - \Pi_K u) \leq 0 \quad \forall w \in K$$

Démonstration

Par définition $(u - \Pi_K u, u - \Pi_K u) \leq (u - v, u - v) \quad \forall v \in K$. Soit w quelconque dans K on considère la combinaison convexe $tw + (1-t)\Pi_K u$, avec $0 \leq t \leq 1$ qui appartient également à K. Donc, par définition du projeté :

$$(u - \Pi_K u, u - \Pi_K u) \leq (u - tw - (1-t)\Pi_K u, u - tw - (1-t)\Pi_K u) \quad \forall w \in K$$

On développe et on obtient

$$2\,(u - \Pi_K u, w - \Pi_K u) \leq t\,(\Pi_K u - w, \Pi_K u - w)$$

et le résultat en faisant tendre t vers zéro dans l'inégalité.

Autres propriétés

1) Π_K est idempotente, i.e. $\Pi_K^2 = \Pi_K$
2) Π_K est monotone, i.e. $(\Pi_K u - \Pi_K v, u - v) \geq 0 \quad \forall u, v \in H$.
3) Π_K est faiblement contractante, i.e. $\|\Pi_K u - \Pi_K v\| \leq \|u - v\| \quad \forall u, v \in H$.

Démonstration

1) $\Pi_K u \in K$, donc évidemment $\Pi_K(\Pi_K u) = \Pi_K u$
2 et 3) On utilise la propriété caractéristique :

$$(u - \Pi_K u, w - \Pi_K u) \leq 0 \quad \forall w \in K$$

$$(v - \Pi_K v, w - \Pi_K v) \leq 0 \quad \forall w \in K$$

On choisit $w = \Pi_K v$ dans la première inégalité et $w = \Pi_K u$ dans la seconde, et on obtient par addition

$$\|\Pi_K u - \Pi_K v\|^2 \leq (u - v, \Pi_K u - \Pi_K v)$$

ceci entraîne 2) et on obtient 3) par Schwarz.

$$(u - v, \Pi_K u - \Pi_K v) \leq \|u - v\| \, \|\Pi_K u - \Pi_K v\|$$

B.3.2 Projection sur un sous-espace vectoriel fermé

Un sous-espace est un sous ensemble convexe, donc les résultats précédents s'appliquent. On a de plus le théorème suivant :

Théorème B.3.2 *Soit F un sous-espace fermé non vide d'un espace de Hilbert H, on note Π_F la projection sur F. On a les résultats suivants :*

1) $\forall u \in H, \quad \Pi_F u$ est caractérisé par

$$(u - \Pi_F u, v) = 0 \quad \forall v \in F$$

2) L'application projection Π_F est linéaire continue de norme 1, son noyau est l'orthogonal de F dans H noté $F^\perp$
3) Il existe un couple unique d'applications linéaires Π_F et $\Pi_{F^\perp}$ qui appliquent respectivement H dans F et H dans $F^\perp$ telles que :

$$u = \Pi_F u + \Pi_{F^\perp} u$$

Démonstration

1) Conséquence directe de la propriété caractéristique de la projection sur un convexe en prenant successivement $w = \Pi_K u + v$ et $w = \Pi_K u - v$
2) La linéarité se déduit simplement de la propriété caractéristique précédente. La continuité de l'inégalité 3) et la norme 1 du choix d'un $u \in F$ dans cette inégalité. Enfin

$$(u, v) = (\Pi_F u, v) \quad \forall v \in F$$

entraîne $(\Pi_F u, v) = 0 \quad \forall v \in F$ si $u \in F^\perp$, d'où le résultat que le noyau de Π_F est $F^\perp$
3) L'existence de Π_F est acquise, on montre simplement par

$$(u - \Pi_F u, v) = 0 \quad \forall v \in F$$

que $(u - \Pi_F u) \in F^\perp$ et le résultat car $u - (u - \Pi_F u) = \Pi_F u$ est orthogonal à $F^\perp$.

B.4 Bases hilbertiennes

Définition B.4.1 *Un espace de Hilbert est dit séparable s'il contient une suite d'éléments (un sous ensemble dénombrable) dense.*

Définition B.4.2 *Une suite d'éléments $\{e_i\}$ d'un Hilbert H est une base hilbertienne de H si*

1) elle est orthonormée :

$$(e_i, e_j) = 0 \quad \forall\, i \neq j$$
$$\|e_i\| = 1 \quad \forall\, i$$

2) elle est totale, c'est à dire que l'ensemble des combinaisons linéaires des e_i est dense dans H.

B.4.1 Coefficients de Fourier

Soit $\{e_i\}$ une base hilbertienne de H, pour tout $x \in H$, les nombres (x, e_i) sont les coordonnées ou coefficients de Fourier de x dans la base $\{e_i\}$. On a le théorème :

Théorème B.4.1 *Soient $\{e_1, e_2, ...e_n\}$, n vecteurs quelconques de la base hilbertienne et soit H_n le sous-espace de dimension finie n qu'ils engendrent dans H.*

Pour tout x appartenant à H, notons $x_i = (x, e_i) \quad \forall i$, les coefficients de Fourier de x. Le vecteur

$$\sum_{i=1,n} x_i e_i$$

est la projection $\Pi_{H_n} x$ de x dans H_n. C'est donc la meilleure approximation de x dans H_n au sens de la norme de H.

Démonstration

En effet $\forall x \in H$, et $\forall e_j$ vecteur de base de H_n on a

$$(x - \sum_{i=1,n} x_i e_i, e_j) = 0$$

par construction. Donc le vecteur

$$x - \sum_{i=1,n} x_i e_i$$

est orthogonal à tout vecteur de base de H_n, donc à tout vecteur de H_n. Ceci montre bien que

$$\sum_{i=1,n} x_i e_i$$

est la projection $\Pi_{H_n} x$ de x dans H_n

On en déduit en faisant tendre n vers l'infini :

1) que la série : $\sum_i x_i e_i$ converge et que sa somme est x,

2) que la série : $\sum_i x_i^2$ converge et que sa somme est $\|x\|^2$ (C'est l'égalité de **Bessel - Parseval**).

Le théorème précédent est fondamental. Il signifie que tout espace de Hilbert séparable, admettant une base hilbertienne, sera donc isomorphe et isométrique à l'espace l^2 des suites de carrés sommables. Ceci est la généralisation naturelle en dimension infinie de l'isomorphisme isométrique naturel entre un espace euclidien de dimension finie N et l'espace $\mathbb{R}^N$.

D'autre part, le théorème précédent donne une construction pratique de la meilleure approximation dans un sous espace de dimension finie par l'utilisation d'une base orthonormée. Ceci a de nombreuses et importantes applications : approximations polynomiales au sens des moindres carrés, polynômes orthogonaux, développements de Fourier etc.

B.4.2 Quelques exemples de bases hilbertiennes

1) Soit l'espace $L^2(-\pi, +\pi)$ des fonctions de carré sommable sur $(-\pi, +\pi)$, la suite des fonctions

$$\{\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\, cos(x), \frac{1}{\sqrt{\pi}}\, sin(x), ..., \frac{1}{\sqrt{\pi}}\, cos(nx), \frac{1}{\sqrt{\pi}}\, sin(nx), ...\}$$

est une base hilbertienne de l'espace des fonctions périodiques de période $2\,\pi$ de carré sommable sur $(-\pi, +\pi)$

2) Dans l'espace $L^2(-1, 1)$ des fonctions de carré sommable sur $(-1, 1)$, la suite des fonctions monômes

$$\{1, x, x^2, ..., x^n, ...\}$$

base de l'espace des polynômes qui est dense dans $L^2(-1, 1)$, n'est pas orthonormée. Pour obtenir à partir de cette base canonique une base hilbertienne, il suffit de l'orthonormer en utilisant le procédé de **Gram-Schmidt**.

B.4.3 Procédé de Gram-Schmidt

A partir d'une base non orthonormée $\{e_1, e_2, ..e_n, ..\}$, on peut construire une base orthonormée par l'algorithme suivant.

$$e_1^* = \frac{e_1}{\|e_1\|}$$

$$\tilde{e}_2 = e_2 - (e_2, e_1^*)e_1^* \quad \text{et} \quad e_2^* = \frac{\tilde{e}_2}{\|\tilde{e}_2\|}$$

$$\tilde{e}_n = e_n - \sum_{i=1}^{n-1}(e_n, e_i^*)e_i^* \quad \text{et} \quad e_n^* = \frac{\tilde{e}_n}{\|\tilde{e}_n\|}$$

Remarque B.4.1 *L'orthonormalisation pratique utilise la technique de Householder plus stable que l'algorithme de Gram-Schmidt (voir Lascaux-Théodor par exemple).*

Application.

Orthonormons la base canonique $\{1, x, x^2, ..\}$ au sens du produit scalaire de $L^2[-1, 1]$. On obtient alors la base des polynômes orthogonaux de Legendre.

B.5 Exemples d'espaces fonctionnels en dimension un

B.5.1 Espaces de fonctions continues

$\underline{C[a, b]}$

$C[a, b]$ est l'espace des fonctions continues sur l'intervalle fermé borné $[a, b]$ de $\mathbb{R}$, muni de la norme ∞

$$\|f\|_{0,\infty} = \sup_{x\in[a,b]} |f(x)|$$

$\underline{C^k[a, b]}$

$C^k[a, b]$ est l'espace des fonctions continues et dont les dérivées jusqu'à l'ordre k inclus sont continues sur l'intervalle fermé borné $[a, b]$ de $\mathbb{R}$, muni de la norme

$$\|f\|_{k,\infty} = \sum_{0\leq j\leq k} \|f^{(j)}\|_{0,\infty}$$

où $f^{(j)}$ représente la dérivée d'ordre j de f.

On notera $C^\infty[a,b]$ l'espace des fonctions continues et à dérivées continues quel que soit l'ordre de dérivation.

B.5.2 Espaces de fonctions de carré sommable

$\underline{L^2[a,b]}$

$L^2[a,b]$ est l'espace des fonctions de carré sommable sur $]a,b[$ muni de la norme

$$\|f\|_{0,2} = \big(\int_a^b f(x)^2\,dx\big)^{\frac{1}{2}}$$

C'est un espace de Hilbert pour le produit scalaire

$$(f,g) = \int_a^b f(x)\,g(x)\,dx$$

$\underline{H^k[a,b]}$

$H^k[a,b]$ est l'espace des fonctions de carré sommable et dont les dérivées, au sens des distributions, jusqu'à l'ordre k inclus sont de carré sommable sur $]a,b[$ muni de la norme

$$\|f\|_{k,2} = \big(\sum_{0\le j\le k} \|f^{(j)}\|_{0,2}^2\big)^{\frac{1}{2}}$$

où $f^{(j)}$ représente la dérivée d'ordre j de f. C'est un espace de Hilbert pour le produit scalaire

$$(f,g) = \sum_{0\le j\le k} \int_a^b f^{(j)}(x)\,g^{(j)}(x)\,dx$$

B.5.3 Propriétés d'inclusion

Pour tout $k \ge 1$ on a les inclusions évidentes :

$$C^\infty[a,b] \subset C^{k+1}[a,b] \subset C^k[a,b] \subset C[a,b]$$

et

$$H^{k+1}[a,b] \subset H^k[a,b] \subset L^2[a,b]$$

On a de plus les inclusions suivantes en dimension un :

$$H^1[a,b] \subset C[a,b] \subset L^2[a,b]$$

et en général

$$H^{k+1}[a,b] \subset C^k[a,b]$$

B.5.4 L'espace $H_0^1[a,b]$

L'inclusion $H^1[a,b] \subset C[a,b]$ permet de définir l'espace $H_0^1[a,b]$ des fonctions de $H^1[a,b]$ nulles aux points a et b. C'est un sous-espace fermé de l'espace de Hilbert $H^1[a,b]$ pour le même produit scalaire et la même norme. Il aura une grande importance pour la formulation de problèmes aux limites dans lesquels la solution est fixée aux bornes de l'intervalle. On a alors besoin de l'inégalité fondamentale suivante :

B.5.5 Inégalité de Poincaré

Théorème B.5.1 *Soit $[a,b]$ un intervalle fermé borné de $\mathbb{R}$, on a pour tout $v \in H_0^1[a,b]$ l'inégalité*

$$\|v\|_{0,2} \leq \frac{1}{\sqrt{2}}\,(b-a)\,\|v'\|_{0,2}$$

D'où l'on déduit :

$$\|v\|_{1,2} \leq (1 + \frac{(b-a)^2)}{2})^{\frac{1}{2}}\,\|v'\|_{0,2}$$

et donc que $|v|_{1,2} = \|v'\|_{0,2}$ définit une norme sur $H_0^1[a,b]$ équivalente à la norme H^1.

Démonstration

pour tout $x \in [a,b]$, on a :

$$v(x) = \int_a^x v'(t)\,dt$$

donc

$$|v(x)| \leq \sqrt{x-a}\,\|v'\|_{0,2}$$

On en déduit

$$\int_a^b v(x)^2\,dx \leq (\int_a^b (x-a)\,dx)\,\|v'\|_{0,2}^2 \leq \frac{(b-a)^2}{2}\,\|v'\|_{0,2}^2$$

et le résultat. L'équivalence des normes se déduit de la définition de $\|v\|_{1,2}$:

$$\|v\|_{1,2} = \left(\|v\|_{0,2}^2 + \|v'\|_{0,2}^2\right)^{\frac{1}{2}}$$

B.5.6 Quelques résultats de densité

L'espace des fonctions polynômes est dense dans l'espace $C[a,b]$ (Théorème de Weierstrass).

L'espace des fonctions polynômes est dense dans l'espace $L^2[a,b]$. On peut alors construire des bases hilbertiennes de polynômes orthogonaux engendrant un sous-espace dense dans $L^2[a,b]$

L'espace des fonctions continues, polynomiales par morceaux, est dense dans $H^1[a,b]$. Cet espace sera à la base de la méthode des éléments finis en dimension un.

B.6 Exemples d'espaces fonctionnels en dimension deux et trois

B.6.1 Rappels

Opérateurs différentiels en dimension trois

Soit u une fonction de 3 variables (x,y,z) à valeurs réelles définie sur un domaine Ω de $\mathbb{R}^3$. On appelle gradient de u et on note $\mathbf{grad}(u)$ ou ∇u, le vecteur :

$$\mathbf{grad}(u) = \nabla u = \begin{pmatrix} \dfrac{\partial u}{\partial x} \\ \dfrac{\partial u}{\partial y} \\ \dfrac{\partial u}{\partial z} \end{pmatrix}$$

Soit $\mathbf{V}$ une fonction vectorielle de 3 variables (x,y,z) définie sur un domaine Ω de $\mathbb{R}^3$ et à valeurs V_1, V_2, V_3 dans $\mathbb{R}$. On appelle divergence du vecteur $\mathbf{V}$ et on note $div(\mathbf{V})$ ou $\nabla.\mathbf{V}$, le scalaire :

$$div(\mathbf{V}) = \nabla.\mathbf{V} = \frac{\partial V_1}{\partial x} + \frac{\partial V_2}{\partial y} + \frac{\partial V_3}{\partial z}$$

Soit u une fonction de 3 variables (x,y,z) à valeurs réelles définie sur un domaine Ω de $\mathbb{R}^3$. On appelle laplacien de u et on note Δu ou $\nabla^2 u$, le scalaire :

$$\Delta u = div(\mathbf{grad}(u)) = \nabla.\nabla(u) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}$$

Soit $\mathbf{V}$ une fonction vectorielle de 3 variables (x,y,z) définie sur un domaine Ω de $\mathbb{R}^3$ et à valeurs V_1, V_2, V_3, on appelle rotationnel de V le vecteur

$$\mathbf{rot}\mathbf{V} = \nabla \wedge \mathbf{V} = \begin{bmatrix} \frac{\partial V_3}{\partial y} - \frac{\partial V_2}{\partial z} \\ \frac{\partial V_1}{\partial z} - \frac{\partial V_3}{\partial x} \\ \frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} \end{bmatrix}$$

Quelques formules d'analyse vectorielle

On a :

$$div(\mathbf{rot}\mathbf{V}) = 0$$

Un champ de rotationnel est à divergence nulle. En particulier pour un fluide, un champ de rotationnel représente la vitesse d'un fluide incompressible.

$$\mathbf{rot}(\mathbf{grad}f) = 0$$

Un champ de gradient (qui dérive d'un potentiel) est irrotationnel.

$$\mathbf{grad}(f\,g) = f\ \mathbf{grad}(g) + g\ \mathbf{grad}(f)$$
$$div(a\mathbf{V}) = a\,div(\mathbf{V}) + \mathbf{V}\ \mathbf{grad}(f)$$
$$\mathbf{rot}(a\mathbf{V}) = a\,\mathbf{rot}(\mathbf{V}) + \mathbf{grad}(a) \wedge \mathbf{V}$$

Opérateurs en coordonnées cylindriques et sphériques

En coordonnées cylindriques l'opérateur Laplacien prend la forme suivante :

$$\Delta U = \frac{\partial^2 U}{\partial r^2} + \frac{1}{r}\frac{\partial U}{\partial r} + \frac{1}{r^2}\frac{\partial^2 U}{\partial \theta^2} + \frac{\partial^2 U}{\partial z^2}$$

avec $U(r, \theta, z) = u(r\cos\theta, r\sin\theta, z)$.

En coordonnées sphériques l'opérateur Laplacien prend la forme suivante :

$$\Delta U = \frac{\partial^2 U}{\partial r^2} + \frac{2}{r}\frac{\partial U}{\partial r} + \frac{1}{r^2}\frac{\partial^2 U}{\partial \theta^2} + \frac{\cos\theta}{r^2\sin\theta}\frac{\partial U}{\partial \theta} + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 U}{\partial \varphi^2}$$

avec $U(r, \varphi, \theta) = u(r\sin\theta\cos\varphi, r\sin\theta\sin\varphi, r\cos\theta)$.

On peut montrer que dans le cas particulier où la fonction U ne dépend que de la distance à l'origine, le Laplacien dans R^n a la forme suivante :

$$\Delta U(r) = \frac{d^2 U}{dr^2} + \frac{(n-1)}{r}\frac{dU}{dr}$$

Définition B.6.1 (Dérivée directionnelle) *Soit V un vecteur unitaire de $\mathbb{R}^3$. On appelle dérivée directionnelle de f au point $M_0 = (x_0, y_0, z_0)$ dans la direction de V le produit scalaire*

$$\nabla f(M_0)\cdot V$$

On utilise souvent la notion de dérivée normale sur le bord d'un domaine Ω de $\mathbb{R}^3$. Notons $\mathbf{n}$ le vecteur normal unitaire orienté vers l'extérieur du domaine Ω en chaque point de son bord $\partial\Omega$. On obtient :

$$\frac{\partial f}{\partial n} = \nabla f\cdot\mathbf{n}$$

Dérivation des fonctions composées

On rappelle les formules de dérivation des fonctions composées à plusieurs variables. Par exemple dans le cas

$$g(s,t) = f(x(s,t), y(s,t), z(s,t))$$

on a

$$\begin{cases} \dfrac{\partial g}{\partial s} = \dfrac{\partial f}{\partial x}\dfrac{\partial x}{\partial s} + \dfrac{\partial f}{\partial y}\dfrac{\partial y}{\partial s} + \dfrac{\partial f}{\partial z}\dfrac{\partial z}{\partial s} \\ \dfrac{\partial g}{\partial t} = \dfrac{\partial f}{\partial x}\dfrac{\partial x}{\partial t} + \dfrac{\partial f}{\partial y}\dfrac{\partial y}{\partial t} + \dfrac{\partial f}{\partial z}\dfrac{\partial z}{\partial t} \end{cases}$$

Ce que l'on peut également noter :

$$\begin{cases} g_s = f_x x_s + f_y y_s + f_z z_s \\ g_t = f_x x_t + f_y y_t + f_z z_t \end{cases}$$

Formule de Taylor

L'étude locale d'une fonction se fait à l'aide de la formule de Taylor, qui s'écrit de la façon suivante pour une fonction de trois variables :

$$f(x+h, y+k, z+l) = f(x,y,z) + [h,k,l] \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial x} \end{bmatrix}$$

$$+\tfrac{1}{2}[h,k,l] \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{bmatrix} \begin{bmatrix} h \\ k \\ l \end{bmatrix} + (h^2+k^2+l^2)\epsilon(h,k,l)$$

La matrice 3×3 des dérivées partielles secondes est symétrique si f est deux fois continuement dérivable (ce que l'on supposera) et s'appelle le **Hessien** de f ou matrice hessienne de f.

Intégrales doubles et changement de variables

On rappelle que pour un changement de variables correspondant à une transformation $(x,y) \in \Omega \longrightarrow (s,t) \in D$ définie par :

$$x = f_1(s,t) \quad y = f_2(s,t)$$

on appelle déterminant jacobien de la transformation le déterminant :

$$J(f_1, f_2) = \begin{vmatrix} \dfrac{\partial f_1}{\partial s} & \dfrac{\partial f_1}{\partial t} \\ \dfrac{\partial f_2}{\partial s} & \dfrac{\partial f_2}{\partial t} \end{vmatrix}$$

que l'on note également : $\dfrac{D(x,y)}{D(s,t)}$

On a alors l'égalité :

$$\iint_\Omega f(x,y)\,dxdy \;=\; \iint_D f(x(s,t),y(s,t))\left|\frac{D(x,y)}{D(s,t)}\right| dsdt$$

Intégrales curvilignes

Soit une courbe C d'extrémités A et B et de longueur L, orientée de A vers B. On note $s(M)$ l'abscisse curviligne d'un point M de C. Soit f une fonction qui à tout point M de C associe le réel $f(M) = f(s)$. On définit l'intégrale curviligne de f sur C par l'intégrale simple suivante :

$$\int_C f(M)ds = \int_0^L f(s)ds$$

Si C admet une représentation paramétrique de paramètre t, nous aurons, en posant $f(s(t)) = g(t)$:

$$\int_C f(M)ds = \int_{t_A}^{t_B} g(t)\,\frac{ds}{dt}\,dt$$

où t_A et t_B sont respectivement les valeurs du paramètre t associées aux points A et B de la courbe C.

B.6.2 Formules de Green

On se place, pour fixer les idées, dans le cas de la dimension deux, mais tout ce qui suit passe sans difficulté en dimension trois.

Formule de base

Soit $\Omega \subset I\!R^2$ de frontière Γ, C^1 par morçeaux, et u et v fonctions de $H^1(\Omega)$, on a l'égalité suivante pour chaque composante i = 1, 2 :

$$\iint_\Omega \frac{\partial u}{\partial x_i}\,v\,dx_1dx_2 = -\iint_\Omega \frac{\partial v}{\partial x_i}\,u\,dx_1dx_2 + \int_\Gamma u\,n_i\,v\,d\gamma$$

avec n_i, i^{ieme} composante du vecteur normal unitaire à Γ orienté vers l'extérieur de Ω et noté $\mathbf{n}$ et γ abscisse curviligne sur Γ orientée dans le sens direct.

Formules déduites

En notant en caractère gras les vecteurs (en particulier **u** désigne ici un vecteur), on déduit aisément les formules suivantes :

— la formule de la divergence

$$\iint_\Omega div\mathbf{V}\, d\Omega = \int_\Gamma \mathbf{V}.\mathbf{n}\, d\gamma$$

Cette formule est aussi appelée **formule d'Ostrogradski-Gauss** ou aussi **formule du flux**

— on en déduit la formule d'intégration "par parties" suivantes :

$$\iint_\Omega div\mathbf{u}\, v\, dxdy = -\iint_\Omega \mathbf{u}\,\mathbf{grad}v\, dxdy + \int_\Gamma \mathbf{u}\,\mathbf{n}\, v\, d\gamma$$

— en posant $\mathbf{u} = a\,\mathbf{grad}(u)$ où a est une fonction $C^1(\bar{\Omega})$ et $\dfrac{\partial u}{\partial n} = \mathbf{grad}(u).\mathbf{n}$:

$$\iint_\Omega div(a\,\mathbf{grad}u)\, v\, dxdy = -\iint_\Omega a\,\mathbf{grad}u\,\mathbf{gradv}\, dxdy + \int_\Gamma a\,\frac{\partial u}{\partial n}\, v\, d\gamma$$

— enfin dans le cas $a = 1$ on obtient la formule classique :

$$-\iint_\Omega \Delta u\ v\, dxdy = \iint_\Omega \mathbf{grad}u\,\mathbf{grad}v\, dxdy - \int_\Gamma \frac{\partial u}{\partial n}\, v\, d\gamma$$

— et la troisième formule de Green

$$\iint_\Omega \Delta u\ v\, dxdy - \iint_\Omega \Delta v\ u\, dxdy = \int_\Gamma \frac{\partial u}{\partial n}\, v\, d\gamma - \int_\Gamma \frac{\partial v}{\partial n}\, u\, d\gamma$$

— formule du rotationnel ou de la circulation

$$\iint_\Omega \mathbf{rot}\mathbf{V}\,.\,\mathbf{n}d\Omega = \int_\Gamma \mathbf{V}\,.\,\tau d\gamma$$

avec τ, vecteur tangent unitaire à Γ orienté dans le sens direct.

B.6.3 Espaces de fonctions continues en dimensions supérieures à un

$\underline{C[\bar{\Omega}]}$

$C[\bar{\Omega}]$ est l'espace des fonctions continues sur le fermé borné $\bar{\Omega}$ de $I\!R^2$ ou $I\!R^3$, muni de la norme ∞

$$\|f\|_{0,\infty} = \sup_{x\in\bar{\Omega}} |f(x)|$$

ici x désigne selon le cas le point (x_1, x_2) de $I\!R^2$ ou le point (x_1, x_2, x_3) de $I\!R^3$

$\underline{C^k[\bar{\Omega}]}$

$C^k[\bar{\Omega}]$ est l'espace des fonctions continues et dont les dérivées partielles jusqu'à l'ordre k inclus sont continues sur le fermé borné $\bar{\Omega}$ de $I\!R^2$ ou $I\!R^3$, muni de la norme

$$\|f\|_{k,\infty} = \sum_{\substack{0\le j\le k \\ j_1+j_2=j}} \Big\|\frac{\partial^j f}{\partial x_1^{j_1} x_2^{j_2}}\Big\|_{0,\infty}$$

pour le cas de $I\!R^2$ et de la norme

$$\|f\|_{k,\infty} = \sum_{\substack{0\le j\le k \\ j_1+j_2+j_3=j}} \Big\|\frac{\partial^j f}{\partial x_1^{j_1} x_2^{j_2} x_3^{j_3}}\Big\|_{0,\infty}$$

dans le cas de $I\!R^3$.

On notera $C^\infty[\bar{\Omega}]$ l'espace des fonctions continues et à dérivées partielles continues quel que soit l'ordre de dérivation.

B.6.4 Espaces de fonctions de carré sommable

$\underline{L^2[\Omega]}$

$L^2[\Omega]$ est l'espace des fonctions de carré sommable sur l'ouvert Ω muni de la norme

$$\|f\|_{0,2} = \big(\int_\Omega f(x)^2\, dx\big)^{\frac{1}{2}}$$

où dx représente respectivement l'élément d'aire dx_1dx_2 ou l'élément de volume $dx_1dx_2dx_3$. C'est un espace de Hilbert pour le produit scalaire

$$(f, g) = \int_\Omega f(x)\, g(x)\, dx$$

$\underline{H^1[\Omega]}$

$H^1[\Omega]$ est l'espace des fonctions de carré sommable et dont les dérivées partielles (au sens des distributions) jusqu'à l'ordre 1 inclus sont de carré sommable sur Ω muni de la norme

$$\|f\|_{1,2} = \big(\|f\|_{0,2}^2 + \|\frac{\partial f}{\partial x_1}\|_{0,2}^2 + \|\frac{\partial f}{\partial x_2}\|_{0,2}^2 + \|\frac{\partial f}{\partial x_3}\|_{0,2}^2\big)^{\frac{1}{2}}$$

C'est un espace de Hilbert pour le produit scalaire

$$(f,g) = \int_\Omega f(x)\, g(x)\, dx + \int_\Omega \mathbf{grad} f(x)\, \mathbf{grad} g(x)\, dx$$

$\underline{H^k[\Omega]}$

$H^k[\Omega]$ est l'espace des fonctions de carré sommable et dont les dérivées partielles (au sens des distributions) jusqu'à l'ordre k inclus sont de carré sommable sur Ω muni de la norme

$$\|f\|_{k,2} = \big(\sum_{0\le j\le k} \|D^{(j)} f\|_{0,2}^2\big)^{\frac{1}{2}}$$

où $D^{(j)} f$ représente le vecteur des dérivées partielles d'ordre j de f :

$$D^{(j)} f = \frac{\partial^j f}{\partial x_1^{j_1} x_2^{j_2} x_3^{j_3}}$$

avec $j_1 + j_2 + j_3 = j$.

C'est un espace de Hilbert pour le produit scalaire

$$(f,g) = \sum_{0\le j\le k} \int_\Omega D^{(j)} f(x)\, D^{(j)} g(x)\, dx$$

B.6.5 Propriétés d'inclusion

Pour tout $k \ge 1$ on a les inclusions évidentes :

$$C^\infty[\bar\Omega] \subset C^{k+1}[\bar\Omega] \subset C^k[\bar\Omega] \subset C[\bar\Omega]$$

et

$$H^{k+1}[\Omega] \subset H^k[\Omega] \subset L^2[\Omega]$$

ATTENTION : On n' a plus par contre l' inclusion

$$H^1[\Omega] \subset C[\bar{\Omega}]$$

mais seulement, pour des domaines Ω de dimension 2 ou 3,

$$H^2[\Omega] \subset C[\bar{\Omega}]$$

B.6.6 L'espace $H_0^1[\Omega]$

Le fait que l'on n'ait plus l'inclusion

$$H^1[\Omega] \subset C[\bar{\Omega}]$$

en dimension 2 ou 3 pose le problème de la définition de conditions aux limites fixées et en particulier de l'espace $H_0^1[\Omega]$ des fonctions de $H^1[\Omega]$ "nulles" sur la frontière Γ de Ω. On admettra le résultat suivant :

Théorème B.6.1 (théorème de trace) *Si le domaine Ω a une frontière Γ assez régulière, les fonctions de $H^1[\Omega]$ sont de carré sommable sur la frontière Γ du domaine Ω de $\mathbb{R}^2$ ou $\mathbb{R}^3$, donc appartiennent à l'espace $L^2[\Gamma]$. Et l'on a la majoration :*

$$\|f\|_{L^2(\Gamma)} \leq C \, \|f\|_{H^1(\Omega)} \quad \forall \, f \in H^1[\Omega]$$

On définira alors $H_0^1[\Omega]$ comme l'espace des fonctions de $H^1[\Omega]$ nulles (au sens des traces) sur la frontière Γ de Ω.

C'est un sous-espace fermé de l'espace de Hilbert $H^1[\Omega]$ pour le même produit scalaire et la même norme. Il apparaıt dans la formulation variationnelle de problèmes aux limites dans lesquels la solution est fixée aux bornes de l'intervalle.

On admettra alors l'inégalité fondamentale suivante :

B.6.7 Inégalité de Poincaré

Théorème B.6.2 *Soit Ω un ouvert borné (au moins dans une direction) de $\mathbb{R}^2$ ou $\mathbb{R}^3$, on a pour tout $v \in H_0^1(\Omega)$ l'inégalité*

$$\|v\|_{0,2} \leq C(\Omega) \, \|grad \, v\|_{0,2}$$

D'où l'on déduit :

$$\|v\|_{1,2} \leq (1 + C(\Omega)^2)^{\frac{1}{2}} \, \|grad \, v\|_{0,2}$$

et donc que $|v|_{1,2} = \|grad \, v\|_{0,2}$ est une norme sur $H_0^1[\Omega]$ équivalente à la norme H^1.

B.6.8 Quelques résultats de densité

L'espace des fonctions polynômes variables est dense dans l'espace $L^2[\Omega]$. On peut alors construire des bases hilbertiennes de polynômes orthogonaux engendrant un sous-espace dense dans $L^2[\Omega]$. Ceci n'est pratiquement utile que dans le cas de domaines simples, par exemple rectangulaires ou parallélépipédiques.

L'espace des fonctions continues, polynomiales par morceaux, est dense dans $H^1[\Omega]$. Cet espace sera à la base de la méthode des éléments finis en dimension 2 et 3.

Suggestions non-exhaustives de lectures complémentaires

Nous donnons quelques indications de lecture en complément de cet ouvrage. Ceci n'est bien entendu pas une liste exhaustive. Certains ouvrages peuvent faire appel à des connaissances mathématiques relativement avancées chez le lecteur, nous les avons identifié avec une (*).

Tout d'abord, sur l'ensemble des techniques numériques, dont les éléments finis :

B. Mohammadi, J.-H. Saiac, *Pratique de la simulation numérique,* Dunod, 2003.

Sur l'analyse mathématique des EDP

(*) H. Brezis , *Analyse fonctionnelle,* Masson, 1983.

(*) R. Dautray, J-L Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques,* Masson 1984

(*) J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires,* Dunod, 1969.

(*) L. Schwartz, *Méthodes mathématiques de la physique,* Hermann, 1965.

I.P. Stavroulakis, S.A. Tersian, *Partial differential equations,* World Scientific, 1999.

Sur les méthodes numériques de base

G. Forsythe, M. Malcolm, C. Moler, *Computer methods for mathematical computations,* Prentice-Hall, 1977.

P. Moin, *Fundamentals of Engineering Numerical Analysis,* Cambridge University Press, 2001.

W. Press, S. Teukolsky, W. Vetterling, P. Flannery, *Numerical Recipes in C*, 2nd ed., Cambridge 1992.

Sur la résolution numérique des équations différentielles

M. Crouzeix, A. Mignot, *Analyse numérique des équations différentielles,* Masson, 1983.

(*) E. Hairer and S. Nørsett and G. Wanner, *Solving ordinary differential equations I, Nonstiff problems,* 2nd ed., Springer, 1993.

(*) E. Hairer and G. Wanner, *Solving ordinary differential equations II, Stiff and differential-algebraic problems,* Springer, 1991.

Sur l'analyse numérique matricielle

A. Björck, *Numerical methods for least squares problems,* SIAM, 1996.

P. Ciarlet, *Analyse numérique matricielle,* Masson, 1994.

P. Joly, *Analyse numérique matricielle,* Editions Cassini, 2002.

P. Lascaux, R. Théodor, *Analyse numérique matricielle,* Masson, 1993.

B.N. Parlett, *The symmetric Eigenvalue problem*, SIAM, 1998.

Y. Saad, *Numerical Methods for Large Eigenvalue Problems,* Manchester University Press, 1992.

Y. Saad, *Iterative methods for sparse linear systems,* PWS, 1996.

G. Strang, *Introduction to applied mathematics,* Wellesley-Cambridge Press, 1986.

N. Trefethen, D. Bau, *Numerical linear algebra*, SIAM, 1997.

J.H. Wilkinson, *The algebraic eigenvalue problem,* Clarendon Press, 1965.

Sur les différences finies

R. D. Richtmyer, K. W. Morton, *Difference Methods for Initial Value Problems,* Interscience, Wiley, 1967.

Sur les éléments finis

J.L. Batoz, G. Dhatt, *Modélisation des structures par éléments finis,* Hermès, 1992.

(*) F. Brezzi, M. Fortin, *Mixed and hybrid finite element methods,* Springer Verlag, 1991.

(*) P.G. Ciarlet, *The finite element method for elliptic problems,* North-Holland, 1978.

T.J.R. Hughes, *The Finite Element method,* Prentice Hall, 1987.

J.F. Imbert, *Analyse des structures par éléments finis,* Cepadues, 1984.

C. Johnson, *Numerical Solutions of Partial Differential Equations by the Finite Element Method,* Cambridge University Press, 1987.

B. Lucquin, O. Pironneau, *Introduction au calcul scientifique,* Masson-Wiley, 2000.

P.A. Raviart, J.M. Thomas, *Introduction à l'analyse numérique des équations aux dérivées partielles,* Masson, 1983.

O.C. Zienkiewicz, R.L. Taylor, *La méthode des éléments finis,* AFNOR, 1991.

Sur les volumes finis

(*) E. Godlewski, P.-A. Raviart, *Numerical approximation of hyperbolic systems of conservation laws,* Springer-Verlag, New-York, 1996.

R.J. LeVeque, *Numerical methods for conservation laws,* Birkhauser, 1999.

(*) J. Smoller, *Shock waves and reaction-diffusion equations,* Second edition, Springer-Verlag, New-York, 1994.

Sur l'optimisation et notions associées

(*) J. Céa, *Optimisation, théorie et algorithmes,* Dunod, 1971.

K. Chadan, D. Colton, L. Paivarina, W. Rundell, *An introduction to inverse scattering and inverse spectral problems,* SIAM Monographs on Mathematical Modeling and Computation, Philadelphia, PA, 1997.

R. Fletcher, *Practical methods of optimization,* Wiley, 2nd ed. 1987.

A. Griewank, *Evaluating derivatives*, SIAM,2000.

B. Mohammadi, O. Pironneau, *Applied shape Optimization for fluids,* Oxford University Press, 2001.

G. N. Vanderplaats, *Numerical optimization techniques for engineering design,* Mc Graw-Hill, 1986.

Sur l'analyse fréquentielle

R.N. Bracewell, *The Fourier Transform and its applications,* McGraw-Hill, New-York. 2ème édition, 1986.

C. Gasquet, P. Witomski, *Analyse de Fourier et applications,* Masson, 1990.

S. Mallat, *A wavelet tour of signal processing*, Academic Press Inc., 1998.

Sur la programmation scientifique

D. Bernardi, F.Hecht, O. Pironneau, *freefem+ documentation*, on the web at `www.freefem.org`, 2001.

F. Hecht, O. Pironneau, *Analyse numérique en C++,* collection SMAI, Springer, 2003.

W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling, *Numerical Recipes,* Cambridge University Press New York, 1986.

Sur le maillage

P.J. Frey, P.L. George, *Maillages*, Hermès, 1999.
P.L. George, *Automatic triangulation*, Wiley, 1996.

Sur les modèles mathématiques en finance

R. Jarrow, A. Rudd, *Option pricing*, R. Irwin publishing co. Illinois, 1983.
P. Wilmott, x. Howinson, J. Dewynne, *The mathematics of financial derivatives, a student introduction*, Cambridge U press, 1997.

Sur les modèles mathématiques et numériques en mécanique des fluides

(*) R. Glowinski, *Numerical Methods for Nonlinear Problems* , Springer-Verlag, 1984.
C. Hirsch, *Numerical Computation of Internal and External Flows,* Wiley, 1976.
(*) B. Mohammadi, O. Pironneau, *Analysis of the k-epsilon turbulence model,* Wiley, 1992.
R. Peyret, T.D. Taylor, *Computational methods for fluid flow,* Springer Verlag, 1982.
O. Pironneau, *Méthode des éléments finis en fluide,* Masson, 1995.

Index