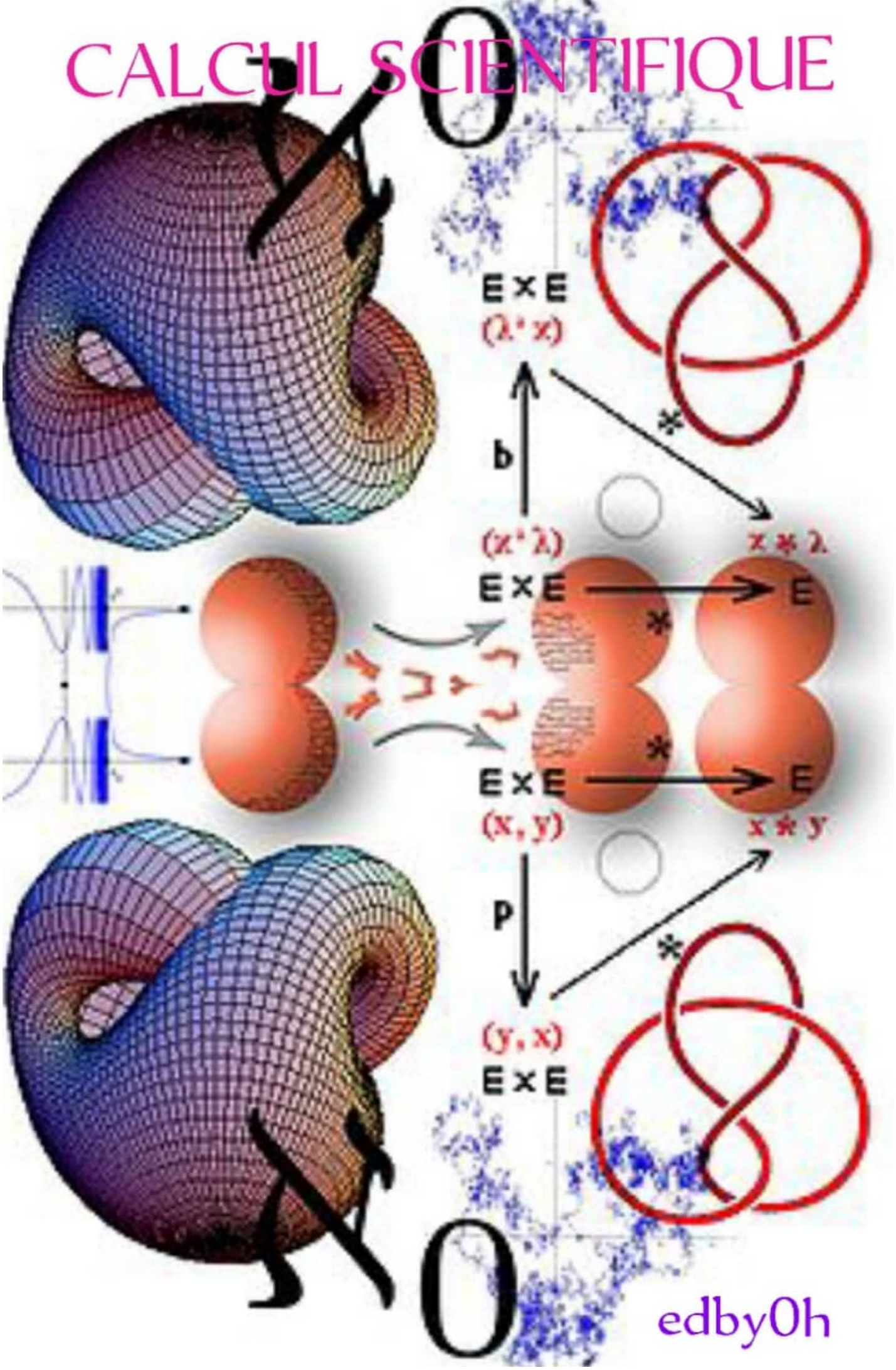
CALCUL SCIENTIFIQUE



edby0h

CALCUL SCIENTIFIQUE

Alexandre ERN et Gabriel STOLTZ

Décembre 2014

Chapitre 1

Avant-propos

1.1 Qu'est-ce que le calcul scientifique ?

Le calcul scientifique est une discipline aux contours pas toujours franchement définis, mais qui regroupe un ensemble de champs mathématiques et informatiques permettant la simulation numérique des phénomènes de la physique, chimie, biologie, et sciences appliquées en général. Son corollaire, la simulation numérique, fournit un outil efficace (parfois incontournable !) afin de prédire, comprendre, optimiser, voire contrôler le comportement de systèmes physiques relevant des sciences de l'ingénieur.

1.1.1 Anatomie d'un champ scientifique

L'approche d'un problème par le biais du calcul scientifique est une démarche globale, qui se passe en plusieurs temps successifs (avec en pratique des aller-retours d'une étape à l'autre), et dont toutes les étapes sont nécessaires :

(i) cela commence par la modélisation du système, qui consiste à décrire les phénomènes observés par le biais d'équations mathématiques, souvent en collaboration avec les scientifiques des disciplines applicatives concernées. Il est essentiel de connaître les hypothèses qui sous-tendent un modèle, leur domaine de validité et le comportement qualitatif que l'on attend des solutions avant d'en chercher une approximation. De nombreux modèles des sciences de l'ingénieur sont posés sous forme d'une équation aux dérivées partielles (EDP) ou d'un système d'EDPs ;

(ii) vient ensuite l'analyse théorique du modèle, et l'étude de ses propriétés (existence/unicité de la solution). Ceci peut faire intervenir des résultats profonds d'analyse, de théorie spectrale, de théorie des probabilités, etc ;

(iii) on propose ensuite une méthode numérique adaptée aux propriétés théoriques du modèle (préservant certains invariants par exemple), et on en fait l'analyse numérique. Ceci permet de déterminer la vitesse de convergence de la méthode numérique, sa stabilité. On peut également chercher des estimations d'erreurs *a priori* et *a posteriori* ;

(iv) vient enfin l'implémentation informatique de la méthode (avec éventuellement sa parallélisation sur un gros cluster de calcul), et sa validation sur des cas tests académiques pour vérifier le comportement de la méthode dans des situations bien connues ;

(v) si le travail s'arrête souvent là pour les mathématiciens, c'est en revanche à ce stade que commence la vraie aventure scientifique pour les chercheurs ou ingénieurs des domaines d'application, qui vont utiliser la nouvelle méthode sur des cas réels (et possiblement jusque dans ses derniers retranchements).

Comme cette sommaire description l'indique, le calcul scientifique est par essence un domaine interdisciplinaire, tant au sein des sciences en général qu'au sein des mathématiques (puisqu'il repose sur des champs aussi divers que l'analyse, l'analyse numérique, la théorie des probabilités, etc). Il est donc important d'avoir une bonne culture scientifique générale pour travailler dans ce domaine, et les étudiants des écoles d'ingénieurs française sont particulièrement qualifiés pour cela !

1.1.2 Moyen d'action : les méthodes numériques

Au cœur d'une démarche de calcul scientifique se trouve une méthode numérique, qui permet de calculer de manière approchée une propriété d'intérêt. Une telle méthode est fondée sur un algorithme implémenté informatiquement, son bras armé en quelque sorte. Un algorithme est une suite de tâches élémentaires qui s'enchaînent selon des règles précises, et qui est exécuté automatiquement par un langage informatique. Ce n'est pas une recette de cuisine ! Un élément important d'un algorithme est sa complexité, qui se reflète bien souvent dans le temps d'exécution. On évalue en particulier le nombre d'opérations arithmétiques élémentaires et le coût du stockage en mémoire.

1.1.3 Exemples d'applications

Un premier champ d'action pour le calcul scientifique concerne les situations où l'on ne peut pas (complètement) réaliser une expérience : ce qui se passe dans une centrale nucléaire qui s'emballe, la résistance de ladite centrale à un crash d'avion, la simulation du fonctionnement des armes nucléaires en l'absence d'essais, la conception d'un centre de stockage définitif des déchets nucléaires, le calcul de trajectoires de satellites, fusées. Dans toutes ces situations, une bonne simulation numérique permettra de donner quelques indications sur le comportement attendu de l'objet de la simulation – sans garantie totale que tout se passe comme prévu, bien sûr...

Il y a également des situations où il est moins cher de réaliser des tests numériques préliminaires : la simulation moléculaire des principes actifs pour l'industrie pharmaceutique, les tests de résistance mécanique (crash tests) dans l'industrie automobile, la synthèse de nouveaux matériaux pour l'industrie (alliages, polymères). Dans ces cas, la simulation numérique ne remplace pas une expérience, mais elle la complète, en suggérant des processus ou des comportements à tester spécifiquement de manière expérimentale.

Signalons enfin les situations où l'on cherche à anticiper des événements par le biais de la simulation. Cela concerne les méthodes numériques pour la finance (trading), et plus traditionnellement, la prévision météorologique ou climatique.

1.1.4 Objectifs du calcul scientifique

En fonction du problème que l'on cherche à résoudre, on peut avoir des objectifs différents. Pour mettre un peu de corps sur une phrase aussi générique, distinguons quatre points de vue :

(i) on peut souhaiter assurer la convergence de la méthode numérique : l'erreur sur le résultat final peut être rendue arbitrairement petite en y mettant les moyens ;

(ii) on peut lui préférer la précision : assurer que les erreurs que l'on commet sont petites par rappport à une tolérance fixée ;

(iii) on peut plutôt privilégier la fiabilité, qui est moins contraignante que la précision. Dans ce cas, on souhaite simplement garantir que l'erreur globale est en dessous d'une certaine tolérance. Ceci demande typiquement une validation de la méthode sur des cas tests ;

(iv) on peut enfin se concentrer sur l'efficacité de la méthode, et assurer que son coût de calcul est aussi petit que possible.

Evidemment, on souhaiterait avoir des résultats aussi convergés que possibles, et qui soient fiables dans tous les cas. Ce n'est cependant pas toujours possible, notamment lorsque l'on cherche à faire des estimations en temps réel (ou presque). Dans ce dernier cas, une fiabilité minimale mais une efficacité maximale seront plus opportunes. Il appartient à l'ingénieur de définir un compromis entre ces critères.

1.1.5 Analyse des sources d'erreurs

Le premier objectif à atteindre pour analyser les erreurs d'une simulation numérique est déjà de reconnaitre les différentes sources d'erreurs possibles ! On peut penser déjà aux erreurs dues à la modélisation mathématique du problème, résultant d'approximations dans la physique du problème – par exemple, négliger la viscosité et travailler avec les équations d'Euler plutôt que Navier-Stokes si le fluide est peu visqueux. Ces erreurs peuvent être évaluées lors de discussions avec les scientifiques des domaines applicatifs concernés.

On distingue également les erreurs dans les données d'entrée du problème : paramètres estimés par une mesure expérimentale ou par un autre modèle mathématique, conditions initiales, etc. Vous n'y pouvez rien *a priori*, mais il faut quand même faire quelque chose, ne serait-ce qu'étudier comment les incertitudes sur les données d'entrée se répercutent sur les données de sortie (valeurs des inconnues).

Mentionnons enfin les erreurs dans les algorithmes et méthodes numériques que l'on utilise : en pratique, on résout un problème approché, l'approximation résultant par exemple de la discrétisation d'un problème continu. Dans cette situation, nous avons des outils pour nous aider à quantifier précisément les erreurs introduites :

(a) les erreurs d'arrondi dues à la représentation machine des nombres et aux opérations arithmétiques effectuées (discutées sur un cas particuler en Section 2.1.6) ;

(b) les erreurs d'approximation des méthodes numériques, qui est le gros du travail pour un mathématicien appliqué. C'est le domaine par excellence de l'analyse numérique ;

(c) ne pas oublier les erreurs humaines... Mêmes développées et implémentées par des professionnels qualifiés, il se peut que les méthodes numériques soient entâchées d'un bug interne ! Pour éviter cela, on ne peut que recommander la validation des résultats de simulation par des approches variées et complémentaires.

1.2 Objectifs et organisation de ce cours

L'objectif de ce cours est de présenter de manière introductive trois « piliers » du calcul scientifique moderne :

(1) des méthodes numériques d'**intégration**, allant du calcul d'intégrales à la résolution numérique d'équations aux dérivées partielles (chapitre 2) ;

(2) l'**optimisation** (chapitre 3) avec ou sans contrainte et quelques algorithmes d'optimisation numérique ;

(3) la **méthode des éléments finis** (chapitre 4) et son application à l'approximation de problèmes elliptiques (linéaires et stationnaires) issus de la modélisation mécanique ou thermique.

Chaque chapitre contient, outre le cours proprement dit, une dizaine d'exercices d'application avec un corrigé détaillé.

Nous exprimons nos remerciements à Miguel Fernandez, Ludovic Goudenège, Laurent Monasse et Rachida Chakir pour leur contribution en tant que membres de l'équipe pédagogique. Merci également à Sébastien Boyaval, Eric Cancès, Michel de Lara, Daniele Di Pietro, Jean-Frédéric Gerbeau, Antoine Gloria, Tony Lelièvre, Olivier le Maître, Serge Piperno et Bruno Sportisse pour leur contribution passée.

1.3 Bibliographie

La littérature sur le calcul scientifique est extrêmement vaste. Voici une première liste (réduite) d'ouvrages, en français ou en anglais, qui complètent et prolongent ce cours.

- R. L. Burden et J. D. Faires, *Numerical analysis*, 5th Edition, PWS Publishing Company, Boston (1993).
- J.-P. Demailly, *Analyse numérique et équations différentielles*, Collection Grenoble Sciences (EDP Sciences, 2006).
- A. Ern, *Aide-mémoire des éléments finis*, Dunod, collection L'Usine Nouvelle (2005).
- A. Ern et J.-L. Guermond, *Eléments finis : théorie, applications, mise en œuvre*, Springer, collection SMAI mathématiques et applications, **36**, Heidelberg (2002).
- E. Godlewski et P.-A. Raviart, *Numerical approximation of hyperbolic systems of conservation laws*, Applied Mathematical Sciences, **118**, Springer, New York (1996).
- D. Goldberg, What every computer scientist should know about floating-point arithmetic, *ACM Computing Surveys* **23**(1) (1991).
- G.H. Golub et C.F. van Loan, *Matrix computations*, John Hopkins University Press, Baltimore (1983).
- R.J. LeVeque, *Numerical methods for conservation laws*, Birkhäuser, Basel (1992).
- A. Quarteroni, R. Sacco et F. Saleri, *Numerical mathematics*, Texts in Applied Sciences, **37**, Springer, New York (2000).
- J. Rappaz et M. Picasso, *Introduction à l'analyse numérique*, Presses polytechniques et universitaires romandes, Lausanne (1998).
- L. Sainsaulieu, *Calcul scientifique*, Masson, Paris (1996).

Chapitre 2

Intégration numérique

Ce chapitre présente des méthodes numériques d'intégration pour les équations différentielles. Plus précisément, on considère

(i) l'intégration en temps d'un problème de Cauchy émanant d'une équation différentielle ordinaire (Section 2.1). Dans ce cas, seule une discrétisation de la variable temporelle est nécessaire ;

(ii) l'intégration en temps d'un problème de Cauchy émanant d'une équation aux dérivées partielles (Section 2.2), pour lesquelles une discrétisation à la fois en espace et en temps est requise.

En complément, nous évoquons le calcul d'intégrales de fonctions sur un domaine donné (Section 2.3), pour lesquelles seule une discrétisation spatiale doit être considérée. Comme nous allons le voir, le mot-clé « intégration numérique » recouvre ainsi des réalités très différentes en fonction du contexte !

2.1 Intégration des équations différentielles ordinaires

L'objectif de cette section est de présenter quelques méthodes numériques pour approcher des solutions d'équations différentielles ordinaires (EDOs), qui sont un problème de Cauchy de la forme

suivante :

$$\dot{y}(t) = f(t, y(t)), \qquad y(0) = y_0, \tag{2.1}$$

où $y : \mathbb{R}_+ \to \mathbb{R}^d$ est une fonction du temps $t \geqslant 0$ à valeurs vectorielles, et $f : \mathbb{R}^d \to \mathbb{R}^d$ est un champ de vecteurs. On peut réécrire ce problème sous une formulation intégrale, qui peut être plus agréable pour l'étude théorique car elle demande moins d'hypothèses de régularité sur les objets en jeu, et qui est également utile pour proposer des schémas numériques :

$$y(t) = y_0 + \int_0^t f(s, y(s))\, ds. \tag{2.2}$$

Cette formulation est équivalente à (2.1) si f est continue.

Nous commençons par présenter quelques applications en Section 2.1.1 (et notamment un problème modèle, la dynamique du système solaire). Nous nous tournons ensuite vers l'étude mathématique du problème continu (2.1) en Section 2.1.2, et discutons en particulier son caractère bien posé. Une fois ceci établi, on peut sereinement se tourner vers l'approximation numérique de (2.1) par le biais des méthodes à un pas, les plus simples (voir Section 2.1.3). L'analyse de l'erreur ainsi commise est menée en Section 2.1.4. On peut également mener plus loin l'étude de systèmes qui ont des propriétés ou une structure particulières, comme les systèmes linéaires dissipatifs (voir la Section 2.1.5). Enfin, pour le lecteur motivé, des éléments d'analyse complémentaires sont rassemblés en Section 2.1.6.

2.1.1 Motivation : dynamique céleste

On a besoin dans certains cas de savoir calculer numériquement avec une grande précision la solution d'une EDO : par exemple pour déterminer la trajectoire de la fusée qui va mettre en orbite un satellite, ou de la sonde spatiale qui va passer au ras de Jupiter pour ensuite aller explorer les confins de notre univers, ou de la météorite que l'on voit arriver près de la Terre (nous touchera-t-elle ou non ?). Dans ces cas, on se donne un horizon de temps fini et on cherche à reproduire au mieux la trajectoire du système sur ce temps.

Il y a d'autres situations où on s'intéresse plutôt à un résultat en temps long, par exemple la convergence vers une trajectoire périodique ou un cycle : c'est cet élément géométrique asymptotique qui sera l'objet de nos soins. Citons par exemple l'équation de Lotka–Volterra, qui est un modèle simplifié de dynamique des populations ; ou l'intégration en temps long de la dynamique Hamiltonienne pour calculer des moyennes microcanoniques en physique statistique, ou en dynamique céleste, pour déterminer la stabilité des orbites d'un système planétaire.

Citons également des problèmes mélangeant plusieurs échelles de temps : un cas frappant est celui de la cinétique chimique entrant dans les modèles de pollution de l'air ou en génie chimique. Certaines transformations dans ces systèmes sont très rapides et/ou oscillantes (constantes de réaction très grandes ou concentrations importantes), alors que d'autres sont très lentes. Une bonne méthode numérique devrait pouvoir traiter toutes ces échelles de temps simultanément, et ne pas caler le pas de temps sur les événements les plus rapides, sans quoi les événements les plus lents ne pourront pas être résolus.

Enfin, les méthodes d'intégration en temps des EDOs constituent une brique fondamentale pour l'intégration en temps d'équations plus compliquées : les équations aux dérivées partielles, où il faut considérer à la fois une discrétisation en espace et en temps (voir Section 2.2) ; ou les équations différentielles stochastiques, rencontrées notamment dans le cadre de la finance quantitative ou en physique statistique numérique.

Un exemple précis : la dynamique céleste

On décrit ici un système qui correspond à la partie « extérieure » du système solaire : on représente Jupiter, Saturne, Uranus, Neptune et Pluton (positions respectives q_i pour $i = 1, \ldots, 5$), et le Soleil auquel on agrège en fait les quatre planètes intérieures que sont Mercure, Vénus, la

Terre et Mars (position q_0). L'énergie potentielle du sytème est

$$V(q) = \sum_{0 \leqslant i < j \leqslant 5} v_{ij}(|q_i - q_j|_{\ell^2}), \qquad v_{ij}(r) = -G\frac{m_i m_j}{r},$$

où $|\cdot|_{\ell^2}$ désigne la norme euclidienne. On utilise des unités réduites pour l'implémentation informatique (afin de manipuler des quantités qui sont toutes d'ordre 1). L'unité de masse est donnée par la masse du soleil $1,9891 \times 10^{30}$ kg ; l'unité de longueur est la distance Terre-Soleil, à savoir 149 597 870 km ; et l'unité de temps est un jour sur Terre, soit $8,64 \times 10^3$ s. Dans ces unités, la constante de gravitation $G = 6,67384 \times 10^{-11}$ $\mathrm{m^3 kg^{-1} s^{-2}}$ vaut $2,95995 \times 10^{-4}$. Les masses des planètes sont notées m_i, et $p_i = m_i v_i$ est leur quantité de mouvement.

L'énergie totale du système dans la configuration $(q,p) \in \mathbb{R}^{6N}$ est donnée par le Hamiltonien

$$H(q,p) = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} + V(q).$$

L'évolution en temps est régie par la dynamique Hamiltonienne

$$\begin{cases} \dot{q}_i(t) = \dfrac{\partial H}{\partial p_i} = \dfrac{p_i(t)}{m_i}, \\ \dot{p}_i(t) = -\dfrac{\partial H}{\partial q_i} = -\nabla_{q_i} V(q(t)). \end{cases} \tag{2.3}$$

Noter que, quitte à introduire l'inconnue $y = (q,p)$, on peut récrire cette dynamique sous la forme générale

$$\dot{y}(t) = f(y), \qquad f(y) = \begin{pmatrix} \nabla_p H \\ -\nabla_q H \end{pmatrix}. \tag{2.4}$$

Un calcul simple (le faire en exercice) montre que l'énergie du système est constante au cours du temps : $H(q(t),p(t)) = H(q_0,p_0)$. Une notion de stabilité pertinente est par exemple la conservation de l'énergie totale en temps long.

2.1.2 Etude du problème continu

Existence et unicité locales

Pour discuter l'existence et l'unicité des solutions du problème (2.1), on utilise le théorème de Cauchy–Lipschitz, qui donne l'existence locale et l'unicité si le champ de force f est localement Lipschitzien : pour tout $(t_0,y_0) \in \mathbb{R}_+ \times \mathbb{R}^d$, il existe $r,\tau,L > 0$ (dépendant de t_0, y_0 *a priori*) tels que

$$\forall (t,y_1,y_2) \in]t_0-\tau, t_0+\tau[\times B(y_0,r)^2, \qquad |f(t,y_1) - f(t,y_2)| \leqslant L|y_1 - y_2|, \tag{2.5}$$

où $B(y_0,r)$ est la boule (ouverte) de centre y_0 et de rayon r et où $|\cdot|$ désigne une norme sur $\mathbb{R}^d$ (toutes les normes sont équivalentes en dimension finie, et le choix de la norme n'a d'incidence que sur la valeur numérique de L). Cela définit une unique solution sur un intervalle de temps maximal $[0, t_{\max}[$. Si on n'a pas de solution globale (au sens où $t_{\max} < +\infty$), alors on a explosion de la solution en temps fini : $|y(t)| \to +\infty$ lorsque $t \to t_{\max}$. La condition (2.5) est vérifiée par exemple si, pour tout $t \in \mathbb{R}_+$, la fonction $f(t,\cdot)$ est localement Lipschitzienne en y : pour tout $y_0 \in \mathbb{R}^d$, il existe $r(t) \geqslant r_\flat > 0$ et $\Lambda_f(t) \leqslant \Lambda_\sharp$ tels que

$$\forall (y_1,y_2) \in B(y_0,r(t))^2, \qquad |f(t,y_1) - f(t,y_2)| \leqslant \Lambda_f(t)|y_1 - y_2|. \tag{2.6}$$

Lorsque la propriété ci-dessus est valable pour tout $(y_1,y_2) \in (\mathbb{R}^d)^2$, on dit que la fonction $f(t,\cdot)$ est globalement Lipschitzienne.

Existence et unicité globales

L'existence et l'unicité de la solution globale est assurée dans certains cas.

(i) Le premier cas est celui où $f(t,\cdot)$ est uniformément Lipschitzienne en y, avec une constante de Lipschitz $L(t)$ continue.

(ii) Le second cas est celui où la croissance de f est au plus affine :

$$|f(t,y)| \leqslant c(t) + L(t)|y|,$$

avec des fonctions c, L localement intégrables. On a en effet, en partant de la formulation intégrale (2.2),

$$|y(t)| \leqslant |y_0| + \int_0^t \Big(c(s) + L(s)|y(s)|\Big) ds. \tag{2.7}$$

En introduisant

$$M(t) = \int_0^t L(s)|y(s)|\,ds \geqslant 0, \qquad a(t) = L(t)\Big(|y_0| + \int_0^t c(s)\,ds\Big),$$

on peut reformuler (2.7) sous la forme $\dot{M}(t) \leqslant a(t) + L(t)M(t)$, d'où, par le lemme de Gronwall, [1]

$$0 \leqslant M(t) \leqslant \int_0^t a(s) \exp\left(\int_s^t L\right) ds.$$

En reportant dans (2.7), on obtient bien une borne supérieure finie pour $|y(t)|$.

(iii) Un dernier cas est celui où il existe une fonction de Lyapounov, c'est-à-dire une fonction $W \in C^1(\mathbb{R}^d)$ telle que $W(x) \to +\infty$ lorsque $|x| \to +\infty$. Comme W est alors uniformément minorée, on peut lui ajouter une constante de telle manière à ce que $W \geqslant 1$. Enfin, on demande que

$$f(x) \cdot \nabla W(x) \leqslant c < +\infty.$$

Dans ce cas, on a alors

$$\frac{d}{dt}\Big[W\Big(y(t)\Big)\Big] = \Big(f \cdot \nabla W\Big)(y(t)) \leqslant c,$$

ce qui montre que $W(y(t)) \leqslant W(y_0)\,\mathrm{e}^{ct}$, et assure ainsi que la norme de la solution ne peut pas exploser en temps fini. Cette technique est très utile pour établir l'existence et l'unicité dans le cas où f n'est pas globalement Lipschitzienne et croît plus que linéairement à l'infini. On peut appliquer cela au système très simple $\dot{y}(t) = -y(t)^3$ pour lequel $W(x) = 1 + x^4$ est une fonction de Lyapounov.

Remarque (Stabilité)**.** Pour étudier le caractère bien posé de (2.1), une analyse standard demande de se pencher sur la stabilité des solutions par rapport à des perturbations de la dynamique. Il existe plusieurs notions de stabilité, telles que la stabilité au sens de Lyapunov (des perturbations au champ de force, uniformément majorées par ε, ne peuvent engendrer que des modifications d'ordre $C(T)\varepsilon$ à la solution au temps T, avec une constante qui dépend typiquement exponentiellement de T) ou la stabilité asymptotique (qui est pertinente pour les systèmes dissipatifs : si les perturbations tendent vers 0 en temps long, alors la solution du système perturbé reste uniformément proche en la taille de la perturbation de la solution du système de référence). □

1. Rappelons rapidement le calcul : on note que

$$\frac{d}{dt}\left[M(t)\,\exp\left(-\int_0^t L(s)\,ds\right)\right] = \Big(\dot{M}(t) - L(t)M(t)\Big)\exp\left(-\int_0^t L(s)\,ds\right) \leqslant a(t)\,\exp\left(-\int_0^t L(s)\,ds\right),$$

d'où, par intégration et en tenant compte du fait que $M(0) = 0$,

$$M(t)\,\exp\left(-\int_0^t M(s)\,ds\right) \leqslant \int_0^t a(s)\,\exp\left(-\int_0^s M(r)\,dr\right)\,ds,$$

ce qui permet de conclure en multipliant les deux membres de l'inégalité par $\exp\left(\int_0^t M(s)\,ds\right)$.

2.1.3 Approximation par les méthodes à un pas

On va à présent décrire des méthodes numériques pour approcher la solution de (2.1) sur un intervalle de temps fini $[0, T]$. Plus précisément, on va considérer des temps $t_0 = 0 < t_1 < \cdots < t_N = T$, et on va noter y^n l'approximation numérique de la solution exacte $y(t_n)$. Par la suite, on notera $\Delta t_n = t_{n+1} - t_n$ les incréments de temps strictement positifs. Souvent, on choisira un pas de temps uniforme $\Delta t > 0$, auquel cas $t_n = n\Delta t$, le nombre total de pas d'intégration étant [2] $N = T/\Delta t$.

Pour rester le plus simple possible, nous évoquerons seulement les méthodes à un pas, et non les méthodes multi-pas, bien qu'elles soient en général plus précises à coût de calcul fixé (cependant, leur stabilité demande une attention particulière).

Principe de l'approximation. La construction des méthodes à un pas repose sur une discrétisation de la formulation intégrale (2.2) sur l'intervalle de temps $[t_n, t_{n+1}]$ par une règle de quadrature. De manière abstraite, on peut ainsi écrire une relation de récurrence permettant de calculer itérativement la trajectoire numérique

$$y^{n+1} = y^n + \Delta t_n \Phi_{\Delta t_n}(t_n, y^n), \tag{2.8}$$

où $\Phi_{\Delta t_n}(t_n, y^n)$ est une approximation de

$$\frac{1}{t_{n+1} - t_n} \int_{t_n}^{t_{n+1}} f\Big(s, y(s)\Big) ds.$$

Les méthodes ainsi obtenues sont appelées méthodes de Runge–Kutta. Elles sont décomposées en deux catégories selon qu'elles sont *explicites* (la nouvelle configuration peut être obtenue directement de la précédente) ou *implicites* (pour obtenir la nouvelle configuration, il faut résoudre un problème linéaire ou nonlinéaire en fonction de la dépendance de f en y). Dans le cas implicite, le schéma numérique n'est pas spontanément écrit sous la forme (2.8). Toutefois, on préfère toujours écrire l'incrément $y^{n+1} - y^n$ comme une fonction de y^n seulement pour mettre en exergue le fait qu'un schéma numérique pour les EDOs fonctionne de manière itérative : il suffit de connaitre une approximation de l'état du système y^n au temps t_n, et l'incrément de temps Δt_n, pour en déduire une approximation de l'état au temps $t_n + \Delta t_n$.

Donnons à présent quelques exemples de méthodes numériques pour illustrer notre propos :

(1) Méthodes explicites :

(i) Euler explicite : $y^{n+1} = y^n + \Delta t_n\, f(t_n, y^n)$;

(ii) méthode de Heun : $y^{n+1} = y^n + \dfrac{\Delta t_n}{2}\Big(f(t_n, y^n) + f\big(t_{n+1}, y^n + \Delta t_n f(t_n, y^n)\big)\Big)$;

(iii) schéma de Runge–Kutta d'ordre 4 : on calcule les points intermédiaires

$$\begin{cases} F_1 = f(t_n, y^n) \\ F_2 = f\left(t_n + \dfrac{\Delta t_n}{2}, y^n + \dfrac{\Delta t_n}{2} F_1\right) \\ F_3 = f\left(t_n + \dfrac{\Delta t_n}{2}, y^n + \dfrac{\Delta t_n}{2} F_2\right) \\ F_4 = f(t_n + \Delta t_n, y^n + \Delta t_n F_3), \end{cases}$$

et on pose

$$y^{n+1} = y^n + \Delta t\, \frac{F_1 + 2F_2 + 2F_3 + F_4}{6} \ ;$$

(2) Méthodes implicites :

2. On suppose que ce nombre est entier ; sinon on peut toujours changer un peu le pas de temps pour que ce soit le cas, en prenant par exemple la partie entière de $T/\Delta t$.

(i) Euler implicite : $y^{n+1} = y^n + \Delta t_n\, f(t_{n+1}, y^{n+1})$;

(ii) méthode des trapèzes (ou Crank–Nicolson) : $y^{n+1} = y^n + \dfrac{\Delta t_n}{2}\Big(f(t_n, y^n) + f(t_{n+1}, y^{n+1})\Big)$;

(iii) méthode du point milieu : $y^{n+1} = y^n + \Delta t_n f\left(\dfrac{t_n + t_{n+1}}{2}, \dfrac{y^n + y^{n+1}}{2}\right)$.

On peut bien sûr construire des méthodes plus compliquées et plus précises, et nous renvoyons le lecteur à la bibliographique pour des méthodes de Runge–Kutta d'ordre supérieur (explicites ou implicites). Pour les méthodes explicites, on identifie assez simplement la fonction d'incrément $\Phi_{\Delta t}$ dans (2.8). Pour les schémas implicites, cette tâche est moins facile. Prenons l'exemple du schéma d'Euler implicite : l'application $\Phi_{\Delta t}$ est définie de manière implicite par la relation

$$\Phi_{\Delta t_n}(t_n, y^n) = f\Big(t_n + \Delta t_n, y^n + \Delta t_n \Phi_{\Delta t_n}(t_n, y^n)\Big).$$

Comme nous allons le montrer dans le point suivant, on peut s'assurer que $\Phi_{\Delta t_n}(t_n, y^n)$ est bien définie si Δt_n est assez petit.

Un mot sur l'implémentation des schémas implicites. Dans les schémas implicites, il faut résoudre une équation nonlinéaire pour obtenir l'approximation y^{n+1} partant de y^n. Il y a donc un surcoût de calcul dans l'utilisation de ces schémas, mais qui vaut souvent le coup car on a une stabilité accrue et on peut utiliser des pas de temps plus grands.

Avant de chercher à résoudre numériquement l'équation donnant y^{n+1}, il faut déjà garantir que y^{n+1} existe et est unique ! On peut utiliser pour ce faire un théorème des fonctions implicites, ou un théorème de point-fixe de Banach. Par exemple, le schéma d'Euler implicite $y^{n+1} = y^n + \Delta t_n\, f(t_{n+1}, y^{n+1})$ peut se réécrire

$$y^{n+1} = F(y^{n+1}), \qquad F(y) = y^n + \Delta t_n f(t_{n+1}, y).$$

On a existence et unicité de y^{n+1} lorsque la fonction $f(t_{n+1}, \cdot)$ est globalement Lipschitzienne, cf. (2.6) pour la notation, et que le pas de temps est suffisamment petit pour satisfaire

$$\Delta t_n \Lambda_f(t_{n+1}) < 1, \tag{2.9}$$

car cette propriété implique que l'application F est contractante :

$$|F(y_1) - F(y_2)| \leqslant \Delta t_n \Lambda_f(t_{n+1})\, |y_1 - y_2|.$$

La condition (2.9) impose donc une borne supérieure sur les pas de temps admissibles. Notons qu'une analyse plus fine permettrait de remplacer la norme L^∞ globale par une majoration locale au voisinage du point y^n, et de considérer des incréments de temps variables (voir Section 2.1.6 pour des stratégies de pas de temps adaptatifs).

Le calcul pratique des itérations d'un schéma implicite peut se faire par une méthode numérique s'inspirant de la stratégie de point fixe utilisée pour montrer l'existence et l'unicité de y^{n+1} : on part d'un premier essai obtenu par une méthode explicite, que l'on affine ensuite par des itérations de point-fixe, d'où le nom de stratégie « prédicteur/correcteur ». Illustrons cela pour le schéma d'Euler implicite. On peut partir d'un état obtenu par un schéma d'Euler explicite

$$z^{n+1,0} = y^n + \Delta t_n f(t_n, y^n),$$

et le corriger ensuite par des itérations de point-fixe selon

$$z^{n+1,k+1} = y^n + \Delta t_n f(t_{n+1}, z^{n+1,k}).$$

Sous les bonnes hypothèses précédentes, on a $z^{n+1,k} \xrightarrow[k\to+\infty]{} y^{n+1}$. En pratique, on n'effectue qu'un nombre fini d'itérations de point-fixe, en utilisant un critère de convergence tel que

$$|z^{n+1,k+1} - z^{n+1,k}| \leqslant \varepsilon |z^{n+1,0}|$$

pour une tolérance $\varepsilon > 0$ donnée (le facteur multiplicatif $|z^{n+1,0}|$ à droite permet de considérer un critère de convergence indépendant de l'échelle ou unités de la variable y). On observe souvent une très bonne convergence dans les itérations de point-fixe, avec des erreurs relatives de l'ordre de 10^{-8} en quelques itérations. Une alternative aux itérations de point-fixe est d'utiliser un algorithme de Newton, qui a une convergence plus rapide, mais qui demande que l'on parte suffisamment près de la solution y^{n+1}.

2.1.4 Analyse d'erreur

L'objectif de l'analyse d'erreur *a priori* est de donner une estimation de l'erreur commise par la méthode numérique en fonction des paramètres du problème (temps d'intégration, pas de temps, champ de force). L'idée générale est de remarquer qu'à chaque pas de temps on commet une erreur d'intégration locale (erreur de troncature dans la discrétisation de l'intégrale, à laquelle s'ajoutent souvent des erreurs d'arrondi), et que ces erreurs locales s'accumulent. Le contrôle de cette accumulation demande l'introduction d'une notion de stabilité adéquate, alors que les erreurs locales sont liées à une notion de consistance. On peut montrer qu'une méthode numérique stable et consistante est convergente. Insistons sur le fait que ce type résultat, quoique courant en analyse numérique, est un pilier de l'analyse d'erreur *a priori*. Pour mesurer l'erreur, on choisit une norme sur $\mathbb{R}^d$ que l'on note $|\cdot|$; toutes les normes étant équivalentes en dimension finie et la dimension étant fixée, le choix particulier de la norme n'est pas essentiel ici.

Erreur de troncature locale

L'erreur de troncature locale est l'erreur résiduelle que l'on obtiendrait si on appliquait le schéma numérique à la solution exacte. Elle est donc définie à l'itération n comme la différence entre la solution exacte à l'itération suivante, à savoir $y(t_{n+1})$, et l'approximation numérique obtenue en partant de $y(t_n)$, à savoir $y(t_n) + \Delta t_n \Phi_{\Delta t_n}(t_n, y(t_n))$, le tout divisé par Δt_n. Ainsi, l'erreur de troncature locale vaut par définition

$$\eta^{n+1} := \frac{y(t_{n+1}) - y(t_n) - \Delta t_n \Phi_{\Delta t_n}(t_n, y(t_n))}{\Delta t_n}. \tag{2.10}$$

Notons que l'erreur de troncature a la même dimension physique que la dérivée en temps de y. La normalisation que nous employons dans ce cours permet d'interpréter η comme une erreur par unité de temps.

Définition 2.1. *On note $\Delta t = \max_{0 \leqslant n \leqslant N-1} \Delta t_n$ le pas de temps maximal. On dit qu'un méthode numérique est consistante si*

$$\lim_{\Delta t \to 0} \left(\max_{1 \leqslant n \leqslant N} |\eta^n| \right) = 0, \tag{2.11}$$

et qu'elle est consistante d'ordre p s'il existe une constante C telle que, pour tout $0 \leqslant n \leqslant N-1$,

$$|\eta^{n+1}| \leqslant C(\Delta t_n)^p. \tag{2.12}$$

La constante C dépend en général de la régularité de la solution exacte.

Les preuves de consistance reposent sur des développements de Taylor de la solution exacte, et demandent donc de la régularité sur le champ de force f. On supposera toujours que le champ de force est aussi régulier que nécessaire par la suite. Notons que la régularité de la solution y découle de celle du champ de force. En effet, si f est continue, alors la solution de (2.1) est C^1. Si f est C^1, on voit alors que le membre de droite de (2.1) est C^1 (par composition) et donc que la solution y est C^2 (par intégration). On peut itérer cet argument et montrer ainsi que $y \in C^{l+1}$ si $f \in C^l$.

Exemple. Le schéma d'Euler explicite est consistant d'ordre 1. L'erreur de troncature s'écrit

$$\eta^{n+1} = \frac{y(t_n + \Delta t_n) - \Big(y(t_n) + \Delta t_n \, f(t_n, y(t_n)) \Big)}{\Delta t_n}.$$

Or, par application d'une formule de Taylor avec reste exact autour de $y(t_n)$, on voit qu'il existe $\theta^n \in [0,1]$ tel que

$$y(t_n + \Delta t_n) - \Big(y(t_n) + \Delta t f(t_n, y(t_n))\Big) = \frac{\Delta t_n^2}{2} y''(t_n + \theta^n \Delta t_n). \tag{2.13}$$

Par ailleurs, la dérivée seconde $y''(t)$ peut s'exprimer en fonction des dérivées de f en dérivant (2.1) par rapport au temps, ce qui donne

$$y''(\tau) = \partial_t f(\tau, y(\tau)) + \partial_y f(\tau, y(\tau)) \cdot f(\tau, y(\tau)). \tag{2.14}$$

On voit ainsi que y'' est uniformément borné en temps sur tout intervalle de la forme $[0,T]$ ($T < +\infty$) si la fonction f et ses dérivées sont continues (la trajectoire restant bornée dans ce cas). On en déduit finalement que

$$\left|\eta^{n+1}\right| \leqslant C \Delta t_n,$$

avec la constante $C = \frac{1}{2} \sup_{t\in[0,T]} |y''(t)|$.

Stabilité

La notion de stabilité quantifie la robustesse de l'approximation numérique par rapport à des perturbations. On donne ici la définition pour des schémas à pas fixe. On fixe un intervalle de temps $[0,T]$ et un pas de temps $\Delta t > 0$ constant pour simplifier, et on note $N = T/\Delta t$ le nombre d'itérations correspondantes.

Définition 2.2 (Stabilité)**.** *On dit qu'une méthode numérique* (2.8) *est stable s'il existe une constante $S(T) > 0$ (qui dépend du temps d'intégration $T = N\Delta t$ mais pas de N ou de Δt tout seul) telle que, pour toute suite $z = \{z^n\}_{1\leqslant n\leqslant N}$ partant de la même condition initiale $z^0 = y^0$ et vérifiant*

$$\begin{cases} y^{n+1} = y^n + \Delta t \Phi_{\Delta t}(t_n, y^n), \\ z^{n+1} = z^n + \Delta t \Phi_{\Delta t}(t_n, z^n) + \Delta t\, \delta^{n+1}, \end{cases} \tag{2.15}$$

on ait

$$\max_{1\leqslant n\leqslant N} |y^n - z^n| \leqslant S(T)\, \Delta t \sum_{n=1}^{N} |\delta^n|. \tag{2.16}$$

L'extension de la notion de stabilité au cas des pas de temps variables est très simple : la majoration (2.16) devient $\max_{1\leqslant n\leqslant N} |y^n - z^n| \leqslant S(T) \sum_{n=1}^{N} \Delta t_n |\delta^n|$.

Il est utile, pour la suite, de noter que l'on peut réécrire la stabilité sous une forme plus condensée, en introduisant les normes suivantes pour des suites de vecteurs $u = \{u^n\}_{1\leqslant n\leqslant N}$: une norme ℓ_t^∞

$$\|u\|_{\ell_t^\infty} = \max_{1\leqslant n\leqslant N} |u^n|\,,$$

et une norme ℓ_t^1 :

$$\|u\|_{\ell_t^1} = \Delta t \sum_{n=1}^{N} |u^n|\,.$$

L'indice t souligne le fait que les normes sont par rapport à des variables temporelles. Cette notation prendra son sens à la Section 2.2 lorsque l'on considérera des schémas pour les équations aux dérivées partielles, pour lesquelles on introduit des discrétisations à la fois en temps et en espace. La condition de stabilité peut alors se récrire sous la forme compacte

$$\|y - z\|_{\ell_t^\infty} \leqslant S(T)\, \|\delta\|_{\ell_t^1}. \tag{2.17}$$

Intéressons nous à présent à l'obtention de conditions suffisantes de stabilité. Un cas simple est celui où $\Phi_{\Delta t}$ est uniformément Lipschitzienne en y, c'est-à-dire qu'il existe $\Lambda_\Phi > 0$ tel que, pour tout $y_1, y_2 \in \mathbb{R}^d$, on ait

$$|\Phi_{\Delta t}(t, y_1) - \Phi_{\Delta t}(t, y_2)| \leqslant \Lambda_\Phi |y_1 - y_2|.$$

On peut alors prendre $S(T) = \mathrm{e}^{\Lambda_\Phi T}$. Cette assertion repose sur l'estimation suivante :

$$|y^{n+1} - z^{n+1}| \leqslant \Delta t|\delta^{n+1}| + (1 + \Delta t\Lambda_\Phi)|y^n - z^n| \leqslant \Delta t|\delta^{n+1}| + \exp(\Lambda_\Phi \Delta t)|y^n - z^n|,$$

et l'utilisation d'un lemme de Gronwall discret (dont on fait la preuve par récurrence) qui fournit l'estimation

$$|y^n - z^n| \leqslant \mathrm{e}^{\Lambda_\Phi T}\Delta t \sum_{m=1}^{n} |\delta^m|.$$

Remarque. Noter que les constantes de stabilité qui apparaissent croissent *a priori* de manière exponentielle avec le temps. En fait, Λ_Φ^{-1} peut être interprété comme un temps caractéristique de variation du système. La constante de stabilité peut donc être très grande si on intègre la dynamique sur des temps T bien supérieurs à Λ_Φ^{-1}. Pour des systèmes particuliers, tels que les systèmes linéaires dissipatifs étudiés à la section 2.1.5, on montre que les constantes de stabilité restent bornées. □

Convergence

Une méthode numérique est convergente si l'erreur globale vérifie

$$\max_{1\leqslant n\leqslant N} |y^n - y(t_n)| \to 0$$

lorsque $y^0 = y(t_0)$ et $\Delta t = \max_{0\leqslant n\leqslant N-1} |t_{n+1} - t_n| \to 0$. On néglige dans cette section les erreurs d'arrondis dues à la représentation machine des nombres réels (ceci sera traité plus loin). On a alors le résultat suivant.

Théorème 2.3. *Une méthode stable et consistante est convergente.*

La preuve de ce résultat est très simple : on remplace z^n dans (2.15) par la solution exacte $y(t_n)$, ce qui correspond à choisir $\delta^{n+1} = \eta^{n+1}$, l'erreur de troncature locale définie en (2.10). On part en revanche de la même condition initiale. La stabilité donne donc

$$\max_{1\leqslant n\leqslant N} |y^n - y(t_n)| \leqslant S(T)\,\Delta t \sum_{n=1}^{N} |\eta^n| \leqslant S(T)\,T \max_{1\leqslant n\leqslant N} |\eta^n|. \tag{2.18}$$

Par définition de la consistance, on voit que le membre de droite tend vers zéro avec Δt. De plus, si la méthode numérique est consistante d'ordre p, alors l'erreur globale est également d'ordre Δt^p puisque

$$\max_{1\leqslant n\leqslant N} |y^n - y(t_n)| \leqslant S(T)\,T\,C\,\Delta t^p. \tag{2.19}$$

On peut même vérifier numériquement que l'erreur globale se comporte dans beaucoup de situations vraiment en $C\Delta t^p$ (*i.e.*, la borne supérieure sur l'erreur ne peut pas être améliorée). Des estimations similaires peuvent être obtenues avec des pas de temps variables.

2.1.5 Analyse de stabilité pour les systèmes linéaires dissipatifs

Nous avons vu que lorsque $\Phi_{\Delta t}$ est Lipschitzienne en y, nous disposons d'un résultat de stabilité avec $S(T) = e^{\Lambda_\Phi T}$. Lorsque le temps de simulation T est fixé à une valeur bien plus grande que Λ_Φ^{-1} (qui représente typiquement la plus petite échelle de temps présente dans le système), la constante $S(T)$ prend des valeurs trop grandes pour que le résultat de stabilité soit pertinent à des fins d'analyse numérique. Afin d'obtenir un meilleur résultat de stabilité, il est nécessaire d'introduire des hypothèses sur la dynamique du système. Nous allons ici nous restreindre à des systèmes linéaires dissipatifs autonomes, ce qui signifie que $f(t, y) = -Ay$ où A est une matrice donnée dans $\mathbb{R}^{d\times d}$ (linéarité et autonomie) supposée positive (dissipativité). On a donc

$$\langle Ay, y\rangle_{\ell^2} \geqslant 0, \qquad \forall y \in \mathbb{R}^d,$$

où $\langle \cdot, \cdot \rangle_{\ell^2}$ désigne le produit scalaire Euclidien dans $\mathbb{R}^d$, la norme Euclidienne étant notée $\|\cdot\|_{\ell^2}$ ainsi que la norme matricielle induite. Notons que l'on ne suppose pas que la matrice A est symétrique.

Pour simplifier, nous allons nous restreindre à l'étude des schémas d'Euler implicite et explicite. Dans le premier cas, le schéma (2.8) s'écrit

$$y^{n+1} = y^n - \Delta t A y^{n+1}, \tag{2.20}$$

ou encore, en posant $B_I := (\mathrm{Id} + \Delta t A)^{-1}$,

$$y^{n+1} = B_I y^n.$$

Lemme 2.4 (Matrice B_I)**.** *La matrice* $(\mathrm{Id} + \Delta t A)$ *est inversible pour tout* $\Delta t \geqslant 0$ *(si bien que la matrice* B_I *est bien définie), et on a* $\|B_I\|_{\ell^2} \leqslant 1$*.*

Démonstration. La matrice $(\mathrm{Id} + \Delta t A)$ est bien inversible car son noyau est réduit à $\{0\}$ (si $x \in \mathbb{R}^d$ est tel que $(\mathrm{Id} + \Delta t A)x = 0$, en prenant le produit scalaire avec x et en utilisant la positivité de A, il vient $\|x\|_{\ell^2}^2 = -\Delta t \langle Ax, x \rangle_{\ell^2} \leqslant 0$ d'où $x = 0$). Montrons maintenant que $\|B_I\|_{\ell^2} \leqslant 1$. Pour cela, nous devons montrer que pour tout $x \in \mathbb{R}^d$, le vecteur $y = B_I x$ est tel que $\|y\|_{\ell^2} \leqslant \|x\|_{\ell^2}$. En observant que $y - x = -\Delta t A y$ dont on prend le produit scalaire avec y, on obtient

$$\frac{1}{2}\|y\|_{\ell^2}^2 + \frac{1}{2}\|y - x\|_{\ell^2}^2 - \frac{1}{2}\|x\|_{\ell^2}^2 = \langle y - x, y \rangle_{\ell^2} = -\Delta t \langle Ay, y \rangle_{\ell^2} \leqslant 0,$$

si bien que $\|y\|_{\ell^2} \leqslant \|x\|_{\ell^2}$. □

Corollaire 2.5 (Stabilité, Euler implicite)**.** *Le schéma d'Euler implicite* (2.20) *est stable avec* $S = 1$ *pour tout* $T > 0$*.*

Démonstration. En considérant les deux suites dans (2.15), on a

$$y^{n+1} - z^{n+1} = B_I(y^n - z^n) + \Delta t\, B_I \delta^{n+1},$$

ce qui conduit par récurrence à l'expression

$$y^n - z^n = \Delta t \sum_{k=1}^{n} B_I^{n-k+1} \delta^k.$$

En utilisant l'inégalité triangulaire pour le membre de droite ainsi que le fait que $\|B_I\|_{\ell^2} \leqslant 1$, il vient

$$\|y - z\|_{\ell_t^\infty(\ell^2)} := \max_{1 \leqslant n \leqslant N} \|y^n - z^n\|_{\ell^2} \leqslant \Delta t \sum_{k=1}^{N} \left\|\delta^k\right\|_{\ell^2} =: \|\delta\|_{\ell_t^1(\ell^2)}.$$

D'où la stabilité avec $S = 1$. □

Dans le cas du schéma d'Euler explicite, le schéma (2.8) s'écrit

$$y^{n+1} = y^n - \Delta t A y^n, \tag{2.21}$$

ou encore, en posant $B_E := (\mathrm{Id} - \Delta t A)$,

$$y^{n+1} = B_E y^n.$$

Afin d'étudier la stabilité du schéma d'Euler explicite, nous introduisons une hypothèse de structure supplémentaire : il existe $\gamma > 0$ tel que pour tout $x \in \mathbb{R}^d$,

$$\langle Ax, x \rangle_{\ell^2} \geqslant \gamma \|Ax\|_{\ell^2}^2. \tag{2.22}$$

La constante γ a les dimensions d'un temps caractéristique ; lorsque la matrice A est de surcroît symétrique, on peut montrer que $\gamma = 1/\|A\|_{\ell^2}$ (voir la remarque ci-dessous).

Lemme 2.6 (Matrice B_E)**.** *On suppose* (2.22)*. Sous la condition* $\Delta t \leqslant 2\gamma$*, on a* $\|B_E\|_{\ell^2} \leqslant 1$*.*

Démonstration. Supposons $\Delta t \leqslant 2\gamma$. Nous devons montrer que pour tout $x \in \mathbb{R}^d$, le vecteur $y = B_E x$ est tel que $\|y\|_{\ell^2} \leqslant \|x\|_{\ell^2}$. En observant que $y - x = -\Delta t A x$ dont on prend le produit scalaire avec x, on obtient

$$\frac{1}{2}\|y\|_{\ell^2}^2 - \frac{1}{2}\|y-x\|_{\ell^2}^2 - \frac{1}{2}\|x\|_{\ell^2}^2 = \langle y - x, x\rangle_{\ell^2} = -\Delta t \langle Ax, x\rangle_{\ell^2},$$

si bien que

$$\begin{aligned}\frac{1}{2}\|y\|_{\ell^2}^2 - \frac{1}{2}\|x\|_{\ell^2}^2 &= \frac{1}{2}\|y-x\|_{\ell^2}^2 - \Delta t\langle Ax, x\rangle_{\ell^2} \\ &= \frac{1}{2}\Delta t^2 \|Ax\|_{\ell^2}^2 - \Delta t \langle Ax, x\rangle_{\ell^2} \\ &\leqslant \frac{\Delta t}{2\gamma}(\Delta t - 2\gamma)\langle Ax, x\rangle_{\ell^2} \leqslant 0,\end{aligned}$$

grâce à l'hypothèse faite sur Δt. □

Corollaire 2.7 (Stabilité, Euler explicite)**.** *On suppose* (2.22)*. Sous la condition* $\Delta t \leqslant 2\gamma$*, le schéma d'Euler explicite* (2.21) *est stable avec* $S = 1$ *pour tout* $T > 0$*.*

Démonstration. En considérant les deux suites dans (2.15), on a

$$y^{n+1} - z^{n+1} = B_E(y^n - z^n) + \Delta t\, \delta^{n+1},$$

ce qui conduit par récurrence à l'expression

$$y^n - z^n = \Delta t \sum_{k=1}^{n} B_E^{n-k}\delta^k.$$

En utilisant l'inégalité triangulaire pour le membre de droite ainsi que le fait que $\|B_E\|_{\ell^2} \leqslant 1$, il vient $\|y - z\|_{\ell_t^\infty(\ell^2)} \leqslant \|\delta\|_{\ell_t^1(\ell^2)}$, d'où la stabilité avec $S = 1$. □

Remarque (Cas symétrique)**.** Lorsque la matrice A est symétrique (et positive), on a, pour tout $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, $\langle Ax, y\rangle_{\ell^2} \leqslant \langle Ax, x\rangle_{\ell^2}^{1/2}\langle Ay, y\rangle_{\ell^2}^{1/2}$ (la preuve est la même que pour l'inégalité de Cauchy–Schwarz). Par suite,

$$\langle Ax, y\rangle_{\ell^2} \leqslant \langle Ax, x\rangle_{\ell^2}^{1/2}(\|Ay\|_{\ell^2}\|y\|_{\ell^2})^{1/2} \leqslant \langle Ax, x\rangle_{\ell^2}^{1/2}\|A\|_{\ell^2}^{1/2}\|y\|_{\ell^2},$$

si bien que

$$\|Ax\|_{\ell^2} = \sup_{y\in\mathbb{R}^d\setminus\{0\}} \frac{\langle Ax, y\rangle_{\ell^2}}{\|y\|_{\ell^2}} \leqslant \|A\|_{\ell^2}^{1/2}\langle Ax, x\rangle_{\ell^2}^{1/2}.$$

Cette majoration montre que l'hypothèse de structure (2.22) sur la matrice A est satisfaite avec $\gamma = 1/\|A\|_{\ell^2}$. Notons également que dans le cas symétrique, A est diagonalisable dans $\mathbb{R}$ et il est possible de raisonner directement sur les valeurs propres de A, que nous notons $0 \leqslant \lambda_1 \leqslant \ldots \leqslant \lambda_d$. Les valeurs propres de la matrice B_E sont

$$1 - \Delta t\lambda_d \leqslant \ldots \leqslant 1 - \Delta t\lambda_1 \leqslant 1,$$

si bien que sous la condition $\Delta t \leqslant 2/\lambda_d$, on a $1 - \Delta t\lambda_d \geqslant -1$ et donc $\|B_E\|_{\ell^2} \leqslant 1$. Comme $\|A\|_{\ell^2} = \lambda_d$, on retrouve bien la condition de stabilité $\Delta t \leqslant 2\gamma$. □

2.1.6 Autres éléments d'analyse (complément)

Les sous-sections suivantes présentent quelques éléments d'analyse moins standards vous permettant de vous initier à des techniques et des préoccupations d'analyse numérique plus avancées.

Description formelle de l'analyse numérique

Nous avons présenté dans la section précédente l'analyse numérique la plus standard des méthodes d'intégration des équations différentielles ordinaires. De manière un peu abstraite, on peut voir le problème à résoudre comme la solution de

$$F(x) = d,$$

où d est une donnée initiale, x la solution recherchée, et $F : X \to D$ une application entre deux espaces X et D. Formellement, $x = F^{-1}(d)$. Pour les EDOs, $D = \mathbb{R}^d$ serait l'ensemble des conditions initiales, alors que X serait l'ensemble des solutions au temps final T (auquel cas $X = \mathbb{R}^d$), ou alors la trajectoire complète du système sur le temps $[0, T]$ (auquel cas $X = C^0([0, T], \mathbb{R}^d)$). Les méthodes numériques résolvent quant à elles

$$F_\Delta(x_\Delta) = d_\Delta,$$

pour une application $F_\Delta : X_\Delta \to D_\Delta$. Le paramètre Δ caractérise la finesse de la méthode numérique – cela peut être un pas de temps, un pas d'espace, l'inverse d'un nombre de nœuds d'intégration pour le calcul d'intégrales, etc. Noter que les espaces X et D sont en général modifés par la procédure numérique. Pour les EDOs, on aurait $D_\Delta = D$, mais, dans le cas où c'est toute la trajectoire qui nous intéresse, $X_\Delta \neq X$: vu que l'on approche la trajectoire par un ensemble de valeurs discrètes $\{y^n\}_{n=0,\dots,T/\Delta t}$ qui sont des approximations de la solution exacte à des temps discrets ; on a dans ce cas $X_\Delta = (\mathbb{R}^d)^{T/\Delta t}$. Nous avons étudié précédemment l'erreur de consistance

$$\eta_\Delta = F_\Delta(\Pi_\Delta x) - F_\Delta(x_\Delta) = F_\Delta(\Pi_\Delta x) - d_\Delta,$$

où $\Pi_\Delta : X \to X_\Delta$ est un opérateur d'interpolation permettant de passer d'une solution du problème continu à sa version discrète. L'analyse d'erreur que nous avons effectuée à la section précédente était l'analyse d'erreur *a priori* directe, qui repose sur la stabilité de l'opérateur discret F_Δ sous la forme

$$\|\Pi_\Delta x - x_\Delta\|_{X_\Delta} \leqslant S_\Delta \|F_\Delta(\Pi_\Delta x) - d_\Delta\|_{D_\Delta} .$$

Lorsque la constante S_Δ peut être bornée uniformément lorsque $\Delta \to 0$ et que l'erreur de consistance $\eta_\Delta \to 0$ dans cette même limite, on peut conclure à la convergence de la méthode numérique.

Une autre méthode importante est l'analyse *a posteriori*, qui permet de donner une estimation de l'erreur commise en fonction des quantités effectivement calculées. De manière abstraite, cela demande d'introduire un opérateur d'extension $E_\Delta : X_\Delta \to X$ permettant de donner un sens à des différences entre x_Δ et x en considérant la différence $E_\Delta x_\Delta - x$. Typiquement, $\Pi_\Delta E_\Delta = \mathrm{Id}_{X_\Delta}$ alors que $E_\Delta \Pi_\Delta = \mathrm{Id}_X + \mathrm{O}(\Delta^\alpha)$ (pour une certaine puissance α). On cherche ensuite à obtenir des estimations de l'erreur en fonction du résidu $F(E_\Delta x_\Delta) - F(x) = F(E_\Delta x_\Delta) - d$, selon

$$\|E_\Delta x_\Delta - x\|_X \leqslant S \|F(E_\Delta x_\Delta) - d\|_D ,$$

qui repose sur une propriété de stabilité de l'opérateur continu F.

Influence des erreurs d'arrondi

On a négligé jusqu'à présent l'influence des erreurs d'arrondi liées à la représentation machine des nombres réels. Une remarque fondamentale est toutefois qu'on ne peut représenter qu'un sous-ensemble fini de $\mathbb{R}$ sur un ordinateur, en fonction de l'espace mémoire que l'on réserve pour représenter les nombres. On emploie en pratique une représentation en virgule flottante :

$$x = (-1)^s \cdot (0.a_1\, a_2 \dots a_t) \cdot \beta^e = (-1)^s \cdot m \cdot \beta^{e-t}$$

où t est le nombre de chiffres significatifs, e est l'exposant, et m la mantisse. Les chiffres a_i sont tels que $0 \leqslant a_i \leqslant \beta - 1$ avec $a_1 \neq 0$. On a également des bornes sur les exposants : $L \leqslant e \leqslant U$. Dans cette représentation, la précision relative est

$$\frac{\Delta x}{x} = \frac{\Delta m}{m} \leqslant \frac{1}{2}\varepsilon_{\text{machine}} = \frac{1}{2}\beta^{1-t}.$$

Les deux formats les plus standards à ce jour sont définis dans la norme IEEE 754 de 1985 :

- nombres en précision simple (*float*). On utilise pour ce faire un codage sur 32 bits : 1 pour le signe, 8 pour l'exposant dont un bit de signe, 23 pour la mantisse. Dans ce cas, le plus petit nombre qui peut être représenté est $x_{\min} \simeq 10^{-38}$, et le plus grand $x_{\max} = 10^{38}$. La précision relative est de 10^{-7} ;
- nombres en double précision (*double*), via un codage sur 64 bits : 1 pour le signe, 11 pour l'exposant dont un bit de signe, 52 pour la mantisse. Dans ce cas, $x_{\min} \simeq 10^{-308}$, $x_{\max} = 10^{308}$, et $\varepsilon_{\text{machine}} \simeq 10^{-16}$.

Les règles de calculs de l'arithmétique machine sont codifiées, mais, et c'est une remarque fondamentale, sont nécessairement approchées : en gros, étant donnés deux nombres en représentation machine, on peut définir leur somme, soustraction, produit, inverse, etc, en effectuant les opérations usuelles, puis en arrondissant le résultat obtenu au plus proche nombre représentable. Toutes les opérations arithmétiques engendent ainsi des erreurs d'arrondi de l'ordre de la précision machine $\varepsilon_{\text{machine}}$.

L'erreur d'arrondi globale est définie comme la différence entre la solution numérique idéale y^n et celle calculée en pratique, notée $\widetilde{y}^n$ dans la suite. Les écarts ont trois origines : (i) la représentation en arithmétique machine de la condition initiale $\widetilde{y}^0 = y^0 + \delta y^0$; (ii) les opérations arithmétiques nécessaires à l'évaluation à chaque pas de temps de l'incrément $\Phi_{\Delta t_n}(t_n, \widetilde{y}^n x)$; et (iii) les opérations arithmétiques liées au calcul de l'itéré suivant. Ainsi, on peut écrire

$$\widetilde{y}^{n+1} = \widetilde{y}^n + \Delta t_n \Big(\Phi_{\Delta t_n}(t_n, \widetilde{y}^n) + \rho_n \Big) + \sigma_n.$$

On suppose que $|\rho_n| \leqslant \rho$ et $|\sigma_n| \leqslant \sigma$. Typiquement, σ est de l'ordre de la précision machine $\varepsilon_{\text{machine}}$ alors que $\rho \sim \kappa \varepsilon_{\text{machine}}$ pour un certain nombre $\kappa > 0$ (appelé conditionnement de la méthode numérique : ce nombre dit comment les erreurs d'arrondi sont amplifiées par les applications $\Phi_{\Delta t_n}$). Pour une méthode stable et $N = T/\Delta t$ pas de temps (avec un pas Δt fixé), l'erreur d'arrondi globale est ainsi

$$\max_{1 \leqslant n \leqslant N} |\widetilde{y}^n - y^n| \leqslant S \left(|\delta y^0| + \sum_{n=0}^{N-1} |\sigma_n| + \Delta t \, |\rho_n| \right) = S \left(|\delta y^0| + T\rho + \frac{T\sigma}{\Delta t} \right).$$

Notons que l'erreur d'arrondi totale diverge lorsque $\Delta t \to 0$ du fait du nombre croissant d'opérations. Sa prise en compte conduit donc à limiter le nombre d'itérations réalisées, et pose donc une borne inférieure sur le pas de temps.

Avec cette information en main, on peut chercher à minimiser l'erreur numérique totale, somme de l'erreur d'approximation totale $C_T \Delta t^p$ donnée par (2.19) et de l'erreur d'arrondi totale. Pour simplifier, on va ne retenir que le terme principal de l'erreur d'arrondi, à savoir $T\sigma/\Delta t$, auquel cas on cherche à minimiser la quantité

$$C_T \Delta t^p + \frac{T\sigma}{\Delta t}.$$

Un calcul simple montre que le pas de temps optimal pour lequel l'erreur totale soit la plus petite est

$$\Delta t_{\text{opt}} = \left(\frac{T\sigma}{pC_T} \right)^{1/(p+1)}.$$

Cela correspond à un objectif de précision maximale.

Analyse *a priori* rétrograde

La philosophie de l'analyse rétrograde est la suivante : « au lieu de considérer un résultat numérique comme la solution approchée d'un problème exact, considérons-le comme la solution exacte d'un problème approché, et étudions ce problème approché ». Dans le cas présent, au lieu de chercher à estimer l'erreur entre la solution exacte $y(t_n)$ et la solution numérique y^n comme le fait l'analyse *a priori* directe, l'objectif de l'analyse *a priori* rétrograde est de considérer la solution numérique comme la solution exacte d'une EDO avec un champ de force modifié $f_{\Delta t}$ proche de f. Il s'agit donc de construire $f_{\Delta t}$ tel que la solution *exacte* de

$$\dot{z}(t) = f_{\Delta t}\big(z(t)\big), \tag{2.23}$$

soit telle que

$$y^n = z(t_n).$$

On étudie alors les propriétés de l'EDO modifiée (2.23) pour en déduire des propriétés du schéma numérique.

On va traiter le cas d'une équation autonome (f ne dépend pas explicitement du temps) intégrée numériquement avec un pas de temps constant. Comme on cherche un champ de force modifié $f_{\Delta t}$ proche de f, on écrit

$$\dot{z} = f_{\Delta t}(z) = f(z) + \Delta t F_1(z) + \Delta t^2 F_2(z) + ..., \qquad z(0) = y^0,$$

avec des fonctions F_i (pour $i \geqslant 1$) à déterminer. On définit également le flot numérique du problème (2.1), qui est l'application

$$y^{n+1} = \Psi_{\Delta t}(y^n)$$

(*i.e.*, $\Psi_{\Delta t}(y) = y + \Delta t\, \Phi_{\Delta t}(y)$). Par exemple, pour le schéma d'Euler explicite, $\Psi_{\Delta t}(y) = y + \Delta t f(y)$ et $|\Psi_{\Delta t}(y^0) - y(\Delta t)| = \mathrm{O}(\Delta t^2)$.

Pour assurer la coincidence $y^n = z(t_n)$, il suffit d'assurer la coincidence sur un pas : si $y^0 = z(0)$ alors $y^1 = z(\Delta t)$. Le problème est donc de trouver $F_1, F_2, \ldots$ tels que

$$z(\Delta t) = \Psi_{\Delta t}(y^0).$$

On utilise pour ce faire un développement en puissances de Δt de la solution exacte de la dynamique modifiée selon

$$z(\Delta t) = z(0) + \Delta t\, \dot{z}(0) + \frac{\Delta t^2}{2} \ddot{z}(0) + ...$$

On va commencer par déterminer F_1, et pour ce faire il suffit de considérer les termes d'ordre Δt et Δt^2 dans le développement ci-dessus. On peut écrire

$$\dot{z}(0) = f(z(0)) + \Delta t F_1(z(0)) + \mathrm{O}(\Delta t^2),$$

et

$$\ddot{z}(0) = \partial_z f(z(0)) \cdot f(z(0)) + \mathrm{O}(\Delta t),$$

d'où

$$z(\Delta t) = z^0 + \Delta t\, f(z^0) + \Delta t^2 \left(F_1(z^0) + \frac{1}{2} \partial_z f(z^0) f(z^0) \right) + \mathrm{O}(\Delta t^3). \tag{2.24}$$

En choisissant

$$F_1(z) = -\frac{1}{2} \partial_z f(z) f(z) \tag{2.25}$$

on voit qu'on a donc $|\Psi_{\Delta t}(y^0) - z(\Delta t)| = \mathrm{O}(\Delta t^3)$. L'erreur de consistance pour le schéma numérique $y^{n+1} = \Psi_{\Delta t}(y^n)$, vu comme une discrétisation avec pas Δt de la dynamique continue $\dot{z} = f_{\Delta t}(z)$, est ainsi

$$\widetilde{\eta}^1 = \frac{|z(\Delta t) - \Psi_{\Delta t}(y^0)|}{\Delta t} = \mathrm{O}(\Delta t^2).$$

On peut formellement [3] itérer l'argument jusqu'à un ordre arbitraire, et montrer qu'un bon choix des perturbations au champ de force conduisent à une erreur de consistance en $\mathrm{O}(\Delta t^l)$ entre la solution numérique et la solution de la dynamique approchée, pour l arbitrairement grand ; alors que l'erreur de consistance entre la solution numérique et la solution exacte

$$\eta^1 = \frac{\left| z(\Delta t) - \left(y^0 + \Delta t\, f(y^0) \right) \right|}{\Delta t}$$

est, rappelons-le, d'ordre $\mathrm{O}(\Delta t)$.

On peut ensuite étudier les propriétés de l'équation continue $\dot{z} = f_{\Delta t}(z)$ (ou au moins de $\dot{z} = f(z) + \Delta t F_1(z)$) et en déduire des propriétés du schéma numérique pour l'EDO $\dot{y} = f(y)$. Ceci est illustré dans l'Exercice 3, où on étudie la conservation de l'énergie pour des dynamiques hamiltoniennes.

3. Ce n'est qu'un calcul formel car la série définissant $f_{\Delta t}$ n'est pas convergente en général.

Contrôle du pas d'intégration et analyse *a posteriori*

Il est également possible de construire des estimateurs d'erreur *a posteriori* (utilisant la trajectoire numérique effectivement calculée), ce qui est utile pour déterminer de manière adaptative le pas de temps d'intégration. En particulier, dans ces méthodes numériques plus avancées, on ne fixe pas le nombre de pas d'intégration *a priori*, mais plutôt un niveau d'erreur total. La discussion de cette section est très proche de la discussion sur les méthodes de quadrature adaptatives pour le calcul d'intégrales (Section 2.3.4), ce qui n'est pas une surprise puisque les schémas numériques que nous avons décrits sont fondés sur la discrétisation de la formulation intégrale de la solution.

Fixons-nous donc une erreur totale ε. Rappelons que l'erreur d'approximation peut être bornée par (2.18) lorsque l'on part de $y^0 = y(0)$ et que l'on néglige les erreurs d'arrondi :

$$\max_{1 \leqslant n \leqslant N} |y^n - y(t_n)| \leqslant S \sum_{n=1}^{N} \Delta t_n \, |\eta^n| \, .$$

On a parfois une estimation de la constante de stabilité S. Même si ce n'est pas le cas, l'inégalité (2.18) suggère que le pas de temps Δt_n doit être choisi pour que l'erreur par unité de temps soit plus ou moins constante :

$$|\eta^n| \leqslant \frac{\kappa \varepsilon}{T}, \tag{2.26}$$

où $\kappa = 1/S$ quand S est connue, ou $1/S^*$ avec S^* une majoration prudente de S sinon. L'idée est alors de partir d'un pas de temps Δt_0 suffisamment petit, et de l'augmenter ou de le réduire pour que (2.26) soit vraie.

On a donc besoin pour cela d'une estimation *a posteriori* de l'erreur de troncature. Rappelons en effet que l'estimation *a priori* est compliquée car demande le calcul de dérivées de f (voir des expressions telles que (2.13)-(2.14)). Il y a deux approches générales pour ce faire : utiliser une même méthode numérique avec deux pas de temps différents (Δt_n et $\Delta t_n/2$), ou des méthodes de Runge–Kutta d'ordres différents et emboîtées (les calculs de force intermédiaires utilisés pour le schéma d'ordre le plus élevé sont également nécessaires pour le schéma d'ordre le plus petit). L'intérêt des méthodes emboîtées est qu'elles n'engendrent pas de surcoût de calcul.

Traitons un cas particulier simple pour illustrer la méthode : le schéma d'Euler explicite, pour lequel on peut obtenir une estimation de l'erreur locale sans recourir à une méthode avec un pas $\Delta t/2$. Plus précisément, l'estimation d'erreur *a posteriori* est obtenue via la différence des forces (voir (2.13)-(2.14)) :

$$\begin{aligned} f(t_{n+1}, y^{n+1}) - f(t_n, y^n) &= \Delta t_n \Big(\partial_t f(t_n, y^n) + \partial_y f(t_n, y^n) \cdot f(t_n, y^n) \Big) + \mathrm{O}(\Delta t_n^2) \\ &= 2\, \eta^{n+1} + \mathrm{O}(\Delta t_n^2). \end{aligned}$$

Ceci montre que

$$|\eta^{n+1}| \simeq \frac{\big| f(t_{n+1}, y^{n+1}) - f(t_n, y^n) \big|}{2}.$$

Si la condition (2.26) n'est pas satisfaite, on diminue le pas de temps Δt_n jusqu'à ce que ce soit le cas (par exemple, par un facteur 0.8) ; si cette condition est satisfaite avec une bonne marge de sécurité, on peut songer à augmenter un peu le pas de temps pour la prochaine itération (par exemple, en le multipliant par 1.25). En pratique, on fixe un intervalle de valeurs admissibles $[\Delta t_{\min}, \Delta t_{\max}]$ pour le pas de temps (la valeur de $\Delta t_{\min}$ étant fixée par les limitations en temps de calcul et l'accroissement des erreurs d'arrondi, voir la discussion de la Section 2.1.4), et on arrête la simulation si Δt passe sous $\Delta t_{\min}$, ce qui est le signe d'une singularité du champ de force.

2.2 Intégration des équations aux dérivées partielles

L'objectif de cette section est de présenter les idées fondamentales d'une méthode d'intégration numérique des solutions d'équations aux dérivées partielles, connue sous le nom de méthode des différences finies. Le nom de cette méthode provient du fait que l'on approche les dérivées en

temps et en espace apparaissant dans l'équation aux dérivées partielles (EDP) à intégrer par des différences discrètes de valeurs de la fonction en des points d'un maillage régulier en espace et en temps. Prévenons tout de suite le lecteur que l'on ne peut ne pas analyser ces méthodes en les considérant comme des EDOs car on souhaite obtenir des résultats de convergence uniformes dans la limite où le pas d'espace tend vers 0, limite dans laquelle la dimension de l'inconnue dans l'EDO tend vers l'infini.

Mentionnons tout de suite que l'intérêt principal des méthodes de différences finies est leur simplicité, mais que plusieurs inconvénients limitent leur usage en pratique, en particulier la nécessaire régularité du maillage en espace (qui empêche en particulier de recourir à de l'adaptivité pour intégrer plus finement autour des seuls points de singularité) et la nécessaire régularité des fonctions en jeu, que ce soit l'inconnue de l'EDP ou les autres fonctions qui apparaissent dans les équations (termes sources ou de forçage, coefficients variables). Mentionnons enfin qu'il est souvent important de bien comprendre la physique du problème pour construire de bons schémas – et réciproquement d'ailleurs : pour bien comprendre la physique d'un problème, de bonnes simulations sont souvent utiles...

2.2.1 Motivation : l'équation d'advection-diffusion

On va considérer comme problème modèle une EDP d'inconnue $u : \mathbb{R}_+ \times \Omega \to \mathbb{R}$ sous la forme

$$\partial_t u(t,x) = \operatorname{div}\Big(D(x)\nabla u(t,x)\Big) - \operatorname{div}\Big(a(x)u(t,u)\Big) + f(t,x), \tag{2.27}$$

assortie d'une condition initiale

$$u(0,x) = u_0(x), \tag{2.28}$$

et de conditions de bord appropriées. On se place par exemple sur un ouvert borné régulier $\Omega = [0,L]^d \subset \mathbb{R}^d$, avec des conditions de bord périodiques. Les arguments de u sont la variable de temps $t \geqslant 0$ et la variable d'espace $x \in \Omega$. La fonction f est un terme source, la fonction à valeurs matricielles $D(x) \in \mathbb{R}^{d\times d}$ tient compte de la possible hétérogénéité de la diffusion, et $a(x) \in \mathbb{R}^d$ est une vitesse (ces deux fonctions étant prises indépendantes du temps pour simplifier). Pour que le problème soit bien posé, on fait l'hypothèse que $D(x)$ est symétrique définie positive en tout point. Ce genre d'EDP est utilisé par exemple en physique statistique pour décrire l'évolution de la loi du processus stochastique décrivant le mouvement de particules browniennes (équation de Fokker–Planck), ou modélise l'évolution de la concentration d'un soluté dans un écoulement. Dans ce dernier cas, on peut faire l'approximation que a est à divergence nulle et D est constant, auquel cas (2.27) se simplifie en

$$\partial_t u = D\Delta u - a \cdot \nabla u + f. \tag{2.29}$$

Le terme de convection $-a \cdot \nabla u$ est lié au transport du soluté par l'écoulement, alors que le terme $D\Delta u$ rend compte de la diffusion libre du soluté du fait de l'agitation thermique. Par la suite, on étudiera principalement une version unidimensionnelle en espace du problème ci-dessus pour alléger les notations : pour un coefficient de diffusion $D \in \mathbb{R}_+$, une vitesse $a \in \mathbb{R}$, et un domaine $\Omega = [0,L]$ (avec $L > 0$) donnés,

$$\partial_t u = D\partial_x^2 u - a\partial_x u + f \text{ pour tout } (t,x) \in [0,T] \times [0,L]. \tag{2.30}$$

Les conditions de bord périodiques se traduisent dans ce cas par

$$u(t,0) = u(t,L) \text{ pour tout } t \in [0,T].$$

Noter que si on choisit $a = 0$, on se retrouve avec un problème de diffusion pure, alors que si $D = 0$, on obtient un problème d'advection pure. On peut également, sur cette équation simple, motiver le choix de la convention de signe pour la vitesse a : en effet, si $D = 0$, on voit que $u(t,x) = u_0(x-at)$ est solution de l'équation. Lorsque $a \geqslant 0$, ceci correspond bien à la propagation des valeurs de la condition initiale vers la droite car la valeur initiale en x se retrouve au temps t au point $x + at$.

Il est utile à ce stade de donner quelques propriétés mathématiques de l'équation (2.30) – propriétés que l'on cherchera à conserver (au moins de manière approchée) pour les schémas numériques associés. Ainsi, l'EDP (2.30) satisfait

(i) une préservation de la moyenne pour $f = 0$ (ou plus généralement lorsque l'intégrale de f sur le domaine Ω est nulle) : notons tout d'abord que, grâce aux conditions de bord périodiques,

$$\int_\Omega \partial_x u(t,x)\,dx = u(t,L) - u(t,0) = 0, \qquad \int_\Omega \partial_x^2 u(t,x)\,dx = \partial_x u(t,L) - \partial_x u(t,0) = 0.$$

En intégrant (2.30) sur Ω et comme $f = 0$, on trouve alors

$$\frac{d}{dt}\left(\int_\Omega u(t,x)\,dx\right) = \int_\Omega f(t,x)\,dx = 0.$$

et donc

$$\int_\Omega u(t,x)\,dx = \int_\Omega u_0(x)\,dx.$$

(ii) une dissipation en norme spatiale $L^2(\Omega)$ pour $f = 0$: en multipliant (2.30) par u, on obtient

$$\partial_t\left(\frac{1}{2}u^2\right) = D\,u\partial_x^2 u - a\partial_x\left(\frac{1}{2}u^2\right) + fu.$$

On intègre cette relation sur le domaine Ω et on utilise une intégration par parties sur le terme de diffusion et la périodicité pour obtenir

$$\frac{d}{dt}\left(\frac{1}{2}\|u(t)\|^2_{L^2(\Omega)}\right) = -D\int_\Omega |\partial_x u(x)|^2 dx + \int_\Omega f(x)u(x)\,dx.$$

Lorsque $f = 0$, on voit que la norme $L^2(\Omega)$ décroît au cours du temps.

(iii) un principe du maximum, qui implique la décroissance de la norme $L^\infty(\Omega)$ lorsque $f = 0$. De manière générale, si la condition initiale u_0 et le forçage f sont positifs, alors la solution est positive : $u(t,x) \geqslant 0$ pour tout $x \in [0,L]$ et tout $t \geqslant 0$. On en déduit en particulier que si $|u^0(x)| \leqslant M$ et $f = 0$, alors $u_0 + M \geqslant 0$ et donc, par le principe du maximum discret, que $u(t,x) + M$, qui est la solution de (2.30) associée à la condition initiale $u_0 + M$, est positive. On montre de même que $u_0 - M \leqslant 0$, d'où au final

$$\|u(t)\|_{L^\infty(\Omega)} = \sup_{x\in\Omega} |u(t,x)| \leqslant \|u_0\|_{L^\infty(\Omega)}, \qquad (f = 0).$$

Les estimations de décroissance en normes L^∞ et L^2 motiveront l'étude de la stabilité des méthodes numériques dans des normes qui en sont le pendant au niveau discret (voir Section 2.2.4).

2.2.2 Principe de la méthode des différences finies

Dans la méthode des différences finies, on approche une fonction en considérant un ensemble de points sur une grille uniforme où on estime la valeur de la solution de l'EDP que l'on considère. Pour ce faire, on commence par se donner un temps maximal d'intégration T, et on introduit un maillage régulier en espace-temps fondé sur un pas de temps Δt et un pas d'espace Δx. On suppose que Δt et Δx sont commensurables [4] avec T et L, respectivement, et on note N et J les entiers tels que $T = N\Delta t$ et $L = J\Delta x$. Les inconnues que l'on considère sont les valeurs u_j^n pour $0 \leqslant n \leqslant N$ et $0 \leqslant j \leqslant J$, qui seront des approximations de la solution exacte $u(t_n, x_j)$ sur les points $(t_n, x_j) = (n\Delta t, j\Delta x)$ de la grille déterminée par Δt et Δx. Les conditions de bord périodiques se traduisent par l'égalité

$$\forall n \geqslant 0, \qquad u_J^n = u_0^n.$$

Il faut ajouter à cela une condition initiale au temps $t = 0$, qui se traduit par une discrétisation de la condition (2.28) selon

$$\forall 1 \leqslant j \leqslant J, \qquad u_j^0 = u_0(x_j),$$

4. Quitte à modifier légèrement ces valeurs, voir une remarque similaire au début de la Section 2.1.3.

Il suffit donc de considérer les valeurs u_j^n pour $1 \leqslant n \leqslant N$ et $1 \leqslant j \leqslant J$, ce que nous ferons par la suite.

Un schéma aux différences finies est une relation de récurrence qui permet de calculer les valeurs $\{u_j^{n+1}\}_{1\leqslant j\leqslant J}$ en fonction des valeurs $\{u_l^m\}_{1\leqslant l\leqslant J,\, m\leqslant n}$ à des temps précédents. L'idée fondamentale qui sous-tend ces relations de récurrence est le remplacement des dérivées partielles par des différences discrètes qui les approchent avec une précision suffisante. Les outils techniques de base pour proposer des différences discrètes appropriées sont les développements de Taylor, dans lesquels on néglige les restes. Plusieurs possibilités apparaissent assez naturellement pour l'opérateur de dérivation en espace : une version décentrée à droite :

$$\partial_x u(t,x) \simeq \frac{u(t,x+\Delta x)-u(t,x)}{\Delta x}, \tag{2.31}$$

ou décentrée à gauche :

$$\partial_x u(t,x) \simeq \frac{u(t,x)-u(t,x-\Delta x)}{\Delta x}, \tag{2.32}$$

ou encore une version centrée :

$$\partial_x u(t,x) \simeq \frac{u(t,x+\Delta x)-u(t,x-\Delta x)}{2\Delta x}. \tag{2.33}$$

Pour une dérivée d'ordre 2 en espace, on peut songer à

$$\partial_x^2 u(t,x) \simeq \frac{u(t,x+\Delta x)-2u(t,x)+u(t,x-\Delta x)}{\Delta x^2}. \tag{2.34}$$

Pour les différences finies en temps, *i.e.*, l'approximation de $\partial_t u(t,x)$, on peut utiliser des schémas similaires à ceux que nous avons étudiés en Section 2.1 pour l'intégration des EDOs. On peut ainsi proposer la discrétisation suivante de l'EDP (2.30) si on considère un schéma d'Euler explicite pour la dérivée en temps en voyant le membre de droite de (2.30) comme le champ de force, discrétisé avec (2.33) pour la dérivée d'ordre 1 en espace et (2.34) pour la dérivée d'ordre 2 en espace :

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} = D\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2} - a\frac{u_{j+1}^n-u_{j-1}^n}{2\Delta x} + f_j^n, \tag{2.35}$$

où $f_j^n = f(t_n, x_j)$. Si on a plutôt recours à un schéma d'Euler implicite pour la discrétisation temporelle, on obtient

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} = D\frac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{\Delta x^2} - a\frac{u_{j+1}^{n+1}-u_{j-1}^{n+1}}{2\Delta x} + f_j^{n+1}. \tag{2.36}$$

Pour symétriser un peu l'évolution en temps, on peut penser à introduire une pondération (1/2,1/2) entre une évolution en temps explicite et une évolution en temps implicite, ce qui conduit à un schéma de type Crank–Nicolson :

$$\begin{aligned}\frac{u_j^{n+1}-u_j^n}{\Delta t} = {} & \frac{1}{2}\left(D\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2} - a\frac{u_{j+1}^n-u_{j-1}^n}{2\Delta x} + f_j^n\right) \\ & + \frac{1}{2}\left(D\frac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{\Delta x^2} - a\frac{u_{j+1}^{n+1}-u_{j-1}^{n+1}}{2\Delta x} + f_j^{n+1}\right).\end{aligned} \tag{2.37}$$

2.2.3 Analyse de convergence

L'analyse de la convergence des schémas suit la même ligne directrice que dans la Section 2.1.4 : on commence par montrer que le schéma numérique est consistant (l'erreur d'intégration est localement petite) et stable, ce qui amènera immédiatement la convergence. De manière plus précise, on s'intéresse à l'ordre de convergence, c'est-à-dire la détermination de bornes supérieures de l'erreur entre la solution numérique et la solution exacte dans une norme appropriée, en fonction de puissances de Δx et de Δt.

Consistance

La notion de consistance, dans sa version quantitative, donne un premier critère fondamental permettant de discriminer parmi l'infinité de possibilités qui s'ouvrent à nous quant à la discrétisation des opérateurs différentiels. Dans la définition ci-dessous, on note $\mathscr{L}$ l'opérateur différentiel en espace et en temps associé à l'EDP $\mathscr{L}u = f$. Une approximation de cet opérateur différentiel par un opérateur aux différences finies est notée $\mathscr{L}_\Delta$, où l'indice Δ rappelle les paramètres de discrétisation Δt et Δx. Cet opérateur aux différences finies prend comme argument une suite discrète $u_\Delta = \{u_j^n\}_{0\leqslant n\leqslant N, 1\leqslant j\leqslant J}$ dans $\mathbb{R}^{(N+1)\times J}$ et renvoie une suite discrète dans $\mathbb{R}^{N\times J}$ (le schéma est écrit pour n allant de 1 à N uniquement). Par exemple, pour le problème modèle (2.30), on considère $\mathscr{L} = \partial_t - D\partial_x^2 + a\partial_x$ et l'opérateur discret associé au schéma (2.35) est

$$(\mathscr{L}_\Delta u_\Delta)_j^{n+1} = \frac{u_j^{n+1} - u_j^n}{\Delta t} - D\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} + a\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x}, \tag{2.38}$$

pour tout $0 \leqslant n \leqslant N-1$ et tout $1 \leqslant j \leqslant J$. Rappelons que la condition initiale est traitée en supposant que $u_j^0 = u^0(x_j)$ pour tout $1 \leqslant j \leqslant J$. Le schéma (2.35) s'écrit sous la forme compacte

$$\mathscr{L}_\Delta u_\Delta = \widehat{\Pi}_\Delta f, \tag{2.39}$$

avec $\left(\widehat{\Pi}_\Delta f\right)_j^{n+1} = f(t_n, x_j)$. Par la suite, on écrira toujours les schémas numériques sous la forme (2.39), en modifiant de manière appropriée les définitions de $\mathscr{L}_\Delta$ et $\widehat{\Pi}_\Delta$. Pour le schéma (2.37) par exemple, on a

$$\left(\widehat{\Pi}_\Delta f\right)_j^{n+1} = \frac{1}{2}\Big(f(t_n, x_j) + f(t_{n+1}, x_j)\Big).$$

La consistance mesure l'erreur que l'on produit lorsque l'on injecte la solution exacte dans le schéma. Pour définir rigoureusement cette notion, on généralise l'action des opérateurs discrets $\mathscr{L}_\Delta$ à des fonctions en introduisant un opérateur d'interpolation qui associe à une fonction u ses valeurs aux points du maillage :

$$\Pi_\Delta u = \Big\{u(t_n, x_j)\Big\}_{0\leqslant n\leqslant N, 1\leqslant j\leqslant J}.$$

Par exemple, pour l'équation d'advection-diffusion, on peut ainsi considérer les différences discrètes fondées sur la solution exacte :

$$\begin{aligned}(\mathscr{L}_\Delta \Pi_\Delta u)_j^{n+1} = {} & \frac{u(t_{n+1}, x_j) - u(t_n, x_j)}{\Delta t} \\ & - D\frac{u(t_n, x_{j+1}) - 2u(t_n, x_j) + u(t_n, x_{j-1})}{\Delta x^2} \\ & + a\frac{u(t_n, x_{j+1}) - u(t_n, x_{j-1})}{2\Delta x}.\end{aligned} \tag{2.40}$$

On souhaite comparer les valeurs obtenues par les différences discrètes de la solution exacte aux valeurs obtenues en appliquant les opérateurs différentiels, lorsque l'on se place aux points du maillage. Il faut pour cela se doter d'une norme espace-temps discrète.

En espace, on travaille avec les normes $\|\cdot\|_{\ell_x^p}$ définies de la manière suivante : Pour tout vecteur $y = \{y_j\}_{1\leqslant j\leqslant J} \in \mathbb{R}^J$, pour tout $1 \leqslant p < +\infty$,

$$\|y\|_{\ell_x^p} := \left(\Delta x \sum_{j=1}^J |y_j|^p\right)^{1/p}, \tag{2.41}$$

et, pour $p = +\infty$,

$$\|y\|_{\ell_x^\infty} = \max_{1\leqslant j\leqslant J} |y_j|. \tag{2.42}$$

Notons que ces normes dépendent du pas d'espace Δx, explicitement de par le préfacteur dans (2.41), mais aussi implicitement de par l'index maximal de sommation $J = L/\Delta x$. En pratique, on utilise surtout $p = 2$ et $p = +\infty$. On utilise la même notation pour les normes matricielles induites dans $\mathbb{R}^{J\times J}$. En temps, on travaille avec les normes $\|\cdot\|_{\ell^1_t}$ et $\|\cdot\|_{\ell^\infty_t}$ comme pour les équations différentielles ordinaires (voir la Section 2.1.4). Ceci conduit aux normes espace-temps suivantes : pour un vecteur $y_\Delta \in \mathbb{R}^{N\times J}$ de composantes $\{y_j^n\}_{1\leqslant n\leqslant N, 1\leqslant j\leqslant J}$,

$$\|y_\Delta\|_{\ell^\infty_t(\ell^p_x)} := \max_{1\leqslant n\leqslant N} \|y^n_\Delta\|_{\ell^p_x},$$

$$\|y_\Delta\|_{\ell^1_t(\ell^p_x)} := \Delta t \sum_{n=1}^{N} \|y^n_\Delta\|_{\ell^p_x},$$

où y^n_Δ est le vecteur de $\mathbb{R}^J$ de composantes $\{y_j^n\}_{1\leqslant j\leqslant J}$. La définition précise de la consistance est alors la suivante. Noter que l'erreur de consistance est majorée uniformément en temps et en espace.

Définition 2.8 (Consistance)**.** *L'erreur de consistance associée au schéma* (2.39) *est définie par*

$$\eta_\Delta = \mathscr{L}_\Delta \Pi_\Delta u - \widehat{\Pi}_\Delta f = \Big(\mathscr{L}_\Delta \Pi_\Delta - \widehat{\Pi}_\Delta \mathscr{L}\Big) u. \tag{2.43}$$

Le schéma est dit consistant (pour un choix de norme spatiale ℓ^p_x) si

$$\lim_{\Delta t, \Delta x\to 0} \|\eta_\Delta\|_{\ell^\infty_t(\ell^\infty_x)} = 0. \tag{2.44}$$

L'approximation est consistante d'ordre r en temps et q en espace s'il existe une constante C (dépendant de u, T et L) telle que

$$\|\eta_\Delta\|_{\ell^\infty_t(\ell^\infty_x)} \leqslant C\Big(\Delta t^r + \Delta x^q\Big). \tag{2.45}$$

Remarque (Scaling de l'erreur de consistance)**.** On pourrait arbitrairement décider de multiplier les deux membres de l'égalité (2.39) définissant le schéma numérique par des puissances de $\Delta t, \Delta x$. Par convention, l'erreur de consistance est normalisée de sorte que son unité soit celle de u divisée par un temps. □

Exemple. Traitons le cas du schéma (2.38) avec $\Big(\widehat{\Pi}_\Delta f\Big)_j^{n+1} = f(t_n, x_j)$. Pour cela, on part de (2.40) et on utilise les développements de Taylor en un élement $(t, x) \in [0, T] \times \Omega$ (ici, au nœud du maillage (t_n, x_j)) :

$$\begin{cases} u(t+\Delta t, x) = u(t,x) + \Delta t\, \partial_t u(t,x) + \dfrac{\Delta t^2}{2}\partial_t^2 u(t,x) + \mathrm{O}(\Delta t^3), \\ u(t, x+\Delta x) = u(t,x) + \Delta x\, \partial_x u(t,x) + \dfrac{\Delta x^2}{2}\partial_x^2 u(t,x) + \dfrac{\Delta x^3}{6}\partial_x^3 u(t,x) + \dfrac{\Delta x^4}{24}\partial_x^4 u(t,x) + \mathrm{O}(\Delta x^5), \end{cases}$$

pour obtenir (toutes les fonctions sont évaluées en (t_n, x_j))

$$(\mathscr{L}_\Delta \Pi_\Delta u)_j^{n+1} = \partial_t u - D\partial_x^2 u + a\partial_x u + \frac{\Delta t}{2}\partial_t^2 u - \Delta x^2\left(\frac{D}{12}\partial_x^4 u - \frac{a}{6}\partial_x^3 u\right) + \mathrm{O}\Big(\Delta t^2 + \Delta x^3\Big).$$

Comme $\mathscr{L} u = \partial_t u - D\partial_x^2 u + a\partial_x u = f$, on montre ainsi que

$$\Big(\eta_\Delta\Big)_j^{n+1} = -\frac{\Delta t}{2}\partial_t^2 u + \Delta x^2\left(\frac{D}{12}\partial_x^4 u - \frac{a}{6}\partial_x^3 u\right) + \mathrm{O}\Big(\Delta t^2 + \Delta x^3\Big),$$

où les fonctions du membre de droite sont encore évaluées au point (t_n, x_j). Par régularité de la solution, il existe ainsi une constante $C_u > 0$ (indépendante de $\Delta t, \Delta x$ lorsque ces quantités sont assez petites) telle que

$$\left|\Big(\eta_\Delta\Big)_j^{n+1}\right| \leqslant C_u(\Delta t + \Delta x^2).$$

Ceci montre que la condition (2.45) est satisfaite avec $p = 1$ et $q = 2$.

L'exercice 4 propose de montrer la consistance d'autres schémas numériques pour l'équation (2.30).

Stabilité

Comme pour l'analyse de convergence des EDOs, nous avons besoin d'un critère permettant d'assurer que les erreurs locales de consistance ne s'accumulent pas trop vite dans le temps.

Définition 2.9 (Stabilité)**.** *On dit que l'opérateur discret $\mathscr{L}_\Delta$ est stable pour la norme $\|\cdot\|_{\ell_x^p}$ en espace s'il existe une constante $S(T) > 0$ telle que, pour toute suite $z_\Delta \in \mathbb{R}^{(N+1)\times J}$ avec $(z_\Delta)_j^0 = 0$ pour tout $1 \leqslant j \leqslant J$, on ait, en définissant $y_\Delta := \mathscr{L}_\Delta z_\Delta$,*

$$\|z_\Delta\|_{\ell_t^\infty(\ell_x^p)} \leqslant S(T)\,\|y_\Delta\|_{\ell_t^1(\ell_x^p)}. \tag{2.46}$$

Si l'inégalité (2.46) *n'a lieu que pour des pas de temps Δt et des pas d'espace Δx astreints à certaines inégalités, on dit que le schéma est conditionnellement stable pour la norme $\|\cdot\|_{\ell_x^p}$.*

L'étude de la stabilité d'un schéma aux différences finies est un point délicat, auquel nous consacrons la Section 2.2.4. En pratique, la première étape consiste à remarquer que comme l'opérateur discret $\mathscr{L}_\Delta$ est linéaire et correspond à une marche en temps, l'égalité $y_\Delta = \mathscr{L}_\Delta z_\Delta$ peut se récrire génériquement sous la forme suivante : pour tout $1 \leqslant n \leqslant N$, la suite $\{z_\Delta^n\}_{0\leqslant n\leqslant N}$ avec $z_\Delta^n = \{(z_\Delta)_j^n\}_{1\leqslant j\leqslant J}$ satisfait la relation de récurrence

$$z_\Delta^n = M z_\Delta^{n-1} + \Delta t\,\widetilde{M} y_\Delta^n, \tag{2.47}$$

avec des matrices $M, \widetilde{M} \in \mathbb{R}^{J\times J}$ à déterminer en fonction du schéma considéré. Le facteur Δt est motivé par le fait que dans l'expression de $\mathscr{L}_\Delta$, on a une approximation de la dérivée temporelle, et il est ainsi naturel de voir y_Δ^n comme une différence entre les vecteurs z_Δ^n et z_Δ^{n-1}, divisée par un temps. On déduit par récurrence (et linéarité) et le fait que $z_\Delta^0 = 0$ que

$$z_\Delta^n = \Delta t \sum_{k=1}^{n} M^{n-k}\widetilde{M} y_\Delta^k.$$

d'où l'on tire la condition suffisante [5] de stabilité

$$\|M\|_{\ell_x^p} \leqslant 1,$$

avec la constante $S = \|\widetilde{M}\|_{\ell_x^p}$ (indépendante de T).

Exemple. Pour l'opérateur discret associé au schéma (2.35), on a $M = \mathrm{Id} - \Delta t A$ et $\widetilde{M} = \mathrm{Id}$ avec

$$A = \frac{D}{\Delta x^2} B + \frac{a}{2\Delta x} C, \tag{2.48}$$

où les matrices $B, C \in \mathbb{R}^{J\times J}$ sont données par (noter la prise en compte des conditions de périodicité)

$$B = \begin{pmatrix} 2 & -1 & 0 & \dots & -1 \\ -1 & 2 & -1 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & -1 & 2 & -1 \\ -1 & \dots & 0 & -1 & 2 \end{pmatrix}, \qquad C = \begin{pmatrix} 0 & 1 & 0 & \dots & -1 \\ -1 & 0 & 1 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & -1 & 0 & 1 \\ 1 & \dots & 0 & -1 & 0 \end{pmatrix}. \tag{2.49}$$

Par ailleurs, pour l'opérateur discret associé au schéma (2.36), on a $M = \widetilde{M} = (\mathrm{Id} + \Delta t A)^{-1}$ (la matrice $(\mathrm{Id} + \Delta t A)$ étant bien inversible car la matrice A est dissipative, voir Lemme 2.4).

5. On observe en pratique que cette condition s'avère également nécessaire...

Convergence

Le résultat principal concernant la convergence des schémas aux différences finies est que *la stabilité et la consistance impliquent la convergence*. C'est une extension tout à fait naturelle du théorème 2.3. Bien noter que la converegnce se fait à T, L fixés.

Définition 2.10 (Convergence)**.** *On dit qu'un schéma aux différences finies est convergent dans la norme $\|\cdot\|_{\ell_x^p}$ si, pour tout L, T fixés,*

$$\lim_{\Delta x, \Delta t \to 0} \|e_\Delta\|_{\ell_t^\infty(\ell_x^p)} = \lim_{\Delta x, \Delta t \to 0} \left(\max_{1 \leqslant n \leqslant T/\Delta t} \|e_\Delta^n\|_{\ell_x^p} \right) = 0,$$

où le vecteur erreur $e_\Delta^n \in \mathbb{R}^J$ a pour composantes $(e_\Delta^n)_j = u_j^n - u(t_n, x_j)$ pour tout $1 \leqslant j \leqslant J$.

Théorème 2.11 (Lax)**.** *On suppose que la solution u de l'EDP (2.30) que l'on considère est suffisamment régulière.*[6] *Un schéma stable et consistant est convergent dans la norme de stabilité, la vitesse de convergence étant donnée par l'ordre de consistance.*

Démonstration. Notons que $\mathscr{L}_\Delta e_\Delta = \eta_\Delta$ et que $e_\Delta^0 = 0$. De par la stabilité,

$$\|e_\Delta\|_{\ell_t^\infty(\ell_x^p)} \leqslant S(T) \, \|\mathscr{L}_\Delta e_\Delta\|_{\ell_t^1(\ell_x^p)} = S(T) \, \|\eta_\Delta\|_{\ell_t^1(\ell_x^p)} \leqslant S(T) \, T \, L^{1/p} \, \|\eta_\Delta\|_{\ell_t^\infty(\ell_x^\infty)},$$

et on conclut grâce à la consistance. □

Remarque (Comparaison des normes $\|\cdot\|_{\ell_x^p}$)**.** L'étude de stabilité dépend du choix de l'exposant $1 \leqslant p \leqslant +\infty$ (et ne peut se déduire de celle pour $p = +\infty$). En revanche, la convergence en norme l_x^∞ implique la convergence en norme l_x^p. Il est donc possible que les conditions permettant d'assurer la convergence en norme l_x^∞ soient plus restrictives que les conditions de convergence en norme l_x^2. □

2.2.4 Etude de stabilité

L'étude de la consistance (aussi pénible puisse-t-elle être) ne posant en général pas de problème, le Théorème 2.11 montre que la condition importante à satisfaire pour un schéma numérique est la stabilité. Pour une norme $\|\cdot\|_{\ell_x^p}$ donnée, on obtient des critères *suffisants* de stabilité en procédant par récurrence (en pratique, ces critères s'avèrent également le plus souvent nécessaires). Nous allons nous concentrer sur l'étude de stabilité en norme $\|\cdot\|_{\ell_x^2}$; en fin de section, quelques éléments sont fournis en complément dans le cas de la norme $\|\cdot\|_{\ell_x^\infty}$.

Rappelons que le but du jeu est de récrire l'opérateur discret associé au schéma sous la forme (2.47), puis de montrer que $\|M\|_{\ell_x^2} \leqslant 1$. Nous allons détailler deux approches pour estimer cette norme matricielle, la première basée sur une approche directe dans l'espace physique et la deuxième basée sur une approche dans le domaine fréquentiel en utilisant les séries de Fourier (cette deuxième approche est connue sous le nom d'analyse de Von Neumann). De plus, afin de bien séparer les difficultés, nous allons distinguer le cas de la diffusion pure ($a = 0$ dans l'EDP (2.29)) puis celui de l'advection pure ($D = 0$ dans (2.29)).

Diffusion pure

Le point clé à retenir concernant la diffusion est que l'approche implicite pour l'opérateur discret conduit à une stabilité ℓ_x^2 inconditionnelle, alors que l'approche explicite conduit à une stabilité conditionnelle où le pas de temps est limité par le carré du pas d'espace. Les schémas implicite et explicite dans le cas de la diffusion pure s'écrivent respectivement sous la forme

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = D \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{\Delta x^2} + f_j^{n+1}, \tag{2.50}$$

6. Ceci est en fait donné par des résultats d'analyse standards sur les équations paraboliques lorsque $D > 0$, si la condition initiale et le forçage sont suffisamment réguliers.

et

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} = D\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} + f_j^n. \tag{2.51}$$

Commençons par considérer l'approche directe dans l'espace physique. Les schémas (2.50) et (2.51) conduisent respectivement aux matrices suivantes d'ordre J dans (2.47) :

$$M_I = (\mathrm{Id} + \nu_D B)^{-1}, \qquad M_E = \mathrm{Id} - \nu_D B,$$

la matrice B étant définie dans (2.49) et où on a introduit le nombre de Courant diffusif

$$\nu_D := \frac{D\Delta t}{\Delta x^2}.$$

En définissant le produit scalaire associé à la norme $\|\cdot\|_{\ell_x^2}$ par $\langle y, z\rangle_{\ell_x^2} = \Delta x \sum_{j=1}^J y_j z_j$ pour tout $y, z \in \mathbb{R}^J$, nous constatons que la matrice B est *dissipative* puisque pour tout $y \in \mathbb{R}^J$,

$$\langle By, y\rangle_{\ell_x^2} = \sum_{j=1}^J (y_j - y_{j-1})^2,$$

avec la convention $y_0 = y_J$ par périodicité. Nous pouvons donc raisonner comme dans la dernière partie de la Section 2.1.4. Nous en déduisons que le schéma d'Euler implicite (2.50) est inconditionnellement stable. Pour le schéma d'Euler explicite (2.51), commençons par montrer la condition structurelle (2.22) pour la matrice $(D/\Delta x^2)B$. Nous constatons que pour tout $y \in \mathbb{R}^J$,

$$\begin{aligned}
\|By\|_{\ell_x^2}^2 &= \Delta x \sum_{j=1}^J \{(y_{j+1} - y_j) - (y_j - y_{j-1})\}^2 \\
&\leqslant 2\Delta x \left(\sum_{j=1}^J (y_{j+1} - y_j)^2 + \sum_{j=1}^J (y_j - y_{j-1})^2 \right) \\
&= 4\Delta x \sum_{j=1}^J (y_j - y_{j-1})^2 = 4\langle By, y\rangle_{\ell_x^2}.
\end{aligned}$$

On en déduit donc que la condition (2.22) est satisfaite pour la matrice $(D/\Delta x^2)B$ avec $\gamma = \Delta x^2/(4D)$. La stabilité conditionnelle du schéma d'Euler explicite (2.51) est donc garantie sous la condition

$$\Delta t \leqslant 2\gamma = \frac{\Delta x^2}{2D}.$$

Cette condition signifie que, pour un pas de discrétisation en espace fixé, on ne peut pas prendre des pas de temps trop grands. Ce type de condition est appelée condition CFL, du nom de ses inventeurs Courant, Friedrichs et Lewy. C'est une des remarques les plus simples, mais aussi les plus importantes de l'analyse numérique. Notons qu'elle fut découverte en 1928, avant l'apparition des premiers ordinateurs et des premières simulations numériques.

Considérons maintenant l'approche fréquentielle par série de Fourier. Cette approche (une fois qu'on en a compris le principe et le mode opératoire) est plus simple à manipuler que l'approche directe car elle évite les manipulations matricielles. Le mode opératoire est très simple : on cherche une solution numérique pour $f = 0$ de la forme

$$u_j^n(k) = \hat{U}^n(k)\, \mathrm{e}^{2\mathrm{i}\pi k j \Delta x/L}, \qquad k \in \mathbb{Z}. \tag{2.52}$$

Il s'agit d'une onde plane de nombre d'onde k. En insérant cette expression dans le schéma, on obtient une relation de la forme

$$\hat{U}^{n+1}(k) = \mathscr{A}(k)\hat{U}^n(k),$$

où $\mathscr{A}(k) \in \mathbb{C}$ est appelé le coefficient d'amplification pour le nombre d'onde k. La condition de stabilité ℓ^2_x de Von Neumann s'écrit sous la forme suivante :

$$|\mathscr{A}(k)| \leqslant 1, \qquad \forall k \in \mathbb{Z}. \tag{2.53}$$

Considérons à titre d'exemple le schéma d'Euler implicite (2.50), que nous récrivons, pour $f = 0$, sous la forme

$$u_j^{n+1} + \nu_D(-u_{j+1}^{n+1} + 2u_j^{n+1} - u_{j-1}^{n+1}) = u_j^n.$$

En insérant l'expression (2.52), il vient

$$\left(1 + \nu_D(-\mathrm{e}^{2\mathrm{i}\pi k\Delta x/L} + 2 - \mathrm{e}^{-2\mathrm{i}\pi k\Delta x/L})\right) \hat{U}^{n+1}(k) = \hat{U}^n(k).$$

D'où en posant $\xi_k = \pi k \frac{\Delta x}{L}$, $\hat{U}^{n+1}(k) = \mathscr{A}(k)\hat{U}^n(k)$ avec

$$\mathscr{A}(k) = \frac{1}{1 + 4\nu_D \sin^2(\xi_k)},$$

montrant (à nouveau) la stabilité inconditionnelle du schéma d'Euler implicite pour la diffusion pure. En procédant de même pour le schéma d'Euler explicite, on trouve

$$\mathscr{A}(k) = 1 - 4\nu_D \sin^2(\xi_k).$$

La condition de stabilité (2.53) est satisfaite si $\mathscr{A}(k) \geqslant -1$, ce qui fournit (à nouveau) la condition $\nu_D \leqslant 1/2$ ou encore $\Delta t \leqslant \Delta x^2/(2D)$.

Justifions maintenant cette analyse de stabilité dans le domaine fréquentiel. Nous commençons par définir l'opérateur linéaire $\mathcal{F} : \mathbb{R}^J \to \mathbb{C}^{\mathbb{Z}}$ de la manière suivante. Soit $y \in \mathbb{R}^J$. Nous lui associons tout d'abord la fonction reconstruite $R(y) : [0, L] \to \mathbb{R}$ constante par morceaux telle que

$$R(y)(x) := y_j \quad \text{si} \quad x \in K_j = \left]\left(j - \frac{1}{2}\right)\Delta x, \left(j + \frac{1}{2}\right)\Delta x\right[, \qquad \forall 1 \leqslant j \leqslant J,$$

avec la convention de périodicité $K_J = [0, \Delta x/2[\cup]L - \Delta x/2, L]$. La fonction $R(y)$ est dans $L^2_{\mathrm{per}}(]0, L[; \mathbb{R})$. On peut ainsi la décomposer en série de Fourier selon

$$R(y)(x) = \sum_{k\in\mathbb{Z}} \widehat{y}(k)\, \mathrm{e}^{2\mathrm{i}\pi kx/L},$$

où

$$\widehat{y}(k) = \frac{1}{L}\int_0^L R(y)(x)\, \mathrm{e}^{-2\mathrm{i}\pi kx/L}\, dx \in \mathbb{C},$$

est appelé le k-ième mode de la fonction $R(y)$. Lorsque comme ici la fonction $R(y)$ est à valeurs réelles, on a la propriété de symétrie $\widehat{y}(-k) = \overline{\widehat{y}(k)}$. On pose enfin

$$\mathcal{F}(y)_k := \widehat{y}(k), \qquad \forall k \in \mathbb{Z}.$$

En équipant l'espace $\mathbb{C}^{\mathbb{Z}}$ de la norme

$$\|\widehat{z}\|_{\ell^2(\mathbb{Z})} := \left(L\sum_{k\in\mathbb{Z}} |\widehat{z}(k)|^2\right)^{1/2}$$

pour tout $\widehat{z} = \{\widehat{z}(k)\}_{k\in\mathbb{Z}} \in \mathbb{C}^{\mathbb{Z}}$, l'opérateur linéaire $\mathcal{F}$ définit une isométrie puisque, pour tout $y \in \mathbb{R}^J$,

$$\|\mathcal{F}(y)\|_{\ell^2(\mathbb{Z})} = \left(\int_0^L |R(y)(x)|^2\, dx\right)^{1/2} = \left(\Delta x \sum_{j=1}^J |y_j|^2\right)^{1/2} = \|y\|_{\ell^2_x}, \tag{2.54}$$

la première égalité résultant de la formule de Plancherel et la deuxième de la définition de la fonction $R(y)$.

Soit $y \in \mathbb{R}^J$ et $z = My$. Notre but est de montrer que $\|z\|_{\ell^2_x} \leqslant \|y\|_{\ell^2_x}$. Grâce à la propriété d'isométrie (2.54), cela revient à montrer que $\|\mathcal{F}(z)\|_{\ell^2(\mathbb{Z})} \leqslant \|\mathcal{F}(y)\|_{\ell^2(\mathbb{Z})}$ ou encore que

$$|\hat{z}(k)| \leqslant |\hat{y}(k)\|, \qquad \forall k \in \mathbb{Z}.$$

En posant $\hat{U}^n(k) = \hat{y}(k)$, le fait que $z = My$ et que $u_j^{n+1}(k)$ se déduit du schéma sans terme source à partir de $u_j^n(k)$ impliquent que $\hat{z}(k) = \hat{U}^{n+1}(k)$. Par suite,

$$\hat{z}(k) = \mathscr{A}(k)\hat{y}(k),$$

ce qui justifie la condition de stabilité (2.53).

Advection pure

Le point clé à retenir concernant l'advection est lié à la discrétisation spatiale plus qu'à la discrétisation temporelle (où nous nous contentons pour l'instant de considérer une discrétisation explicite). L'approche par différences finies centrées en espace,

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = f_j^n, \tag{2.55}$$

conduit à un schéma inconditionnellement instable, alors que l'approche par différences finies décentrées

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_j^n - u_{j-1}^n}{\Delta x} = f_j^n, \tag{2.56}$$

conduit à un schéma conditionnellement stable où le pas de temps est limité de manière linéaire par le pas d'espace. On suppose ici que $a > 0$ pour fixer les idées ; lorsque $a < 0$, le décentrement s'opère en combinant les indices j et $(j+1)$ en espace.

Commençons par considérer l'approche directe dans l'espace physique. Le schéma centré (2.55) conduit à la matrice d'ordre J donnée par

$$M = \mathrm{Id} - \frac{\nu_a}{2}C,$$

la matrice C étant définie dans (2.49) et en définissant le nombre de Courant advectif

$$\nu_a := \frac{a\Delta t}{\Delta x}.$$

Soit $y \in \mathbb{R}^J$ et $z = My$. Comme $z - y = -(\nu_a/2)Cy$ et que la matrice C est anti-symétrique, il vient en prenant le produit scalaire avec y,

$$\|z\|^2_{\ell^2_x} - \|y\|^2_{\ell^2_x} - \|z-y\|^2_{\ell^2_x} = 2\langle z-y, y\rangle_{\ell^2_x} = -\nu_a\langle Cy, y\rangle_{\ell^2_x} = 0,$$

et donc

$$\|z\|^2_{\ell^2_x} - \|y\|^2_{\ell^2_x} = \|z-y\|^2_{\ell^2_x} = \frac{\nu_a^2}{4}\|Cy\|^2_{\ell^2_x} \geqslant 0,$$

avec en général inégalité stricte (sauf dans le cas particulier où $Cy = 0$). Ceci montre que le schéma centré est inconditionnellement instable. Le schéma centré (2.56) conduit quant à lui à la matrice d'ordre J donnée par

$$M = \mathrm{Id} - \nu_a C_\sharp, \qquad C_\sharp = \begin{pmatrix} 1 & 0 & 0 & \dots & -1 \\ -1 & 1 & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & -1 & 1 & 0 \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}.$$

On vérifie aisément que la matrice $C_\sharp$ est dissipative avec, pour tout $y \in \mathbb{R}^J$,

$$\langle C_\sharp y, y\rangle_{\ell^2_x} = \Delta x \sum_{j=1}^{J} y_j^2 - y_{j-1}y_j = \frac{\Delta x}{2}\sum_{j=1}^{J}\frac{y_j^2}{2} - y_{j-1}y_j + \frac{y_{j-1}^2}{2} = \frac{\Delta x}{2}\sum_{j=1}^{J}(y_j - y_{j-1})^2 = \frac{1}{2}\langle By, y\rangle_{\ell^2_x}.$$

Nous pouvons donc chercher à satisfaire la condition structurelle (2.22) pour la matrice $C_\sharp$. Un calcul simple montre que

$$C_\sharp C_\sharp^T = C_\sharp^T C_\sharp = B,$$

et donc $\|C_\sharp y\|^2_{\ell^2_x} = \langle By, y\rangle_{\ell^2_x} = 2\langle C_\sharp y, y\rangle_{\ell^2_x}$. La condition (2.22) est ainsi satisfaite avec $\gamma = 1/2$, d'où la stabilité du schéma décentré sous la condition $\nu_a \leqslant 2\gamma = 1$, soit

$$\Delta t \leqslant \frac{\Delta x}{a}.$$

Retrouvons ces résultats par l'approche fréquentielle. Pour le schéma centré (2.55), un calcul direct montre que

$$\mathscr{A}(k) = 1 - \mathrm{i}\nu_a \sin(2\pi k \Delta x/L),$$

d'où l'instabilité inconditionnelle car $|\mathscr{A}(k)|^2 = 1 + \nu_a^2 \sin^2(2\pi k\Delta x/L) \geqslant 1$. Pour le schéma décentré (2.56), un calcul direct montre que

$$\mathscr{A}(k) = 1 - \nu_a(1 - \mathrm{e}^{-2\mathrm{i}\pi k\Delta x/L}) = 1 - \nu_a\big(1 - \cos(2\xi_k)\big) - \mathrm{i}\nu_a \sin(2\xi_k),$$

où on a introduit $\xi_k := \pi k\Delta x/L$. En notant que $1-\cos(2\xi_k) = 2\sin^2(\xi_k)$ et $\sin^2(2\xi_k) = 4\sin^2(\xi_k)(1-\sin^2(\xi_k))$, on montre facilement que

$$|\mathscr{A}(k)|^2 = 1 - 4\nu_a(1-\nu_a)\sin^2(\xi_k) \leqslant 1,$$

sous la condition $\nu_a \leqslant 1$.

Advection-diffusion

L'enseignement des sections précédentes est qu'il est bon d'utiliser un décentrage sur la partie advection, et un schéma implicite pour éviter de trop limiter le pas de temps du fait de la diffusion. On suppose à nouveau $a \geqslant 0$, et introduit le schéma

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = D\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{\Delta x^2} - a\frac{u_j^{n+1} - u_{j-1}^{n+1}}{\Delta x} + f_j^{n+1}. \tag{2.57}$$

Pour l'étude de stabilité, il faut considérer la matrice $M = (\mathrm{Id} + \nu_D B + \nu_a C_\sharp)^{-1}$. Soit $y \in \mathbb{R}^J$ et $z = My$. Comme les matrices B et $C_\sharp$ sont dissipatives, on peut raisonner comme pour le schéma d'Euler implicite pour la diffusion pure et conclure directement à la stabilité inconditionnelle en norme ℓ^2_x du schéma (2.57). On peut retrouver ce résultat avec l'analyse de Fourier. En notant toujours $\xi_k = \pi k\Delta x/L$, un calcul simple montre que facteur d'amplification est

$$\mathscr{A}(k) = \frac{1}{1 + (4\nu_D + 2\nu_a)\sin^2(\xi_k) + \mathrm{i}\nu_a \sin(2\xi_k)},$$

qui est clairement de module inférieur à 1 pour tout choix de pas de temps Δt et d'espace Δx (on a utilisé $\nu_a \geqslant 0$).

On montre également avec une analyse de Fourier que le schéma (2.35) est conditionnellement stable en norme l^2_x, alors ques les schémas implicites (2.36) et (2.37) sont inconditionnellement stables en norme l^2_x (voir les Exercices 5 et 6).

Complément : stabilité ℓ^∞_x.

La stabilité en norme ℓ^∞_x est liée à un principe du maximum discret, qui est le pendant numérique d'une propriété satisfaite au niveau continu par l'équation d'advection (2.30) (voir la discussion suivant cette équation).

Définition 2.12. *Pour un pas de temps $\Delta t > 0$ et un pas d'espace $\Delta x > 0$ fixés, on dit qu'un schéma numérique satisfait un principe du maximum discret si et seulement si la solution numérique $(u^n)_{1\leqslant n\leqslant N}$ est positive (i.e., $u^n_j \geqslant 0$ pour tout $0 \leqslant n \leqslant N$ et $1 \leqslant j \leqslant J$) lorsque la donnée initiale u^0 et le forçage f sont toutes deux des fonctions positives.*

Vu qu'avec (2.47) on peut réécrire le schéma numérique sous la forme

$$u^{n+1} = Mu^n + \Delta t\, \widetilde{f}^n, \quad \text{avec} \quad \widetilde{f}^n = N\left(\widehat{\Pi}_\Delta f\right)^n,$$

la première remarque est que, pour satisfaire le principe du maximum discret, il faut et il suffit que la matrice M apparaissant dans (2.47) ait tous ses coefficients positifs si les termes de forçage $\widetilde{f}^n_j$ du schéma numérique sont bien positifs lorsque f est positive (ce dernier point se vérifie au cas par cas vu que l'expression du terme de forçage dépend du schéma). En effet, il est clair que si M a toutes ses entrées $(M_{k,l})_{1\leqslant k,l\leqslant J}$ positives ou nulles, et si $u^n_j, \widetilde{f}^n_j \geqslant 0$ pour tout $1 \leqslant j \leqslant J$, alors on a $u^{n+1}_j \geqslant 0$ pour tout $1 \leqslant j \leqslant J$. Si on au contraire il existe (k^*, l^*) tel que $M_{(k^*,l^*)} < 0$, alors en choisissant un forçage nul ($f = 0$) et une condition initiale positive concentrée autour du point $l^*\Delta x$, mais nulle en les autres points, on voit que $u^1_j = 0$ si $j \neq k^*$ mais $u^1_{k^*} = M_{(k^*,l^*)}u^0_{l^*} < 0$, en contradiction avec le principe du maximum discret.

L'intérêt du principe du maximum discret est qu'il permet d'énoncer une condition *suffisante* de stabilité.

Proposition 2.13. *Un schéma numérique qui vérifie le principe du maximum discret est stable en norme ℓ^∞_x.*

Ceci montre que pour obtenir une condition suffisante de stabilité ℓ^∞_x d'un schéma numérique, il suffit de s'assurer que la matrice M apparaissant dans (2.47) a tous ses coefficients positifs ou nuls. Bien que la Proposition 2.13 ait été énoncée comme un principe général, il faut en fait la prouver au cas par cas, en fonction du schéma numérique.

Considérons pour ce faire le schéma d'Euler explicite, et cherchons tout d'abord des conditions suffisantes garantissant qu'un principe du maximum discret est satisfait. On a montré que pour le schéma (2.35), la matrice M dans (2.47) est $M_E = \mathrm{Id} - \Delta t\, A$ avec A donnée par (2.48). Le principe du maximum discret est donc satisfait si et seulement si tous les coefficients de la matrice $M = \mathrm{Id} + \Delta t\, A$ sont positifs, à savoir

$$\Delta t \leqslant \frac{\Delta x^2}{2D}, \qquad \frac{|a|\Delta x}{2D} \leqslant 1. \tag{2.58}$$

On retrouve ainsi la condition CFL de stabilité ℓ^2_x (voir (2.72)), ainsi qu'une autre condition traduisant le fait que l'advection ne doit pas être trop grande devant la diffusion.

Preuve de la Proposition 2.13 pour le schéma (2.35). Supposons que les conditions (2.58) soient satisfaites (*i.e.*, le principe du maximum discret est satisfait), et vérifions que le schéma (2.35) est stable en norme ℓ^∞_x. Il suffit pour ce faire de montrer que la norme matricielle de $M = \mathrm{Id} + \Delta t\, A$ est plus petite que 1. Considérons donc deux vecteurs U, V tels que $U = (\mathrm{Id} + \Delta t\, A)V$ et montrons que $\|U\|_\infty \leqslant \|V\|_\infty$. On introduit le vecteur W de composantes $W_j = U_j - \|V\|_\infty$, qui est tel que

$$W = (\mathrm{Id} + \Delta t\, A)\Big(V - \|V\|_\infty\Big),$$

où on utilise le fait que la somme des coefficients de A sur une ligne est nulle. Comme par ailleurs le vecteur $V - \|V\|_\infty$ a toutes ses composantes négatives, on en déduit par le principe du maximum

discret que W a toutes ses composantes négatives. On fait de même en introduisant cette fois $\widetilde{W} = U + \|V\|_\infty$. Au final, on obtient ainsi que $|U_j| \leqslant \|V\|_\infty$, ce qui donne bien le résultat annoncé. □

En conclusion, on a stabilité si les deux conditions (2.58) sont satisfaites, ce qui montre que le schéma d'Euler explicite (2.35) n'est que conditionnellement stable en norme ℓ_x^∞.

On peut également étudier la stabilité des schémas (2.36) ou (2.37) (voir l'Exercice 7). En fait, quitte à modifier un peu le schéma (2.36) en utilisant ce qu'on appelle un décentrage, on peut obtenir une stabilité inconditionnelle. C'est un résultat typique des schémas implicites qui, s'ils sont plus lourds à implémenter, récompensent l'usager par une stabilité accrue.

2.2.5 Généralisations (complément)

Nous évoquons rapidement dans cette dernière section quelques généralisations de l'étude de la discrétisation de l'équation d'advection-diffusion unidimensionnalle :

- la première remarque est qu'on peut bien sûr considérer d'autres conditions de bord : des conditions de Dirichlet homogènes ou non (par exemple $u_0^n = h^n$ pour un forçage $h(t)$ donné à gauche ; idem à droite) ou des conditions de Neumann homogènes ou non (auquel cas on a par exemple des conditions du type $(u_1 - u_0)/\Delta x = h^n$ à gauche ; idem a droite) ;
- on peut également considérer des schémas dits multi-niveaux, où les valeurs de u^{n+1} dépendent non seulement de u^n, mais également d'itérations précédentes u^{n-1}, etc ;
- on peut bien sûr considérer des EDPs posées en dimension $d \geqslant 2$, ce sont d'ailleurs la majorité des situations physiques intéressantes ! Dans ce cas, pour éviter des formules de Taylor compliquées et des problèmes matriciels en dimension très grande, on peut recourir à des méthodes de décomposition (aussi appelées directions alternées en français, ou *splitting* en anglais) et intégrer l'EDP direction spatiale par direction spatiale si cela s'y prête ;
- on a souvent affaire à des EDPs à coefficients non constants : dans ce cas, les schémas numériques demandent plus de soin dans leur conception ;
- un dernier cas, et ce n'est pas le moindre, est celui des EDPs nonlinéaires, telle que l'équation de Burgers qui modélise de manière réduite la dynamique des fluides incompressibles sous la forme

$$\partial_t u + \partial_x \left(\frac{u^2}{2} \right) = 0.$$

  Pour ce type d'équation, des singularités peuvent apparaître en temps fini, indépendamment de la régularité de la condition initiale. La notion de valeur ponctuelle n'a alors plus trop de sens, et il faut se tourner vers des méthodes numériques dédiées comme les volumes finis et les schémas conservatifs.

2.3 Compléments : calcul d'intégrale

Il est nécessaire, dans beaucoup de domaines d'application, de calculer numériquement des intégrales. Ce travail ne semble pas passionnant *a priori*, d'autant plus que les logiciels de calcul scientifique tels que Matlab ou Scilab font très bien tout ça tout seuls dans la majorité des cas... Oui, mais justement : il peut arriver que les solveurs par défaut n'arrivent pas à calculer l'intégrale voulue, ou alors, de manière moins dramatique, que l'on se demande quelle est la fiabilité du résultat. Il est donc important de comprendre comment est calculée l'approximation d'une intégrale. On peut également motiver la pertinence de l'intégration numérique par des applications comme le calcul de quantités moyennes en physique statistique numérique (voir Section 2.3.1).

On se limite ici aux méthodes déterministes qui permettent un calcul précis en petite dimension. Dans la majorité des cas, la brique de base est une formule de quadrature interpolaire, utilisant des points équirépartis (méthode de Newton–Cotes) ou des points de Gauss (qui vérifient certaines propriétés d'optimalité). Il est cependant conseillé de leur superposer des méthodes d'extrapolation de type Richardson/Romberg ou des méthodes automatiques (adaptatives ou non) pour assurer que les quantités calculées le sont avec une précision suffisante.

Pour les problèmes en grande dimension, les méthodes déterministes ne sont plus utilisables, et on leur préfère des méthodes stochastiques. Ces méthodes stochastiques peuvent être des méthodes directes (auquel cas il faut souvent les améliorer avec des techniques de réduction de variance), des méthodes de quasi-Monte Carlo, ou, dans les cas les plus compliqués, des méthodes fondées sur des chaînes de Markov ergodiques. Nous n'évoquerons toutefois aucune de ces approches ici.

2.3.1 Motivation : calcul de propriétés moyennes en physique statistique

Présentons pour commencer une situation physique intéressante dans laquelle il s'agit de calculer des intégrales de fonctions. A l'échelle microscopique, la matière n'est pas continue mais est composée d'atomes, qui ont des positions et des vitesses données. Un système classique [7] de N atomes est décrit par sa configuration microscopique ou microétat :

$$(q,p) = (q_1,\dots,q_N,\ p_1,\dots,p_N) \in \mathcal{D}^N \times \mathbb{R}^{dN},$$

où les positions q_i sont dans un domaine $\mathcal{D}$ (typiquement, une boite, avec des conditions de bord périodiques) et p_i est l'impulsion de la particule i (vitesse multipliée par la masse m_i). L'énergie du système dans cette configuration est donnée par le Hamiltonien

$$H(q,p) = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} + V(q_1,\dots,q_N).$$

L'expression de l'énergie cinétique est la même pour tous les systèmes : toute la physique est contenue dans l'expression de V. Rappelons également quelques ordres de grandeur. Les distances interatomiques typiques sont de quelques Å (soit 10^{-10} m), les énergies à température ambiante de l'ordre de $k_{\mathrm{B}}T \simeq 4 \times 10^{-21}$ J, et le nombre de molécules dans un échantillon macroscopique de l'ordre de $\mathcal{N}_A = 6,02 \times 10^{23}$ atomes ! On ne peut donc pas considérer tous les degrés de libertés d'un système macroscopique car ils sont bien trop nombreux ; et c'est également inutile car souvent un comportement moyen se dégage. Pour toutes ces raisons, on décrit le système par le biais d'un macroétat ou ensemble thermodynamique. Mathématiquement, c'est une mesure de probabilité sur l'ensemble des configurations accessibles.

Les propriétés moyennes ou grandeurs thermodynamiques d'équilibre sont obtenues en prenant la moyenne de fonctions des microétats par rapport à la mesure de probabilité décrivant l'état du système :

$$\langle A \rangle = \int_{\mathcal{D}^N \times \mathbb{R}^{dN}} A(q,p)\, d\mu(q,p).$$

On voit donc que cela demande de calculer une intégrale en dimension très grande, ladite intégrale ne pouvant se calculer analytiquement que dans des cas très simples (et rarement pertinents pour donner une information quantitative précise, même s'ils permettent parfois de discuter qualitativement des comportements physiques). Pour donner des ordres de grandeur des calculs effectués en pratique, disons qu'on peut simuler informatiquement de nos jours des systèmes allant de quelques centaines à plusieurs milliards d'atomes.

Pour préciser cette discussion, présentons un exemple : le calcul de la loi d'état de l'argon, *i.e.*, la courbe donnant la pression en fonction de la densité et de la température. L'observable associée à la pression est

$$A(q,p) = \frac{1}{d|\mathcal{D}|} \sum_{i=1}^{N} \left(\frac{p_i^2}{m_i} - q_i \cdot \nabla_{q_i} V(q) \right).$$

Par ailleurs, la mesure canonique décrit un système à température constante et densité fixée :

$$\mu_{\mathrm{NVT}}(dq\, dp) = \frac{1}{Z} \exp\left(-\frac{H(q,p)}{k_{\mathrm{B}}T} \right),$$

où la fonction de partition Z est une constante de normalisation qui assure que μ_{NVT} est bien une mesure de probabilité. Il ne reste plus qu'à préciser qui est le potentiel d'interaction V. C'est là

7. Au sens de non quantique.

qu'intervient une erreur de modélisation puisqu'on remplace souvent $V(q)$, qui en toute rigueur devrait être calculé par le biais de la physique quantique comme l'énergie fondamentale de l'opérateur de Schrödinger associé aux atomes placés aux positions q_i, par une formule empirique. Pour l'argon dans les conditions thermodynamiques usuelles, une bonne approximation consiste à utiliser des interactions de paires de type Lennard–Jones :

$$V(q_1, \ldots, q_N) = \sum_{1 \leqslant i < j \leqslant N} V_0(|q_j - q_i|),$$

avec

$$V_0(r) = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right],$$

et $\sigma = 3,405 \times 10^{-10}$ m, $\varepsilon/k_{\mathrm{B}} = 119,8$ K. En fait, seule la partie d'interactions à longue portée en r^{-6} a un sens physique (interactions de Van der Waals), l'objectif de la partie à courte portée étant d'empêcher que deux atomes ne puissent trop se rapprocher (ce qui modélise la répulsion des nuages électroniques associés aux atomes).

2.3.2 Principe de base des méthodes déterministes

Dans toute cette section, on se limite au cas d'intégrales à calculer en dimension 1, sur des domaines bornés $[a, b]$, et pour des fonctions régulières f. On cherche donc à approcher

$$I(f) = \int_a^b f(x)\, dx. \tag{2.59}$$

Il y a bien sûr plein d'extensions possibles pour traiter les cas plus compliqués qui ne rentrent pas dans le jeu de nos hypothèses (fonctions f à intégrer non bornées ou possédant des singularités, domaines d'intégration non bornés, etc), mais nous nous limitons volontairement au cas le plus simple pour faire ressortir le cœur des méthodes. Notons également que, quitte à faire un changement de variable affine et à remplacer la fonction à intégrer par $g(t) = f(a + (b-a)t)$, on peut se ramener à l'intégration de fonctions sur l'intervalle $[0, 1]$.

Toute méthode numérique déterministe d'approximation d'une intégrale est fondée sur une formule de quadrature utilisant des nœuds $x_i \in [a, b]$ et des poids $\omega_i \in \mathbb{R}$:

$$\int_a^b f(x)\, dx \simeq \sum_{i=0}^{n} \omega_i f(x_i). \tag{2.60}$$

Définition 2.14. *On dit qu'une méthode est d'ordre k si l'égalité* (2.60) *a lieu pour tout polynôme de degré au plus k.*

On définit également l'erreur de quadrature

$$E(f) = \int_a^b f(x)\, dx - \left(\sum_{i=0}^{n} \omega_i f(x_i) \right).$$

L'idée de base qui sous-tend le choix des poids et des nœuds dans (2.60) est de remplacer la fonction à intégrer par une fonction plus simple qui coïncide avec cette fonction en un certain nombre de points (fonction interpolante), les poids étant ensuite déterminés par l'intégration analytique de la fonction interpolante. Des exemples simples sont présentés ci-dessous.

Si la fonction f ou ses dérivées varient significativement entre a et b, il peut être opportun de remplacer une formule de quadrature globale telle que (2.60) par une formule dite composite : on découpe l'intégrale sur $[a, b]$ en M intégrales élémentaires

$$\int_a^b f(x)\, dx = \sum_{m=1}^{M} \int_{\alpha_{m-1}}^{\alpha_m} f(x)\, dx$$

avec $a = \alpha_0 < \alpha_1 < \cdots < \alpha_M = b$. Chaque intégrale élémentaire apparaissant dans la somme du membre de droite est ensuite approchée par une formule du type (2.60) (obtenue en pratique par un changement de variable affine de (2.60) donnée sur $[0, 1]$). Là aussi, des exemples simples sont présentés ci-dessous.

Quelques exemples simples

Commençons par présenter quelques exemples classiques qui permettront de fixer les idées, et seront également l'occasion de faire un peu d'analyse d'erreur *a priori*. Rappelons toutefois avant de commencer la formule de Taylor avec reste exact, que nous utiliserons de manière répétée : pour une fonction régulière $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2}f''(0) + \cdots + \frac{x^n}{n!}f^{(n)}(0) + \frac{x^{n+1}}{(n+1)!}f^{(n+1)}(\theta_x),$$

avec $\theta_x \in [0, x]$.

Exemple. Pour la *méthode du point milieu*, la fonction à intégrer est approchée par une constante, la valeur de cette constante étant la valeur de la fonction au milieu de l'intervalle. L'intégration analytique de la fonction interpolante est triviale, et donne

$$I_0(f) = (b-a)f\left(\frac{a+b}{2}\right).$$

La version composite de cette méthode est la suivante. On pose $h = (b-a)/M$ et on considère les nœuds $x_m = a + (m + 1/2)h$ avec $m = 0, \ldots, M-1$. Dans ce cas, l'intégrale de la fonction est approchée par la somme discrète

$$I_{0,M}(f) = h\sum_{m=0}^{M-1} f(x_m).$$

Essayons à présent d'obtenir une estimation de l'erreur commise en remplaçant l'intégrale par $I_0(f)$ ou $I_{0,M}(f)$. On va supposer que f est assez régulière, $C^2([a, b])$ dans le cas présent. L'outil technique de base est d'utiliser une formule de Taylor avec reste exact : pour tout $x \in [a, b]$, il existe $\theta(x) \in [a, b]$ tel que

$$f(x) = f\left(\frac{a+b}{2}\right) + f'\left(\frac{a+b}{2}\right)\left(x - \frac{a+b}{2}\right) + \frac{1}{2}f''\left(\theta(x)\right)\left(x - \frac{a+b}{2}\right)^2.$$

En intégrant cette égalité sur $[a, b]$, on obtient

$$\int_a^b f(x)\,dx = (b-a)f\left(\frac{a+b}{2}\right) + \frac{1}{2}\int_a^b f''\left(\theta(x)\right)\left(x - \frac{a+b}{2}\right)^2 dx.$$

Le second terme du membre de droite peut se réécrire comme

$$\begin{aligned}\int_a^b f''\left(\theta(x)\right)\left(x - \frac{a+b}{2}\right)^2 dx &= (b-a)^3\int_0^1 f''\Big(\theta(a + (b-a)t)\Big)\left(t - \frac{1}{2}\right)^2 dt \\ &= (b-a)^3 f''(\xi)\int_0^1 \left(t - \frac{1}{2}\right)^2 dt,\end{aligned}$$

où on a utilisé le théorème de la valeur moyenne pour l'intégration pour obtenir la dernière égalité : avec $F_2(t) = f''(\theta[a + (b-a)t])$, il existe t_ξ tel que

$$\int_0^1 F_2(t)\left(t - \frac{1}{2}\right)^2 dt = F_2(t_\xi)\int_0^1 \left(t - \frac{1}{2}\right)^2 dt.$$

Au final, en posant $H = (b-a)/2$, on peut donc dire qu'il existe $\xi \in [a,b]$ tel que l'erreur de quadrature s'écrive

$$E(f) = \int_a^b f(x)\,dx - I_0(f) = \frac{H^3}{3} f''(\xi).$$

On a donc une méthode d'ordre 1. Pour la méthode composite correspondante, on montre que l'erreur de quadrature est

$$E_M(f) = \frac{(b-a)h^2}{24} f''(\eta) \tag{2.61}$$

pour un certain $\eta \in [a,b]$. Notons que l'erreur varie en $1/M^2$, où M est le nombre de nœuds de la quadrature.

Exemple. La *méthode des trapèzes* est obtenue en interpolant la fonction à intégrer par une fonction affine prenant les mêmes valeurs aux extrémités de l'intervalle. Une intégration analytique de la fonction interpolante donne alors

$$I_1(f) = \frac{b-a}{2}\Big(f(a) + f(b)\Big).$$

La formule composite correspondante utilise les nœuds $x_m = a + mh$ pour $m = 0, \ldots, M$ et $h = (b-a)/M$:

$$I_{1,M} = h\left(\frac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{M-1}) + \frac{1}{2}f(x_M)\right). \tag{2.62}$$

On montre que les erreurs de quadrature sont respectivement

$$E_1(f) = \int_a^b f(x)\,dx - I_1(f) = -\frac{(b-a)^3}{12} f''(\xi), \qquad E_{1,M}(f) = -\frac{(b-a)h^2}{12} f''(\eta), \tag{2.63}$$

pour des nombres $\xi, \eta \in [a,b]$. On obtient donc une méthode numérique d'ordre 1 aux performances très similaires à celle de la méthode du point milieu.

Exemple. La *méthode de Cavalieri–Simpson* est une méthode d'ordre 3, qui est à la base de nombre de techniques plus raffinées. Son principe de base est d'approcher la fonction à intégrer par une parabole en utilisant les trois points d'interpolation $a, (a+b)/2, b$. L'intégration analytique de la fonction interpolante donne

$$I_2(f) = \frac{b-a}{6}\Big(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\Big).$$

On peut montrer que l'erreur de quadrature est

$$E_2(f) = \int_a^b f(x)\,dx - I_2(f) = -\frac{(b-a)^5}{90} f^{(4)}(\xi).$$

La formule composite correspondante utilise $2M+1$ nœuds $x_m = a + mh/2$ pour $m = 0, \ldots, 2M$ et $h = (b-a)/M$:

$$I_{2,M}(f) = \frac{h}{6}\left(f(x_0) + 2\sum_{r=1}^{M-1} f(x_{2r}) + 4\sum_{s=0}^{M-1} f(x_{2s+1}) + f(x_{2M})\right). \tag{2.64}$$

L'erreur de quadrature associée

$$E_{2,M}(f) = -\frac{(b-a)h^4}{2880} f^{(4)}(\eta), \qquad \eta \in [a,b], \tag{2.65}$$

est ici proportionnelle à $1/M^4$. On a intérêt à utiliser cette quadrature par rapport à la méthode du point milieu ou à la méthode des trapèzes si la fonction à intégrer est plus régulière (dérivée quatrième pas trop grande).

Points équidistants

Nous décrivons rapidement ici le principe des méthodes dites de Newton–Cotes, fondées sur l'interpolation de Lagrange en des points équidistants. On note x_j ($0 \leqslant j \leqslant n$) les nœuds de quadrature : $x_j = a + jh$ avec $h = (b-a)/n$, et P_j les polynômes d'interpolation valant 1 en x_j et 0 aux autres nœuds. La fonction interpolante approchant f est alors

$$f_n(x) = \sum_{j=0}^{n} f(x_j) P_j(x), \qquad P_j(x) = \prod_{k \neq j} \frac{x - x_k}{x_j - x_k},$$

et on obtient alors une approximation de l'intégrale à calculer en évaluant analytiquement

$$\int_a^b f_n(x)\, dx.$$

On construit ainsi des méthodes d'ordre arbitrairement élevé. Prévenons toutefois le lecteur que la montée en ordre peut toutefois être limitée en pratique par des instabilités numériques (liées à l'apparition de poids négatifs pour les formules d'ordres élevés), et ne fait sens que si la fonction est assez régulière, avec des dérivées pas trop grandes.

Points de Gauss

Les nœuds dans la méthode de Gauss sont les racines de polynômes orthogonaux d'ordre croissant. Cette classe de quadrature est importante car elle satisfait une certaine propriété d'optimalité : on peut montrer qu'elle est d'ordre $2n-1$ pour n nœuds. On peut obtenir des formules d'ordre arbitrairement élevé, qui sont numériquement stables (au sens où les poids qui interviennent dans la quadrature sont toujours positifs). En revanche, l'expression analytique des nœuds et des poids est de plus en plus compliquée et coûteuse à évaluer. Un exemple important est la méthode de Gauss d'ordre 3, obtenue par une quadrature à 2 nœuds :

$$I(f) = \int_{-1}^{1} f(x)\, dx \simeq I_{\mathrm{G}}(f) = f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

On peut bien sûr transformer cette quadrature élémentaire en une quadrature sur un intervalle général $[a, b]$ par une transformation affine :

$$I(f) = \int_a^b f(x)\, dx \simeq I_{\mathrm{G}}(f) = \frac{b-a}{2}\left[f\left(\frac{a+b}{2} - \frac{1}{\sqrt{3}}\frac{b-a}{2}\right) + f\left(\frac{a+b}{2} + \frac{1}{\sqrt{3}}\frac{b-a}{2}\right)\right].$$

2.3.3 Extrapolation

Pour gagner en précision, plutôt que de travailler avec des méthodes d'ordre élevé, qui demandent beaucoup de nœuds et sont relativement lourdes à mettre en œuvre, une idée attractive est de partir d'une méthode plus simple et d'améliorer ses résultats de manière systématique par un procédé itératif. L'idée de base sous-tendant cette approche est l'extrapolation de Richardson. Son application à la quadrature numérique donne la méthode de Romberg.

Extrapolation de Richardson

Supposons que l'on sache que la fonction A admette le développement suivant pour t petit :

$$A(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \cdots + \alpha_k t^k + \mathrm{O}(t^{k+1}) \tag{2.66}$$

et qu'on souhaite calculer α_0. L'idée est d'extrapoler la valeur de la fonction A en 0 connaissant ses valeurs pour des arguments $t > 0$. On voit déjà qu'on peut combiner $A(t)$ et $A(\delta t)$ (avec $0 < \delta < 1$) pour éliminer le terme linéaire et avoir une approximation à l'ordre 2 de α_0 :

$$\frac{A(\delta t) - \delta A(t)}{1 - \delta} = \alpha_0 - \alpha_2 \delta t^2 + \ldots$$

On peut ensuite combiner deux approximations d'ordre 2 pour en obtenir une à l'ordre 3, etc. Cela revient à définir successivement les fonctions

$$A_0(t) = A(t),$$
$$A_1(t) = \frac{A_0(\delta t) - \delta A_0(t)}{1-\delta},$$
$$A_2(t) = \frac{A_1(\delta t) - \delta^2 A_1(t)}{1-\delta^2},$$

ce qui donne $A_2(t) = \alpha_0 + \alpha_3\delta^3(1+\delta)t^3 + \dots$ et, de manière générale,

$$A_n(t) = \frac{A_{n-1}(\delta t) - \delta^n A_{n-1}(t)}{1-\delta^n}.$$

On élimine ainsi successivement les termes en t, t^2,..., t^n. Cette procédure est illustrée sur la Figure 2.1, où l'on voit que les fonctions successives $A_1, A_2, ...$ sont de plus en plus "plates".

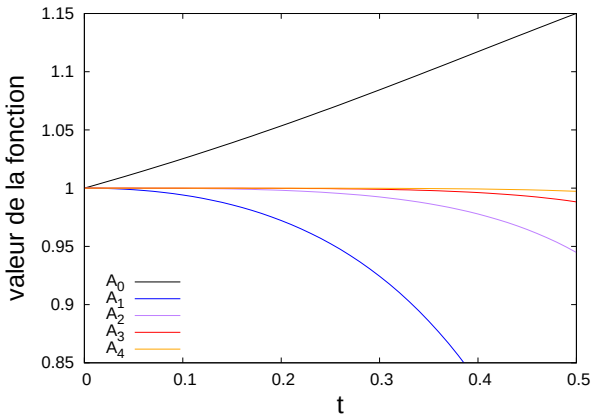


FIGURE 2.1 – Illustration de l'extrapolation de Richardson pour $t = 0.5$ et $\delta = 0.5$, partant d'une fonction A qui est un polynôme de degré 10 avec des coefficients aléatoirement choisis.

On peut réécrire cette procédure de façon à faire apparaitre plus explicitement les quantités que l'on doit effectivement calculer, à savoir les valeurs de la fonction de base $A_0 = A$. Pour ce faire, on définit tout d'abord, pour une valeur t fixée,

$$\mathcal{A}_{m,0} = A(\delta^m t), \qquad m = 0, \dots, n,$$

puis, pour des valeurs croissantes de q, on combine ces valeurs selon

$$\mathcal{A}_{m,q+1} = \frac{\mathcal{A}_{m,q} - \delta^{q+1}\mathcal{A}_{m-1,q}}{1-\delta^{q+1}} = \frac{\delta^{-(q+1)}\mathcal{A}_{m,q} - \mathcal{A}_{m-1,q}}{\delta^{-(q+1)} - 1}, \qquad q = 0, \dots, n-1.$$

Une illustration graphique permet de mieux comprendre ce qui se passe. On calcule les termes de la première colonne, et on en déduit les termes diagonaux, qui sont ceux qui nous intéressent :

$$\begin{array}{lccccccccc}
\mathcal{A}_{0,0} = A(t) & & & & & & & & & \\
 & \searrow & & & & & & & & \\
\mathcal{A}_{1,0} = A(\delta t) & \to & \mathcal{A}_{1,1} & & & & & & & \\
 & \searrow & & \searrow & & & & & & \\
\mathcal{A}_{2,0} = A(\delta^2 t) & \to & \mathcal{A}_{2,1} & \to & \mathcal{A}_{2,2} & & & & & \\
 & \searrow & & \searrow & & \searrow & & & & \\
\mathcal{A}_{3,0} = A(\delta^3 t) & \to & \mathcal{A}_{3,1} & \to & \mathcal{A}_{3,2} & \to & \mathcal{A}_{3,3} & & & \\
\vdots & \ddots & & \ddots & & \ddots & & \ddots & & \\
 & \searrow & & \searrow & & \searrow & & \searrow & & \\
\mathcal{A}_{n,0} = A(\delta^n t) & \to & \mathcal{A}_{n,1} & \to & \mathcal{A}_{n,2} & \to & \mathcal{A}_{n,3} & \dots & \to & \mathcal{A}_{n,n}
\end{array}$$

Une récurrence simple (voir Exercice 8) montre que

$$\mathcal{A}_{m,q} = A_q(\delta^{m-q}t). \tag{2.67}$$

Dans la représentation graphique ci-dessus, l'indice m correspond aux lignes, et l'indice q aux colonnes.

Considérons à présent la vitesse de convergence de cette méthode. Comme $A(t) = \alpha_0 + \mathrm{O}(t)$, on a seulement $\mathcal{A}_{m,0} = \alpha_0 + \mathrm{O}(\delta^m t)$. En revanche, on peut vérifier (voir Exercice 8) que

$$\mathcal{A}_{m,q} = \alpha_0 + \mathrm{O}\left(\delta^{(q+1)(m-q/2)}t^{q+1}\right). \tag{2.68}$$

En particulier, $\mathcal{A}_{n,n} = \alpha_0 + \mathrm{O}\left(\delta^{n(n+1)/2}t^{n+1}\right)$. La convergence de $\mathcal{A}_{n,n}$ est donc $(n+1)/2$ fois plus rapide que celle de $\mathcal{A}_{n,0} = \alpha_0 + \mathrm{O}(\delta^n t)$ (lorsque l'on ne considère que la vitesse de convergence par rapport à δ et que l'on ne s'intéresse pas au facteur t).

Intégration de Romberg

Appliquons à présent l'extrapolation de Richardson aux formules de quadrature dans le cas particulier de la formule composite des trapèzes : c'est l'intégration de Romberg. On doit d'abord établir un développement du type (2.66). Par le développement d'Euler–Maclaurin, on montre que la formule des trapèzes composite (2.62) peut s'écrire, avec $h = (b-a)/M$,

$$\begin{aligned} I_{1,M} &= h\left(\frac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{M-1}) + \frac{1}{2}f(x_M)\right) \\ &= \int_a^b f(x)\,dx + \sum_{i=1}^{k} \frac{B_{2i}}{(2i)!}\Big(f^{(2i-1)}(b) - f^{(2i-1)}(a)\Big)h^{2i} + \mathrm{O}(h^{2(k+1)}), \end{aligned}$$

où les B_n sont les nombres de Bernoulli. Insistons sur le fait que les coefficients qui apparaissent dans le développement ci-dessus ne sont pas importants par eux-mêmes, c'est la forme analytique du développement qui nous importe. Ainsi,

$$I_{1,M} = T(h) = \alpha_0 + \alpha_1 h^2 + \ldots$$

est une série en puissances de h^2. On retrouve donc une expression similaire à (2.66) en remplaçant t par h^2. Lorsque l'on divise le pas du maillage par 2, on doit donc utiliser un procédé d'extrapolation avec $\delta = 1/4$ dans les notations de la section précédente. L'algorithme est alors le suivant : on commence par évaluer le résultat donné par des formules composites avec un pas divisé par 2 à chaque étape, *i.e.*,

$$\mathcal{A}_{m,0} = T\left(\frac{h}{2^m}\right), \qquad m = 0, \ldots, n.$$

On combine ensuite ces approximations selon

$$\mathcal{A}_{m,q+1} = \frac{4^{q+1}\mathcal{A}_{m,q} - \mathcal{A}_{m-1,q}}{4^{q+1} - 1}, \qquad q = 0, \ldots, n-1.$$

Le terme $\mathcal{A}_{m,q}$ approche l'intégrale (2.59) avec un erreur d'ordre $\mathrm{O}\left(2^{q(q+1)}\left(\frac{h}{2^m}\right)^{2(q+1)}\right)$ (remplacer t par h^2 et prendre $\delta = 1/4$ dans (2.67)). En particulier, $\mathcal{A}_{n,n}$ approche l'intégrale avec une erreur d'ordre $\mathrm{O}\left(\left(\frac{h}{2^n}\right)^{n+1}\right)$.

Interprétons enfin cette convergence dans le cas où $h = b - a = 1$: lorsque l'on ne fait pas d'extrapolation, et que l'on calcule $\mathcal{A}_{n,0}$, on utilise un pas d'espace

$$h_n = \frac{1}{2^n},$$

et on a ainsi une erreur de quadrature en h_n^2 au vu de l'erreur sur la formule des trapèzes composite (voir (2.63)). Lorsque l'on fait l'extrapolation, l'erreur est énormément réduite, et est plus précisément d'ordre h_n^{n+1}.

2.3.4 Méthodes automatiques

Une autre manière d'améliorer efficacement des méthodes d'ordre bas est d'utiliser des méthodes automatiques d'intégration assurant que l'erreur de quadrature est plus petite qu'un seuil donné. Ces techniques sont fondées sur des estimateurs d'erreur *a posteriori*. On peut même utiliser si on le souhaite une méthode adaptative, qui détermine de manière automatique comment répartir les nœuds, plutôt que de raffiner uniformément la grille.

Méthode automatique non-adaptative

Commençons par présenter une méthode qui détermine automatiquement le nombre de points M à utiliser dans une formule composite, en procédant par un raffinement uniforme de la grille (avec des points régulièrement espacés). On considère par exemple la formule composite de Cavalieri–Simpson (2.64), dont on rappelle l'estimation d'erreur *a priori* (2.65) :

$$E_M = I - I_M = -\frac{b-a}{2880}\left(\frac{b-a}{M}\right)^4 f^{(4)}(\eta_M).$$

La notation η_M insiste sur le fait que le point η_M pour lequel l'égalité ci-dessus est valide dépend *a priori* de M. Si on suppose toutefois que $f^{(4)}(\eta_M) \simeq f^{(4)}(\eta_{2M})$ (ce qui est effectivement le cas lorsque $M \to +\infty$), on a alors $E_{2M} \simeq E_M/16$. Un calcul simple montre que

$$E_{2M} \simeq \frac{I_{2M} - I_M}{15}. \tag{2.69}$$

Insistons sur le fait que cette estimation *a posteriori* de l'erreur est obtenue en combinant une estimation *a priori* et deux évaluations de I pour des paramètres différents. En pratique, on va donc commencer par une valeur de M_0 donnée, puis doubler cette valeur jusqu'à ce que l'erreur de quadrature estimée par (2.69) soit plus petite que la tolérance que l'on s'est initialement fixée. Au vu du caractère approché de l'estimation *a posteriori* (2.69), il est plus prudent de considérer par exemple un critère d'arrêt tel que $E_{2M} \simeq (I_{2M} - I_M)/10$.

Méthodes adaptatives (complément)

Nous concluons ici notre rapide panorama des méthodes déterministes d'intégration par une méthode de quadrature adaptative, dont des idées sont utilisées dans la définition des algorithmes standards d'intégration des logiciels comme Matlab. Etant donnée une tolérance ε préalablement fixée, l'objectif est d'obtenir une formule de quadrature avec une distribution non-uniforme de nœuds (aussi près les uns des autres qu'il le faut dans certaines régions, mais toujours en prenant soin de limiter au maximum le nombre de ces points), sans fixer le nombre de points au préalable.

Pour ce faire, on souhaite que l'erreur de quadrature estimée sur tout sous-intervalle $[\alpha, \beta]$ soit en $\varepsilon(\beta - \alpha)/(b - a)$. On part initialement de $\alpha = a$ et $\beta = b$. Pour α, β fixés, on calcule l'intégrale sur l'intervalle $[\alpha, \beta]$ avec une formule de Cavalieri–Simpson simple

$$S_f(\alpha, \beta) = \frac{\beta - \alpha}{6}\left(f(\alpha) + 4f\left(\frac{\alpha+\beta}{2}\right) + f(\beta)\right),$$

et on compare le résultat à ce que donne une formule de Cavalieri–Simpson composite avec $M = 2$ sur le même intervalle :

$$S_{f,2}(\alpha, \beta) = S_f\left(\alpha, \frac{\alpha+\beta}{2}\right) + S_f\left(\frac{\alpha+\beta}{2}, \beta\right).$$

On montre comme pour (2.69) que

$$\left|\int_\alpha^\beta f - S_{f,2}(\alpha, \beta)\right| \simeq \frac{1}{15}\left|\mathcal{E}_f(\alpha, \beta)\right|, \qquad \mathcal{E}_f(\alpha, \beta) = S_f(\alpha, \beta) - S_{f,2}(\alpha, \beta).$$

On va donc utiliser le critère d'arrêt suivant afin d'avoir une erreur de quadrature d'ordre $\varepsilon(\beta-\alpha)/(b-a)$ sur l'intervalle $[\alpha,\beta]$ (le facteur 10 ci-dessous remplaçant le 15 de la ligne précédente, pour plus de prudence) :

$$|\mathcal{E}_f(\alpha,\beta)| \leqslant 10\,\varepsilon\,\frac{\beta-\alpha}{b-a}. \tag{2.70}$$

On distingue ensuite deux cas :

- Si (2.70) est vrai, on réalise effectivement la quadrature de Cavalieri–Simpson avec les nœuds $\alpha, (\alpha+\beta)/2, \beta$, et on ajoute la contribution correspondante $S_{f,2}(\alpha,\beta)$ à l'estimation courante de l'intégrale sur l'intervalle $[a,\alpha]$. On obtient ce faisant une estimate courante de l'intégrale sur l'intervalle $[a,\beta]$. On continue ensuite en considérant l'intervalle $[\beta,b]$ sur lequel il reste à estimer l'intégrale de la fonction.
- Sinon, on raffine l'intégration sur l'intervalle $[\alpha,\beta]$ en gardant la borne gauche α et en remplaçant la borne droite par $(\alpha+\beta)/2$, et on reprend la procédure d'estimation avec le critère d'arrêt (2.70).

Il est sain dans toute cette procédure de vérifier que la taille des intervalles sur lesquels on réalise la quadrature ne dégénère pas, *i.e.*, $\beta-\alpha \geqslant h_{\min} > 0$, où $h_{\min}$ est une longueur minimale fixée par le numéricien. Un message d'erreur avertissant l'utilisateur qu'un des intervalles est trop petit est en général un signe que la fonction à intégrer a des singularités.

2.4 Exercices

Exercice 1 (Etude d'un schéma de Runge–Kutta)**.** On considère des réels $0 \leqslant \alpha,\beta,\gamma \leqslant 1$ et le schéma

$$y^{n+1} = y^n + \Delta t_n \Psi_{\alpha,\beta,\gamma}(t_n, y^n; \Delta t_n),$$

avec

$$\Psi_{\alpha,\beta,\gamma}(t,y;\Delta t) = \alpha f(t,y) + \beta f\left(t+\frac{\Delta t}{2}, y+\frac{\Delta t}{2}f(t,y)\right) + \gamma f\big(t+\Delta t, y+\Delta t f(t,y)\big).$$

Pour simplifier, on se placera dans le cas où la dimension spatiale est 1 (*i.e.*, $y \in \mathbb{R}$).

(1) Pour quelles valeurs de (α,β,γ) retrouve-t-on des schémas présentés dans le cours ?

(2) Discuter la consistance de la méthode en fonction des paramètres : quelles sont les relations à satisfaire pour avoir un ordre 1 ou 2 ? Peut-on avoir une méthode d'ordre 3 ou plus ?

(3) Etudier enfin la convergence lorsque f est uniformémement Lipschitzienne.

Exercice 2 (Analyse rétrograde d'une équation explosive)**.** On considère l'EDO $\dot y = f(y)$, partant d'une condition initiale $y^0 > 0$; ainsi qu'une dynamique continue modifiée $z(t)$ partant de la même condition initiale :

$$\dot z(t) = f(z(t)) + \Delta t\, g(z(t)), \qquad z(0) = y^0.$$

L'objectif est de comparer, à des temps multiples de Δt, les solutions *exactes* de la dynamique modifiée $z(n\Delta t)$ aux valeurs y^n, qui sont des approximations *numériques* de $y(n\Delta t)$. C'est ce qu'on appelle l'analyse rétrograde. Cette analyse permet d'obtenir des renseignements qualitatifs sur les solutions numériques. On notera par la suite $y^{n+1} = \Psi_{\Delta t}(y^n)$ l'application déterminant une itération du schéma numérique.

(1) Calculer la solution analytique $y(t)$ lorsque $f(y) = y^2$ et montrer que cette EDO explose en un temps fini noté $T(y^0)$.

(2) Ecrire, dans le cas général, le développement de $z(\Delta t)$ en puissances de Δt (avec un reste d'ordre 3) en fonction de f, g et leurs dérivées évaluées en $z(0) = y^0$.

(3) On considère à présent le schéma d'Euler explicite, pour lequel $\Psi_{\Delta t}(y) = y + \Delta t\, f(y)$.

   (a) Quel est l'entier p tel que $y(\Delta t) - \Psi_{\Delta t}(y^0) = \mathrm{O}(\Delta t^p)$?

   (b) Déterminer g pour que $z(\Delta t) - \Psi_{\Delta t}(y^0) = \mathrm{O}(\Delta t^{p+1})$.

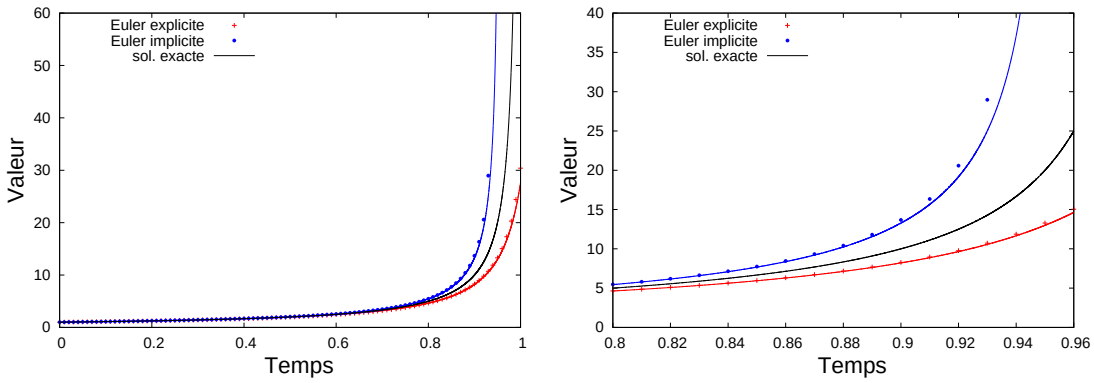


FIGURE 2.2 – Dynamique $\dot{y} = y^2$, intégrée avec Euler explicite ou implicite, et les dynamiques modifiées correspondantes (au premier ordre en Δt) ; les symboles indiquent le résultat des deux schémas numériques.

(c) Montrer que, dans le cas particulier $f(y) = y^2$, on a $z(t) \leqslant y(t)$.
Comme $z(n\Delta t)$ est une bonne approximation de y^n, ceci motive le fait que le temps d'explosion prédit par le schéma d'Euler explicite $T_{\text{EE}}\left(y^0\right)$ sur-estime $T\left(y^0\right)$ (quoiqu'il faut être prudent dans ce genre d'interprétation car on n'a considéré que le premier ordre dans la perturbation).

(4) Reprendre les questions précédentes pour le schéma d'Euler implicite, en montrant que le temps d'explosion est cette fois sous-estimé. On commencera par établir un développement en puissances de Δt de $\Psi_{\Delta t}$.

Exercice 3 (Analyse rétrograde pour les systèmes Hamiltoniens)**.** On considère l'approximation de la dynamique Hamiltonienne (2.3), tout d'abord par des schémas numériques généraux, puis par des schémas numériques spécifiques appelés schémas d'Euler symplectiques

$$\begin{cases} p^{n+1} = p^n - \Delta t\, \nabla V(q^n), \\ q^{n+1} = q^n + \Delta t\, p^{n+1}, \end{cases} \qquad \begin{cases} q^{n+1} = q^n + \Delta t\, p^n, \\ p^{n+1} = p^n - \Delta t\, \nabla V(q^{n+1}). \end{cases}$$

L'objectif de cet exercice est de montrer que ces schémas préservent bien l'énergie $H(q^n, p^n)$ du système. On se place en dimension 1 pour simplifier (*i.e.*, $(q, p) \in \mathbb{R}$), et on prend une masse égale à l'unité.

(1) Déterminer la modification à l'ordre 1 du champ de force (notée F_1 dans (2.24)) pour le schéma d'Euler explicite et le schéma d'Euler implicite. Montrer que le champ de force $f + \Delta t\, F_1$ n'est pas de la forme des champs de force apparaissant dans les dynamiques Hamiltoniennes générales (2.4).

(2) Montrer que le champ de force modifié $f + \Delta t\, F_1$ est au contraire de type Hamiltonien pour les schémas d'Euler symplectiques. En déduire qu'il existe une énergie modifiée préservée à l'ordre 2 en Δt alors que l'énergie exacte n'est préservée qu'à l'ordre 1.

Exercice 4 (Consistance de schémas aux différences finies pour l'équation d'advection-diffusion)**.** Etudier la consistance des schémas numériques (2.36) et (2.37).

Exercice 5 (Stabilité conditionnelle en norme ℓ^2_x d'un schéma explicite)**.** Montrer que le schéma d'Euler explicite (2.35) est stable en norme ℓ^2_x si

$$\forall \xi \in \mathbb{R}, \qquad \left| 1 - 4\nu_D \sin^2\left(\frac{\xi}{2}\right) - \mathrm{i}\nu_a \sin\left(\xi\right) \right| \leqslant 1. \tag{2.71}$$

Montrer que cette condition est satisfaite si et seulement si

$$\nu_a^2 \leqslant 2\nu_D \leqslant 1. \tag{2.72}$$

Pour ce faire, on pourra utiliser l'identité trigonométrique

$$\sin^2(\xi) = 4\sin^2\left(\frac{\xi}{2}\right)\left[1-\sin^2\left(\frac{\xi}{2}\right)\right]$$

pour se ramener à l'étude du signe d'un trinôme en $X(\xi) = \sin^2(\xi/2)$.

Exercice 6 (Stabilité inconditionnelle en norme ℓ^2_x de schémas implicites)**.** Montrer que les schémas implicites (2.36) et (2.37) sont inconditionnellement stables en norme ℓ^2_x.

Exercice 7 (Stabilité en norme ℓ^∞_x de schémas implicites)**.** On s'intéresse dans cet exercice à la stabilité en norme ℓ^∞_x de schémas implicites pour l'équation d'advection-diffusion.

(1) La première étape dans l'étude de la stabilité ℓ^∞_x de schémas implicites est d'obtenir des conditions suffisantes sous lesquelles ces schémas satisfont un principe du maximum discret. Cela demande de trouver des conditions suffisantes sur des matrices telles que $\mathrm{Id} - \Delta t\, A$ pour que leurs inverses aient tous leurs coefficients positifs.
Montrer que si une matrice $B \in \mathbb{R}^{J\times J}$ est telle que

$$B_{i,i} > 0, \qquad B_{i,j} \leqslant 0 \text{ si } i \neq j, \qquad \forall i \in \{1,\dots,J\}, \quad \sum_{j=1}^{J} B_{i,j} > 0,$$

alors B est inversible et B^{-1} est une matrice à coefficients positifs.

(2) A l'aide de ce résultat, montrer que le schéma implicite (2.36), qui peut se réécrire, lorsque $f = 0$, sous la forme $(\mathrm{Id} + \Delta t\, A)u^{n+1} = u^n$ (avec A à préciser), est stable en norme ℓ^∞_x sous la seule condition

$$\frac{|a|\Delta x}{2D} \leqslant 1.$$

(3) Montrer ensuite que le schéma implicite (2.37), qui peut se réécrire lorsque $f = 0$ sous la forme $(\mathrm{Id} + \Delta t\, A_1/2)u^{n+1} = (\mathrm{Id} + \Delta t\, A_2/2)u^n$ (avec A_1, A_2 à préciser), est stable sous la condition

$$\frac{|a|\Delta x}{2D} \leqslant 1, \qquad \Delta t \leqslant \frac{\Delta x^2}{D}.$$

(4) Le schéma (2.36) semble plus intéressant car il impose moins de contraintes sur le pas de temps. Pour s'affranchir de la condition de stabilité restante, qui limite encore le pas de temps, on peut recourir à ce qu'on appelle un décentrage, qui consiste à modifier la discrétisation du terme d'advection $a\partial_x u$ en tenant compte du signe de la vitesse. Supposons par exemple $a > 0$. On modifie (2.36) selon

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = D\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{\Delta x^2} - a\frac{u_{j+1}^{n+1} - u_j^{n+1}}{\Delta x} + f_j^{n+1}.$$

Réécrire ce schéma implicite sous la forme $(\mathrm{Id}+\Delta t\, A)u^{n+1} = u^n$ (avec une matrice A à préciser) lorsque $f = 0$, et vérifier qu'il est inconditionnellement stable.

(5) Quel schéma propose-t-on si $a \leqslant 0$?

Exercice 8 (Extrapolation de Richardson)**.** L'objectif de cet exercice est de vérifier la formule (2.68).

(1) Vérifier la formule (2.67) en procédant par récurrence.

(2) Pour montrer (2.68), on va tout d'abord obtenir une expression de $A_n(t) = \alpha_0 + \alpha_{n+1}^{(n)} t^{n+1} + \alpha_{n+2}^{(n)} t^{n+2} + \dots$.

(a) Montrer que, pour $m \geqslant n+1$, on a la relation de récurrence

$$\alpha_m^{(n)} = \frac{\delta^{m-n} - 1}{1 - \delta^n}\delta^n \alpha_m^{(n-1)}.$$

(b) En déduire que
$$\left|\alpha_m^{(n)}\right| \leqslant |\alpha_m| \prod_{k=1}^{n} \frac{\delta^k}{1-\delta^k}.$$

(c) Montrer qu'il existe une constante $K_\delta > 1$ telle que $\displaystyle\prod_{k=1}^{n}(1-\delta^k) \geqslant \exp\left(-\frac{K_\delta \delta}{1-\delta}\right)$ uniformément en $n \geqslant 1$. Indication : on utilisera le fait que, par concavité de $x \mapsto \ln(1-x)$, il existe $K_\delta > 1$ tel que, pour tout $x \in [0,\delta]$, on ait $\ln(1-x) \geqslant -K_\delta x$.

(d) Montrer qu'il existe une constante $C_\delta > 0$ telle que $\left|\alpha_m^{(n)}\right| \leqslant C_\delta\, |\alpha_m|\, \delta^{n(n+1)/2}$.

(e) Conclure.

Corrigés

Exercice 1. *(Etude d'un schéma de Runge–Kutta)*

(1) On retrouve le schéma d'Euler explicite pour $(\alpha,\beta,\gamma) = (1,0,0)$ et la méthode de Heun pour $(\alpha,\beta,\gamma) = (1/2,0,1/2)$.

(2) Pour étudier la consistance de la méthode, on écrit, pour un paramètre $h > 0$ (qui sera Δt ou $dt2$ par la suite) et un vecteur z (qui sera $f(t,y)$ par la suite),
$$f(t+h,y+hz) = f(t,y) + h\Big(\partial_t f(t,y) + \partial_y f \cdot f(t,y)\Big) + \frac{h^2}{2}\begin{pmatrix}1\\ z\end{pmatrix}^T \begin{pmatrix}\partial_t^2 f & \partial_t\partial_y f\\ \partial_t\partial_y f & \partial_y f\end{pmatrix}\begin{pmatrix}1\\ z\end{pmatrix} + \mathrm{O}(h^3).$$
Ainsi,
$$\begin{aligned}\Psi(t,y) = (\alpha+\beta+\gamma)f(t,y) &+ \Delta t\left(\frac{\beta}{2}+\gamma\right)\Big(\partial_t f + \partial_y f\cdot f\Big)(t,y)\\ &+ \frac{\Delta t^2}{2}\left(\frac{\beta}{4}+\gamma\right)\Big(\partial_t^2 f + 2f\partial_t\partial_y f + f^2\partial_y^2 f\Big)(t,y) + \mathrm{O}(\Delta t^3).\end{aligned}$$
Or, la solution exacte partant d'une configuration $y(t)$ au temps t satisfait
$$y(t) + \Delta t f(t,y(t)) + \frac{\Delta t^2}{2}\Big(\partial_t f(t,y(t)) + \partial_y f(t,y(t)) f(t,y(t))\Big) + \frac{\Delta t^3}{6} y^{(3)}(t) + \mathrm{O}(\Delta t^4).$$
On a donc une méthode consistante d'ordre 1 lorsque $\alpha+\beta+\gamma = 1$, et consistante d'ordre 2 pour $\beta + 2\gamma = 2$. Lorsque l'on calcule l'expression de $y^{(3)}(t)$ en dérivant par rapport au temps l'expression de $y^{(2)}(t) = \partial_t f(t,y(t)) + \partial_y f(t,y(t)) f(t,y(t))$, on constate qu'entre autres des termes de la forme $f(\partial_y f)^2$ apparaissent. Ces termes-là ne peuvent être obtenus avec Ψ, et donc la méthode ne peut pas être d'ordre 3 ou plus.

(3) Il suffit de montrer que le schéma est stable. Pour ce faire, il suffit de montrer que Ψ est uniformément Lipschitzienne. Or, pour $h > 0$ donné (Δt ou $\Delta t/2$ par la suite)
$$\begin{aligned}|f(t+h, y_1 + hf(t,y_1)) - f(t+h, y_2 + hf(t,y_2))| &\leqslant L\left|y_1 + hf(t,y_1) - \big(y_2 + hf(t,y_2)\big)\right|\\ &\leqslant L\Big(|y_1 - y_2| + h|f(y_1) - f(y_2)|\Big)\\ &\leqslant L(1+hL)\,|y_1 - y_2|.\end{aligned}$$
Au final,
$$|\Psi(t,y_1,\Delta t) - \Psi(t,y_2,\Delta t)| \leqslant L\left(\alpha+\beta+\gamma+\left(\frac{\beta}{2}+\gamma\right)L\Delta t\right)|y_1 - y_2|,$$
ce qui permet de conclure à la stabilité.

Exercice 2. *(Analyse rétrograde d'une équation explosive)*

(1) On note que

$$\frac{d}{dt}\left(\frac{1}{y}\right) = -\frac{\dot{y}}{y^2} = -1,$$

et donc

$$\frac{1}{y(t)} = \frac{1}{y^0} - t = \frac{1 - ty^0}{y^0}.$$

Ceci donne

$$y(t) = \frac{y^0}{1 - ty^0}.$$

Le temps d'explosion de la solution analytique est ainsi $T(y^0) = 1/y^0$.

(2) On procède comme dans le poly, en commençant par écrire que

$$z(\Delta t) = z(0) + \Delta t\, \dot{z}(0) + \frac{\Delta t^2}{2}\ddot{z}(0) + \mathrm{O}(\Delta t^3).$$

On remplace ensuite $\dot{z}(0)$ par $f(z(0)) + \Delta t\, g(z(0))$ et on utilise une dérivation composée pour avoir

$$\ddot{z}(0) = f'(z(0)) \cdot f(z(0)) + \mathrm{O}(\Delta t).$$

Au final,

$$z(\Delta t) = y^0 + \Delta t\, f(y^0) + \Delta t^2 \left(g(y^0) + \frac{1}{2}f'(y^0)f(y^0)\right) + \mathrm{O}(\Delta t^3).$$

(3) Pour le schéma d'Euler explicite, on a $\Psi_{\Delta t}(y) = y + \Delta t\, f(y)$.

(a) En reprenant l'expression de la question précédente avec $g = 0$, on voit que

$$y(t) = y^0 + \Delta t\, f(y^0) + \frac{\Delta t^2}{2}f'(y^0)f(y^0) + \mathrm{O}(\Delta t^3).$$

On en déduit que $p = 2$.

(b) Si on choisit $g = -\frac{1}{2}f'f$, on a en effet $z(\Delta t) - \Psi_{\Delta t}(y^0) = \mathrm{O}(\Delta t^3)$.

(c) Lorsque $f(y) = y^2$, on a $g(y) = -y^3$ et donc z satisfait

$$\dot{z} = z^2 - \Delta t\, z^3 \leqslant z^2$$

si $z(t) \geqslant 0$. On en déduit (en procédant comme au point (1), mais en manipulant des inégalités) que $z(t) \leqslant y(t)$. Si $z(t) \leqslant 0$, on a immédiatement $z(t) \leqslant 0 \leqslant y(t)$.

(4) Pour le schéma d'Euler explicite, l'application $\Psi_{\Delta t}$ est définie de manière implicite par la relation $\Psi_{\Delta t}(y) = y + \Delta t\, f(\Psi_{\Delta t}(y))$. Ainsi,

$$\begin{aligned}\Psi_{\Delta t}(y) &= y + \Delta t\, f\Big(y + \Delta t f(\Psi_{\Delta t}(y))\Big) = y + \Delta t\, f(y) + \Delta t^2 f'(y)f(\Psi_{\Delta t}(y)) + \mathrm{O}(\Delta t^3)\\ &= y + \Delta t\, f(y) + \Delta t^2 f'(y)f(y) + \mathrm{O}(\Delta t^3).\end{aligned}$$

On voit donc que $y(\Delta t) - \Psi_{\Delta t}(y^0) = \mathrm{O}(\Delta t^2)$. En revanche, si on choisit $g = \frac{1}{2}f'f$ (l'opposé de la correction pour le schéma d'Euler explicite), on a $z(\Delta t) - \Psi_{\Delta t}(y^0) = \mathrm{O}(\Delta t^3)$. Pour le cas particulier $f(y) = y^2$, on trouve $g(y) = y^3$. La dynamique modifiée est alors

$$\dot{z} = z^2 + \Delta t\, z^3 \geqslant z^2,$$

car, partant de $y^0 \geqslant 0$, on a toujours $z(t) \geqslant 0$. On en déduit que $z(t) \geqslant y(t)$, ce qui motive le fait que y^n devrait être plus grand que $z(n\Delta t)$.

Exercice 3. *(Analyse rétrograde pour les systèmes Hamiltoniens)*

(1) Le champ de force est $f(z) = (p, -V'(q))$. La formule (2.25) donne

$$F_1(z) = -\frac{1}{2}\partial_z f \cdot f(z) = -\frac{1}{2}\begin{pmatrix} 0 & 1 \\ -V''(q) & 0 \end{pmatrix}\begin{pmatrix} p \\ -V'(q) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} V'(q) \\ pV''(q) \end{pmatrix}.$$

On ne peut pas réécrire ce terme comme dérivant d'un Hamiltonien selon

$$F_1(q,p) = \begin{pmatrix} \partial_p H_1(q,p) \\ -\partial_q H_1(q,p) \end{pmatrix}.$$

Si cela était possible, on aurait

$$F_1(q,p) = \begin{pmatrix} F_1^1(q,p) \\ F_1^2(q,p) \end{pmatrix}, \quad \partial_q F_1^1 + \partial_p F_1^2 = 0,$$

ce qui n'est pas le cas. Ceci montre que le système modifié n'est pas de type hamiltonien.
Pour le schéma d'Euler implicite, on considère l'application discrète $\Psi(y^0) = y^1 = y^0 + \Delta t\, f(y^1)$. Le développement en puissances de Δt de ce schéma est

$$\Psi(y^0) = y^0 + \Delta t\, f\Big(y^0 + \Delta t\, f(y^1)\Big) = y^0 + \Delta t\, f(y^0) + \Delta t^2 \partial_y f \cdot f(y^0) + \mathrm{O}(\Delta t^3).$$

On compare cette expression à la solution (2.24) de la dynamique modifiée, ce qui donne

$$F_1(z) = \frac{1}{2}\partial_z f \cdot f(z),$$

qui est l'opposée de la correction obtenue pour le schéma d'Euler explicite. Cette correction n'est donc pas non plus un champ de force dérivant d'un Hamiltonien.

(2) On peut effectuer le même genre de calculs pour les schémas d'Euler symplectiques. Considérons par exemple le second schéma. On a dans ce cas

$$\begin{cases} q^{n+1} = q^n + \Delta t\, p^n, \\ p^{n+1} = p^n - \Delta t\, V'(q^n) - \Delta t^2\, V''(q^n) p^n + \mathrm{O}(\Delta t^3). \end{cases}$$

Or, on a vu que la solution de la dynamique modifiée partant de $y^n = (q^n, p^n)$ est

$$\begin{pmatrix} q^{n+1} \\ p^{n+1} \end{pmatrix} = \begin{pmatrix} q^n \\ p^n \end{pmatrix} + \Delta t \begin{pmatrix} p^n \\ -V'(q^n) \end{pmatrix} + \Delta t^2 \left(F_1(y^n) + \frac{1}{2}\partial_y f(y^n) \cdot f(y^n) \right) + \mathrm{O}(\Delta t^3).$$

La comparaison avec le schéma d'Euler symplectique montre que le choix suivant permet d'assurer que le flot numérique du schéma d'Euler symplectique coïncide avec le flot de la dynamique modifiée à l'ordre 2 :

$$F_1(q,p) = \frac{1}{2}\begin{pmatrix} V'(q) \\ -pV''(q) \end{pmatrix} = \begin{pmatrix} \partial_p H_1(q,p) \\ -\partial_q H_1(q,p) \end{pmatrix},$$

avec

$$H_1(q,p) = \frac{1}{2} p^T V'(q).$$

Dans ce cas, la dynamique modifiée est encore hamiltonienne, et est associée au Hamiltonien

$$H_{\Delta t}(q,p) = H(q,p) + \Delta t H_1(q,p).$$

On vient ainsi de montrer que $H_{\Delta t}(q^1,p^1) = H_{\Delta t}(q^0,p^0) + \mathrm{O}(\Delta t^2)$, alors que $H(q^1,p^1) = H(q^0,p^0) + \mathrm{O}(\Delta t)$. On conserve donc mieux une énergie approchée.

Exercice 4. *(Consistance de schémas aux différences finies pour l'équation d'advection-diffusion)* On commence par le schéma (2.36), pour lequel

$$(\mathscr{L}_\Delta u_\Delta)_j^{n+1} = \frac{u_j^{n+1}-u_j^n}{\Delta t} - D\frac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{\Delta x^2} + a\frac{u_{j+1}^{n+1}-u_{j-1}^{n+1}}{2\Delta x},$$

et $\Big(\widehat{\Pi}_\Delta f\Big)_j^{n+1} = f(t_{n+1},x_j)$. On commence par noter que

$$\frac{u(t+\Delta t,x+\Delta x)-u(t+\Delta t,x-\Delta x)}{2\Delta x} = (\partial_x u)(t+\Delta t,x)+\mathrm{O}(\Delta x^2),$$

ainsi que

$$\frac{u(t+\Delta t,x+\Delta x)-2u(t+\Delta t,x)+u(t+\Delta t,x-\Delta x)}{\Delta x^2} = (\partial_x^2 u)(t+\Delta t,x)+\mathrm{O}(\Delta x^2).$$

Par ailleurs,

$$\frac{u(t+\Delta t,x)-u(t,x)}{\Delta t} = (\partial_t u)(t+\Delta t,x)+\mathrm{O}(\Delta t).$$

Ceci montre que

$$\Big(\mathscr{L}_\Delta \Pi_\Delta u\Big)_j^{n+1} = \Big((\Pi_\Delta \mathscr{L})u\Big)_j^{n+1} + \mathrm{O}\Big(\Delta t+\Delta x^2\Big).$$

et donc

$$(\eta_\Delta)_j^{n+1} = \mathrm{O}\Big(\Delta t+\Delta x^2\Big).$$

Le schéma (2.36) est donc consistant d'ordre 1 en temps et 2 en espace.

Passons à présent au schéma (2.37). En utilisant les calculs précédents, on voit que

$$\begin{aligned}
&\frac{1}{2}\left(\frac{u(t,x+\Delta x)-u(t,x-\Delta x)}{2\Delta x}+\frac{u(t+\Delta t,x+\Delta x)-u(t+\Delta t,x-\Delta x)}{2\Delta x}\right)\\
&\quad=\frac{1}{2}\Big((\partial_x u)(t,x)+(\partial_x u)(t+\Delta t,x)\Big)+\mathrm{O}(\Delta x^2)\\
&\quad=(\partial_x u)(t,x)+\frac{\Delta t}{2}\partial_t\partial_x u(t,x)+\mathrm{O}(\Delta t^2+\Delta x^2).
\end{aligned}$$

On montre de même que

$$\begin{aligned}
&\frac{1}{2}\left(\frac{u(t,x+\Delta x)-2u(t,x)+u(t,x-\Delta x)}{\Delta x^2}+\frac{u(t+\Delta t,x+\Delta x)-2u(t+\Delta t,x)+u(t+\Delta t,x-\Delta x)}{\Delta x^2}\right)\\
&\quad=(\partial_x^2 u)(t,x)+\frac{\Delta t}{2}(\partial_t\partial_x^2 u)(t,x)+\mathrm{O}(\Delta t^2+\Delta x^2).
\end{aligned}$$

Enfin,

$$\frac{u(t+\Delta t,x)-u(t,x)}{\Delta t} = (\partial_t u)(t,x)+\frac{\Delta t}{2}(\partial_t^2 u)(t,x)+\mathrm{O}(\Delta t^2).$$

Au final, on a ainsi

$$\Big(\mathscr{L}_\Delta \Pi_\Delta u\Big)_j^{n+1} = (\mathscr{L}u)(t_n,x_j)+\frac{\Delta t}{2}(\partial_t\mathscr{L}u)(t_n,x_j)+\mathrm{O}(\Delta t^2+\Delta x^2).$$

Par ailleurs,

$$\begin{aligned}
\Big(\widehat{\Pi}_\Delta f\Big)_j^{n+1} &= \frac{1}{2}\Big(f(t_n,x_j)+f(t_{n+1},x_j)\Big)\\
&= f(t_n,x_j)+\frac{\Delta t}{2}\partial_t f(t_n,x_j)+\mathrm{O}(\Delta t^2).
\end{aligned}$$

Lorsque u est solution de $\mathscr{L}u=f$, on a ainsi $\eta_\Delta = \mathrm{O}(\Delta t^2+\Delta x^2)$. Le schéma est d'ordre 2 en espace et en temps.

Exercice 5. *(Stabilité conditionnelle en norme l^2 d'un schéma explicite)* Un calcul direct montre que

$$\begin{aligned}\widehat{u}^{n+1}(k) &= \left(1+\frac{D\Delta t}{\Delta x^2}\left(\mathrm{e}^{2\mathrm{i}\pi k\Delta x/L}-2+\mathrm{e}^{-2\mathrm{i}\pi k\Delta x/L}\right)+\frac{a\Delta t}{2\Delta x}\left(\mathrm{e}^{2\mathrm{i}\pi k\Delta x/L}-\mathrm{e}^{-2\mathrm{i}\pi k\Delta x/L}\right)\right)\widehat{u}^n(k)\\ &= \left(1-\frac{4D\Delta t}{\Delta x^2}\sin^2\left(\frac{\mathrm{i}\pi k\Delta x}{L}\right)-\frac{\mathrm{i}a\Delta t}{\Delta x}\sin\left(\frac{2\mathrm{i}\pi k\Delta x}{L}\right)\right)\widehat{u}^n(k). \qquad (2.73)\end{aligned}$$

On en déduit donc la condition (2.71) de l'énoncé. Notons à présent

$$z(\xi)=1-4\nu_D\sin^2\left(\frac{\xi}{2}\right)-\mathrm{i}\nu_a\sin(\xi).$$

La condition (2.71) est satisfaite si et seulement si $|z(\xi)|^2\leqslant 1$ pour tout $\xi\in\mathbb{R}$. Or, avec les indications de l'énoncé, on trouve

$$|z(\xi)|^2=(1-4\nu_DX(\xi))^2+4\nu_a^2X(\xi)(1-X(\xi))=1+4(\nu_a^2-2\nu_D)X(\xi)+4(4\nu_D^2-\nu_a^2)X(\xi)^2.$$

On définit alors $P(\xi)=|z(\xi)|^2-1$. On souhaite assurer que $P(\xi)\leqslant 0$. Comme $P(0)=0$, il faut déjà que $P'(0)\leqslant 0$, ce qui donne $\nu_a^2\leqslant 2\nu_D$. Une fois que cette condition est satisfaite, il suffit que $P(1)\leqslant 0$, soit $16\nu_D^2-8\nu_D\leqslant 0$, ou encore $2\nu_D\leqslant 1$. Cette seconde condition est la condition de stabilité du schéma explicite pour la diffusion pure.

Exercice 6. *(Stabilité inconditionnelle en norme ℓ_x^2 de schémas implicites)* Pour le schéma (2.36), on obtient, par des calculs similaires à ceux de la question précédente,

$$\widehat{u}^{n+1}(k)=\widehat{u}^n(k)+\left[\frac{D\Delta t}{\Delta x^2}\left(\mathrm{e}^{2\mathrm{i}\pi k\Delta x/L}-2+\mathrm{e}^{-2\mathrm{i}\pi k\Delta x/L}\right)+\frac{a\Delta t}{2\Delta x}\left(\mathrm{e}^{2\mathrm{i}\pi k\Delta x/L}-\mathrm{e}^{-2\mathrm{i}\pi k\Delta x/L}\right)\right]\widehat{u}^{n+1}(k)$$

soit $\widehat{u}^{n+1}(k)=z(\xi_k)\widehat{u}^n(k)$ avec $\xi_k=\pi k\Delta x/L$ et

$$z(\xi)=\frac{1}{1+\dfrac{4D\Delta t}{\Delta x^2}\sin^2(\xi)-\dfrac{\mathrm{i}a\Delta t}{\Delta x}\sin(\xi)}.$$

On voit donc immédiatement que, pour tout $\xi\in\mathbb{R}$,

$$|z(\xi)|^2=\frac{1}{\left[1+\dfrac{4D\Delta t}{\Delta x^2}\sin^2(\xi)\right]^2+\left(\dfrac{a\Delta t}{\Delta x}\right)^2\sin^2(\xi)}\leqslant 1.$$

Ceci établit la stabilité inconditionnelle en norme ℓ_x^2 du schéma.

Pour le schéma (2.37), on trouve

$$z_1(\xi_k)\widehat{u}^{n+1}(k)=z_2(\xi_k)\widehat{u}^n(k),$$

où z_1 est l'expression obtenue pour le schéma implicite, avec Δt remplacé par $\Delta t/2$:

$$z_1(\xi)=1+\frac{2D\Delta t}{\Delta x^2}\sin^2(\xi)-\frac{\mathrm{i}a\Delta t}{2\Delta x}\sin(\xi),$$

alors que l'expression de z_2 est la même que celle du schéma explicite, avec également Δt remplacé par $\Delta t/2$:

$$z_2(\xi)=1-\frac{2D\Delta t}{\Delta x^2}\sin^2\left(\frac{\xi}{2}\right)+\frac{\mathrm{i}a\Delta t}{2\Delta x}\sin(\xi).$$

On peut donc écrire $\widehat{u}^{n+1}(k)=z(\xi_k)\widehat{u}^n(k)$ avec $z(\xi)=z_2(\xi)/z_1(\xi)$. Ces deux nombres complexes sont tels que $|\mathrm{Im}(z_1(\xi))|=|\mathrm{Im}(z_2(\xi))|$, alors que

$$\mathrm{Re}(z_1(\xi))^2=\left(1+\frac{2D\Delta t}{\Delta x^2}\sin^2\left(\frac{\xi}{2}\right)\right)^2\geqslant\left(1-\frac{2D\Delta t}{\Delta x^2}\sin^2\left(\frac{\xi}{2}\right)\right)^2=\mathrm{Re}(z_2(\xi))^2.$$

On en déduit donc que $|z(\xi)|\leqslant 1$ pour tout $\xi\in\mathbb{R}$, d'où la stabilité inconditionnelle en norme ℓ_x^2.

Exercice 7. *(Stabilité en norme ℓ_x^∞ de schémas implicites)*

(1) On considère $x \in \mathbb{R}^J$ tel que $Bx = 0$. Alors, pour tout $i \in \{1, \dots, J\}$, on a

$$B_{ii}x_i = -\sum_{j \neq i} B_{ij}x_j.$$

On choisit l'indice i_0 tel que $|x_{i_0}| = \max_{j=1,\dots,J} |x_j|$. Alors,

$$B_{i_0 i_0}|x_{i_0}| = |B_{i_0 i_0}x_{i_0}| = \left|\sum_{j \neq i_0} B_{i_0 j}x_j\right| \leqslant \left(-\sum_{j \neq i_0} B_{i_0 j}\right)|x_{i_0}|,$$

ce qui permet de conclure que $|x_{i_0}| = 0$ au vu de la dernière condition sur les coefficients de B. On en déduit que B est injective donc inversible.
Pour voir que B^{-1} a ses coefficients positifs, on considère x tel que $Bx = y$ avec $y_i \geqslant 0$ pour tout $i \in \{1, \dots, J\}$, et on montre que $x_i \geqslant 0$ pour tout $i = 1, \dots, J$. Supposons qu'il existe i_0 tel que $x_{i_0} = \min_{j=1,\dots,J} x_j \leqslant 0$. Alors,

$$0 \leqslant y_{i_0} = \sum_{j=1}^{J} B_{i_0 j}x_j \leqslant B_{i_0 i_0}x_{i_0} + \sum_{j \neq i_0} B_{i_0 j}x_j \leqslant \left(\sum_{j=1}^{J} B_{i_0 j}\right) x_{i_0}, \tag{2.74}$$

car, en utilisant le fait que $B_{i_0 j} \leqslant 0$ si $j \neq i_0$,

$$\min_{j=1,\dots,J} x_j \leqslant \frac{\displaystyle\sum_{j \neq i_0} B_{i_0 j}x_j}{\displaystyle\sum_{j \neq i_0} B_{i_0 j}} \leqslant \max_{j=1,\dots,J} x_j,$$

et donc (vu que le dénominateur est négatif)

$$\sum_{j \neq i_0} B_{i_0 j}x_j \leqslant \left(\sum_{j \neq i_0} B_{i_0 j}\right) \min_{j=1,\dots,J} x_j.$$

L'inégalité (2.74) montre que $x_{i_0} = 0$. On peut ainsi conclure que $x_i \geqslant 0$ pour tout $i = 1, \dots, J$.

(2) On voit facilement que $A = (D/\Delta x^2)B + (a/(2\Delta x))C$ où les matrices B, C sont définies en (2.49). Notons $M = \mathrm{Id} + \Delta t\, A$. On voit facilement que $M_{ii} > 0$. Pour avoir $M_{ij} \leqslant 0$ pour $i \neq j$, il faut que les termes sous-diagonaux de M soient négatifs, ce qui est le cas lorsque

$$\frac{|a|}{2\Delta x} \leqslant \frac{D}{\Delta x^2}.$$

Par ailleurs, la somme des coefficients sur une ligne soit strictement positive, ce qui permet de conclure que l'inverse a des coefficients positifs. On retrouve donc la condition de l'énoncé.

(3) On voit que $A_1 = -A_2 = (D/\Delta x^2)B + (a/(2\Delta x))C$. La matrice $M = (\mathrm{Id} + \Delta t\, A_1/2)^{-1}(\mathrm{Id} + \Delta t\, A_2/2)$ est à coefficients positifs lorsque les deux matrices $(\mathrm{Id} + \Delta t\, A_1/2)^{-1}$ et $\mathrm{Id} + \Delta t\, A_2/2$ sont à coefficients positifs. Les conditions de stabilité sont celles du schéma d'Euler explicite pour un pas de temps $\Delta t/2$ et celle du schéma d'Euler implicite pour $\Delta t/2$, ce qui donne bien les deux conditions de l'énoncé.

(4) On voit que $A = (D/\Delta x^2)B + (a/\Delta x)C_\sharp^T$. Les coefficients hors diagonaux de A sont bien négatifs, et les coefficients diagonaux de A sont positifs. Par ailleurs la somme des coefficients de A sur une ligne est égale à 0. On en déduit que $\mathrm{I} - \Delta t\, A$ est inversible et que son inverse a des coefficients positifs. Si $a \leqslant 0$, on décentre dans l'autre sens, par exemple selon

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = D\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{\Delta x^2} + a\frac{u_j^{n+1} - u_{j-1}^{n+1}}{\Delta x}.$$

Exercice 8. *(Extrapolation de Richardson)*

(1) On suppose que $\mathcal{A}_{m-1,q} = A_q(\delta^{m-1-q}t)$ pour $0 \leqslant q \leqslant m-1$ (propriété immédiatement vérifiée pour $m = 1$), et on souhaite montrer que $\mathcal{A}_{m,q} = A_q(\delta^{m-q}t)$ pour $0 \leqslant q \leqslant m$. Pour ce faire, on procède encore de manière itérative, sur q cette fois. On a déjà, par définition, que $\mathcal{A}_{m,0} = A_0(\delta^m t)$. Supposons ensuite que $\mathcal{A}_{m,q} = A_q(\delta^{m-q}t)$ pour $0 \leqslant q \leqslant k$, avec $0 \leqslant k \leqslant m-1$, et considérons $\mathcal{A}_{m,k+1}$:

$$\mathcal{A}_{m,k+1} = \frac{\mathcal{A}_{m,k} - \delta^{k+1}\mathcal{A}_{m-1,k}}{1-\delta^{q+1}} = \frac{A_k(\delta^{m-k}t) - \delta^{k+1}A_k(\delta^{m-1-k}t)}{1-\delta^{k+1}} = A_{k+1}(\delta^{m-k-1}t)$$

par définition de A_{k+1}. Ceci permet donc de conclure la récurrence.

(2) Passons maintenant aux estimations sur les coefficients de A_n.

(a) La relation de récurrence donne, en identifiant les coefficients associés à t^m pour $m \geqslant n+1$:

$$\alpha_m^{(n)} = \frac{\delta^m - \delta^n}{1-\delta^n}\alpha_m^{(n-1)},$$

ce qui est bien la relation recherchée. On en déduit que

$$\alpha_m^{(n)} = \alpha_m \prod_{k=1}^{n} \left(\delta^{m-k} - 1\right) \frac{\delta^k}{1-\delta^k}.$$

On obtient alors la relation demandée en passant aux valeurs absolues et en notant que $|1-\delta^{m-k}| = 1-\delta^{m-k} \leqslant 1$.

(b) En passant au logarithme, on a

$$\ln\left(\prod_{k=1}^{n}(1-\delta^k)\right) = \sum_{k=1}^{n}\ln(1-\delta^k) \geqslant -K_\delta \sum_{k=1}^{n}\delta^k = -K_\delta\delta\frac{1-\delta^n}{1-\delta} \geqslant -K_\delta\frac{\delta}{1-\delta},$$

d'où

$$\prod_{k=1}^{n}(1-\delta^k) \geqslant \exp\left(-\frac{K_\delta\delta}{1-\delta}\right)$$

(c) Comme $\displaystyle\prod_{k=1}^{n}(1-\delta^k)$ est uniformément minoré par une constante positive noté $1/C_\delta$, on obtient

$$\left|\alpha_m^{(n)}\right| \leqslant C_\delta\,|\alpha_m|\left(\prod_{k=1}^{n}\delta^k\right) = C_\delta\,|\alpha_m|\,\delta^{n(n+1)/2},$$

qui est bien la relation annoncée.

(d) Au final, le terme en t^j de $A_q(\delta^{m-q}t)$ est d'ordre $\alpha_j^{(q)}(\delta^{m-q}t)^j$. On voit ainsi que le terme en t^{q+1} est le terme dominant. Ce terme est majoré par

$$\left|\alpha_{q+1}^{(q)}(\delta^{m-q}t)^{q+1}\right| \leqslant C_\delta\,|\alpha_{q+1}|\;\delta^{q(q+1)/2}(\delta^{m-q}t)^{q+1} = C_\delta\,|\alpha_{q+1}|\;\delta^{(q+1)(m-q/2)}t^{q+1},$$

ce qui est bien (2.68).

Chapitre 3

Optimisation

L'objectif de ce chapitre est l'étude de quelques problèmes d'optimisation. Le problème le plus simple, dit *d'optimisation libre* ou *sans contrainte*, peut être décrit par la donnée d'un espace vectoriel V et d'une *fonctionnelle* $J : V \to \mathbb{R}$. Ce problème consiste à

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ J(u) \leqslant J(v), \quad \forall v \in V. \end{cases} \tag{3.1}$$

On observera qu'il s'agit d'un problème de minimisation et qu'un problème de maximisation peut s'y ramener en changeant J en $-J$. La fonctionnelle J est parfois appelée *critère* ou *fonction coût*. Si u est solution de (3.1), on dit que u est un *minimiseur global* de J sur V. On dit par ailleurs que u est un *minimiseur local* de J sur V si

$$\exists \delta > 0, \quad \text{t.q.} \quad J(u) \leqslant J(v), \ \forall v \in \mathcal{B}_V(u, \delta), \tag{3.2}$$

où $\mathcal{B}_V(u, \delta) = \{v \in V, \ \|v - u\|_V \leqslant \delta\}$ désigne la boule (fermée) de centre u et de rayon δ dans la topologie de V. L'espace vectoriel V peut être de dimension finie ou infinie, étant entendu que

lorsqu'on voudra résoudre le problème numériquement, on ne pourra considérer que des problèmes de dimension finie.

Un cas plus général est celui de *l'optimisation sous contraintes*. De manière générale, les contraintes sont formulées en cherchant un minimiseur dans un sous-ensemble K de l'espace vectoriel V. On dit que K est l'ensemble des *états admissibles*. Par la suite, on supposera toujours que l'ensemble K est non-vide. Le problème d'optimisation sous contraintes consiste à

$$\begin{cases} \text{Chercher } u \in K \text{ tel que} \\ J(u) \leqslant J(v), \quad \forall v \in K. \end{cases} \tag{3.3}$$

Si u est solution de (3.3), on dit que u est un *minimiseur global* de J dans K. On dit par ailleurs que u est un *minimiseur local* de J dans K si

$$\exists \delta > 0, \quad \text{t.q.} \quad J(u) \leqslant J(v), \ \forall v \in \mathcal{B}_V(u, \delta) \cap K. \tag{3.4}$$

Dans de nombreuses applications, l'ensemble K s'exprime par le biais d'une fonctionnelle $\Phi : V \to W$. Dans ce chapitre, on se restreindra pour simplifier au cas où l'espace vectoriel W est de dimension finie. On a donc $W = \mathbb{R}^m$ avec $m \geqslant 1$. Deux cas types sont d'une part celui où

$$K = \{v \in V;\ \Phi(v) = 0\}, \tag{3.5}$$

et on parle alors de *contraintes égalité*, et d'autre part celui où

$$K = \{v \in V;\ \Phi(v) \leqslant 0\}, \tag{3.6}$$

et on parle alors de *contraintes inégalité* (la notation $\Phi(v) \leqslant 0$ signifie ici et par la suite que toutes les composantes de $\Phi(v)$ sont négatives). On notera qu'une contrainte égalité peut se reformuler par le biais de deux contraintes inégalité. Toutefois, il est souvent judicieux d'exploiter directement la forme égalité de la contrainte.

La résolution des problèmes (3.1) et (3.3) passe par la réponse à trois questions, qui, de la plus théorique à la plus appliquée, sont :

(1) existe-t-il une solution et est-elle unique ?

(2) comment caractériser cette (ou ces) solution(s) ?

(3) comment approcher de manière efficace une (la) solution ?

Les sections suivantes apporteront des éléments de réponse à ces questions, d'abord pour le problème d'optimisation sans contrainte (sections 3.2 et 3.3) puis pour le problème d'optimisation sous contraintes (section 3.4). En particulier, la notion de différentielle jouera un rôle clé dans la formulation de critères permettant de caractériser un minimiseur local. Mais voyons tout d'abord quelques exemples de problèmes d'optimisation.

3.1 Exemples de problèmes d'optimisation

Cette section présente à un niveau introductif quelques exemples de problèmes d'optimisation afin d'en illustrer le vaste champ d'applications.

3.1.1 Régression au sens des moindres carrés

Étant donné un nuage de $N \geqslant 3$ points de $\mathbb{R}^2$ de coordonnées $\{(x_i, y_i)\}_{1 \leqslant i \leqslant N}$, le problème consiste à déterminer la droite qui s'en approche le plus au sens des moindres carrés. On cherche donc deux réels a et b minimisant la quantité

$$J(a, b) = \sum_{i=1}^{N} [y_i - (a x_i + b)]^2.$$

Il s'agit d'un problème d'optimisation sans contrainte en dimension finie (égale à deux) avec $V = \mathbb{R}^2$ et la fonctionnelle $J : \mathbb{R}^2 \to \mathbb{R}$ définie ci-dessus.

Plus généralement, une matrice $A \in \mathbb{R}^{N,M}$ avec $M < N$ n'est pas de rang plein. Il existe donc des vecteurs $y \in \mathbb{R}^N$ tels que le système linéaire $Av = y$ n'admette pas de solution dans $\mathbb{R}^M$. On peut alors résoudre ce système linéaire au sens des moindres carrés en cherchant un vecteur $v \in \mathbb{R}^M$ minimisant la quantité

$$J(v) = \|Av - y\|^2_{\mathbb{R}^N},$$

où $\|\cdot\|_{\mathbb{R}^N}$ désigne la norme Euclidienne sur $\mathbb{R}^N$. Il s'agit à nouveau d'un problème d'optimisation sans contrainte en dimension finie avec $V = \mathbb{R}^M$ et la fonctionnelle $J : \mathbb{R}^M \to \mathbb{R}$ définie ci-dessus. Le problème de régression au sens des moindres carrés en constitue un cas particulier avec $M = 2$, $y \in \mathbb{R}^N$ de composantes $(y_1, \dots, y_N)$, $v \in \mathbb{R}^2$ de composantes (a, b) et la matrice A donnée par

$$A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix}.$$

3.1.2 Exemples en économie

Reprenons tout d'abord deux exemples vus en cours d'Économie. Considérons une entreprise fabriquant un seul produit. La quantité y de ce produit (ou output) qui est obtenue lorsque l'entreprise utilise des quantités $(x_1, \dots, x_N)$ de N facteurs de production (ou inputs) est donnée par une fonction de production $f(x_1, \dots, x_N)$. On suppose que l'entreprise a un comportement rationnel et qu'elle décide parmi toutes les combinaisons de facteurs de production possibles pour produire une quantité y d'output donnée de choisir celle qui lui revient le moins cher, c'est-à-dire celle qui minimise les coûts de production. Si le prix de l'input i est noté p_i et que l'entreprise en consomme une quantité x_i, le coût de production total, appelé en économie *fonction de dépense*, est donné par

$$J(x_1, \dots, x_N) = \sum_{i=1}^{N} p_i x_i.$$

Nous avons affaire à un problème d'optimisation sous contraintes en dimension finie avec $V = \mathbb{R}^N$, la fonctionnelle J définie ci-dessus, la contrainte égalité $y = f(x_1, \dots, x_N)$ et les contraintes inégalité $x_i \geqslant 0$ pour tout $1 \leqslant i \leqslant N$. Les données du problème sont les prix $\{p_i\}_{1 \leqslant i \leqslant N}$ et le niveau de production y, l'inconnue étant le vecteur de $\mathbb{R}^N$ de composantes $(x_1, \dots, x_N)$.

Considérons maintenant un modèle de consommateur. On suppose que si celui-ci consomme N biens en quantités $(x_1, \dots, x_N)$, son degré de satisfaction peut être quantifié par le biais d'une fonction d'utilité $f(x_1, \dots, x_N)$. Le consommateur, qui dispose d'un capital C alloué aux biens de consommation, cherche à maximiser son utilité dans la limite du budget disponible. Il s'agit à nouveau d'un problème d'optimisation sous contrainte en dimension finie avec $V = \mathbb{R}^N$, la fonctionnelle $J = -f$ et les contraintes inégalité $\sum_{i=1}^N p_i x_i \leqslant C$ et $x_i \geqslant 0$ pour tout $1 \leqslant i \leqslant N$. Les données du problème sont les prix des biens $\{p_i\}_{1 \leqslant i \leqslant N}$ et le capital disponible C, l'inconnue étant le vecteur de $\mathbb{R}^N$ de composantes $(x_1, \dots, x_N)$.

Le dernier exemple de cette section s'inscrit dans le cadre (très important et bien plus général) des problèmes de décision ou de commande optimale. Conduire une voiture, piloter un avion, exploiter durablement des ressources naturelles relèvent de ce cadre. Nous nous contenterons ici d'une illustration économique (dans un cadre très idéalisé !). On considère un consommateur souhaitant utiliser sa richesse de manière optimale sur N périodes de temps de durée unité. Sa richesse à l'instant n est notée w_n et sa consommation sur la période $[n, n+1[$ est notée c_n. On suppose que la richesse non consommée sur cette période est investie dans un portefeuille financier de rendement certain R. On obtient ainsi la dynamique suivante pour la richesse du consommateur :

$$w_{n+1} = R(w_n - c_n), \qquad \forall 0 \leqslant n \leqslant N-1. \tag{3.7}$$

Le « bien-être » du consommateur à l'instant final N est représenté par la somme des utilités de ses consommations successives et de sa richesse finale sous la forme

$$f(c_0, \dots, c_{N-1}) = \sum_{n=0}^{N-1} L_1(c_n) + L_2(w_N),$$

où L_1 et L_2 sont des fonctions données de $\mathbb{R}$ dans $\mathbb{R}$. Le consommateur souhaite maximiser son bien-être. Pour cela, il est amené à résoudre un problème d'optimisation sous contraintes en dimension finie avec $V = \mathbb{R}^N$, $J = -f$ et les contraintes données par la dynamique (3.7) et le fait que les consommations sont positives (celles-ci devraient également être bornées supérieurement par la richesse instantanée, à moins qu'il ne soit possible d'emprunter...). L'inconnue est le vecteur de consommations $(c_0, \dots, c_{N-1})$.

Le problème de commande optimale décrit ci-dessus a été formulé en temps discret et en l'absence d'aléa. Plus généralement, on peut s'intéresser à des problèmes de commande optimale en temps continu et en présence d'incertitudes.

3.1.3 Formulation variationnelle et principe de moindre énergie

Considérons le problème de Poisson avec conditions aux limites de Dirichlet homogènes posé sur un ouvert borné Ω de $\mathbb{R}^d$ avec une donnée $f \in L^2(\Omega)$. La formulation faible de ce problème consiste à

$$\begin{cases} \text{Chercher } u \in H_0^1(\Omega) \text{ tel que} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv, \quad \forall v \in H_0^1(\Omega). \end{cases} \tag{3.8}$$

Ce problème admet une et une seule solution (cf. cours d'Analyse) : en posant $V = H_0^1(\Omega)$ et en introduisant la forme bilinéaire

$$a : V \times V \ni (u, v) \longmapsto a(u, v) = \int_\Omega \nabla u \cdot \nabla v \in \mathbb{R},$$

et la forme linéaire

$$b : V \ni v \longmapsto b(v) = \int_\Omega fv \in \mathbb{R},$$

toutes les hypothèses du théorème de Lax–Milgram sont satisfaites.

En exploitant la symétrie de la forme bilinéaire a, on montre (nous le reprendrons à la section 3.2.3) que résoudre le problème (3.8) équivaut à trouver le minimiseur dans $H_0^1(\Omega)$ de la fonctionnelle d'énergie

$$J(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 - \int_\Omega fv. \tag{3.9}$$

On parle de *formulation variationnelle*. Il s'agit d'un problème d'optimisation sans contrainte posé en dimension infinie. Dans le cadre de la mécanique des milieux continus en élasticité linéaire, la formulation variationnelle exprime le principe de *moindre énergie*, le terme $\frac{1}{2}\int_\Omega |\nabla v|^2$ dans la fonctionnelle d'énergie représentant l'énergie élastique de déformation et le terme $-\int_\Omega fv$ l'énergie potentielle sous le chargement extérieur f.

3.1.4 Problèmes inverses et contrôle optimal

Considérons un système dont l'état peut être décrit par une fonction v dans un espace fonctionnel V. On dit que v est la *variable d'état*. Supposons que l'état du système dépend d'un paramètre y qui peut être un scalaire, un vecteur ou une fonction (de l'espace, du temps ou des deux). L'espace des paramètres est noté Y. La dépendance de l'état du système en le paramètre y se formule par le biais d'une *équation d'état* faisant intervenir une application $\Psi : Y \to V$ sous la forme $v = \Psi(y)$.

Un premier exemple important d'une telle situation est celui où l'état du système est solution dans $H_0^1(\Omega)$ de l'équation

$$-\nabla \cdot (y \nabla v) = f \quad \text{dans } \Omega. \tag{3.10}$$

La fonction f est donnée dans $L^2(\Omega)$ et le paramètre y est (pour simplifier) un réel strictement positif ($Y := \mathbb{R}_+^*$). Pour chaque $y \in Y$ fixé, il existe une et une seule solution au problème ci-dessus. On la note $\Psi(y)$. Considérons maintenant le problème suivant : étant donnée une fonction $v_0 \in H_0^1(\Omega)$, déterminer la valeur du paramètre y minimisant l'écart en norme H^1 entre les fonctions $\Psi(y)$ et v_0. Il s'agit d'un problème d'optimisation sous contraintes posé sur $\mathbb{R}$ avec la fonctionnelle

$$J(y) = \|\Psi(y) - v_0\|_{H^1}^2.$$

La contrainte est $y > 0$. Voici un exemple d'application : considérons un sol occupant un domaine $\Omega \subset \mathbb{R}^3$. Celui-ci est modélisé comme un milieu poreux constitué d'un squelette solide indéformable et d'un réseau de pores au-travers desquels s'écoule un fluide (de l'eau). L'état du système est décrit par le champ de pression v du fluide qui sous certaines hypothèses est régi par l'équation (3.10), la fonction f représentant un terme source (supposé connu) et le paramètre y la perméabilité hydraulique du milieu. L'équation (3.10) porte le nom d'équation de Darcy. Le paramètre y peut être difficile à déterminer expérimentalement. Le problème d'optimisation sous contraintes ci-dessus consiste donc à en déterminer une valeur optimale en cherchant à minimiser l'écart entre le champ de pression issu du modèle et un champ de pression v_0 qui peut être, par exemple, le fruit d'observations sur le terrain. On parle de problème *d'identification de paramètre* ou de *problème inverse*.

Un deuxième exemple est celui où l'état du système est solution dans $H_0^1(\Omega)$ de l'équation

$$-\Delta v = f + y \quad \text{dans } \Omega. \tag{3.11}$$

La fonction f est donnée dans $L^2(\Omega)$ et le paramètre y est une fonction dans $Y := L^2(\Omega)$. Pour chaque $y \in Y$ fixé, il existe une et une seule solution au problème (3.11). On la note à nouveau $\Psi(y)$. Considérons maintenant le problème suivant : étant donnée une fonction $v_0 \in L^2(\Omega)$, déterminer la valeur du paramètre y qui minimise le critère suivant :

$$J(y) = \int_\Omega |\Psi(y) - v_0|^2 + \int_\Omega |y|^2.$$

Voici un exemple d'application : considérons une pièce occupant le volume $\Omega \subset \mathbb{R}^3$ dans laquelle se trouvent plusieurs radiateurs dont nous pouvons faire varier la position et l'intensité. Nous souhaitons placer les radiateurs et déterminer leur intensité de façon à ce que le champ de température dans la pièce soit le plus proche possible du champ de température v_0. De plus, nous souhaitons réaliser cette opération en dépensant le moins d'énergie possible. La conductivité thermique dans la pièce est fixée à un et les pertes de chaleur sont décrites par la fonction $f \in L^2(\Omega)$. La fonction $y \in L^2(\Omega)$ représente les sources d'énergie apportées par les radiateurs. Le premier terme dans le critère J rend compte du souhait de viser le champ de température v_0, le deuxième terme celui de minimiser la dépense. Ces deux termes étant positifs, la minimisation du critère J va conduire à un compromis entre ces deux souhaits (un facteur de pondération aurait pu être introduit entre les deux termes). On parle de problème de *contrôle optimal*.

3.2 Optimisation sans contrainte : bases théoriques

L'objectif de cette section est d'apporter des éléments de réponse aux questions 1 et 2 formulées ci-dessus pour le problème d'optimisation sans contrainte (3.1). Cette section contient des rappels et compléments sur des notions vues en cours d'Analyse. Nous les exposons de manière détaillée afin de disposer d'une présentation complète des notions qui nous seront utiles par la suite.

3.2.1 Existence et unicité

Théorème 3.1 (Existence, dimension finie)**.** *On suppose que l'espace vectoriel V est de dimension* finie*. On suppose de plus que*

(i) *J est* continue *dans V,*

(ii) *J est* coercive *dans V, ce qui signifie que $J(v) \to +\infty$ quand $\|v\|_V \to +\infty$.*
Alors, J admet au moins un minimiseur global dans V.

Preuve. La coercivité de J dans V implique que

$$\forall M \in \mathbb{R}, \quad \exists R \in \mathbb{R}, \quad \|v\|_V \geqslant R \Longrightarrow J(v) \geqslant M.$$

En considérant cette propriété pour $M = J(0)$, il vient

$$\exists R \in \mathbb{R}, \quad \|v\|_V \geqslant R \Longrightarrow J(v) \geqslant J(0).$$

Par suite,

$$\inf_{v \in V} J(v) = \inf_{v \in \mathcal{B}_V(0,R)} J(v).$$

Or, en dimension finie, la boule $\mathcal{B}_V(0, R)$ est compacte. La fonctionnelle J étant continue, elle y atteint son infimum. □

Le théorème 3.1 ne s'étend pas à la dimension infinie, car dans ce cas, la boule $B_R(0)$ n'est jamais compacte. Pour pouvoir formuler un résultat d'existence en dimension infinie, il nous faut rajouter une hypothèse. Un exemple d'une telle hypothèse nous est fourni par la notion de convexité.

Définition 3.2 (Convexité, stricte convexité, forte convexité)**.** *Soit V un espace vectoriel et $J : V \to \mathbb{R}$.*

- *On dit que J est* convexe *dans V si*

$$\forall (v, w) \in V \times V, \quad \forall \theta \in [0, 1], \quad J(\theta v + (1-\theta)w) \leqslant \theta J(v) + (1-\theta)J(w). \tag{3.12}$$

- *On dit que J est* strictement convexe *dans V si l'inégalité ci-dessus est stricte lorsque $\theta \in]0, 1[$ et $v \neq w$.*
- *On dit que J est* fortement convexe *dans V de paramètre $\alpha > 0$ (on dit également que J est α-convexe) si*

$$\forall (v, w) \in V \times V, \quad \forall \theta \in [0, 1], \quad J(\theta v + (1-\theta)w) \leqslant \theta J(v) + (1-\theta)J(w) - \alpha \frac{\theta(1-\theta)}{2} \|v - w\|_V^2. \tag{3.13}$$

Il est clair qu'une fonctionnelle fortement convexe est strictement convexe et qu'une fonctionnelle strictement convexe est convexe. La notion de convexité nous sera utile pour l'existence d'un minimiseur en dimension infinie, celle de stricte convexité pour l'unicité du minimiseur (en dimension finie ou non) et celle de forte convexité pour l'existence, l'unicité et également pour l'étude de la convergence de certains algorithmes d'optimisation numérique.

Un exemple fondamental de fonctionnelle fortement convexe dans un espace de Hilbert V est $J(v) = \|v\|_V^2$ où $\|\cdot\|_V$ désigne la norme induite par le produit scalaire dans V. En effet, dans ce cas, il vient pour tout $(v, w) \in V \times V$ et pour tout $\theta \in [0, 1]$,

$$\begin{aligned} \|\theta v + (1-\theta)w\|_V^2 - \theta\|v\|_V^2 - (1-\theta)\|w\|_V^2 &= \theta(\theta - 1)\|v\|_V^2 + 2\theta(1-\theta)(v, w)_V + (1-\theta)\theta\|w\|_V^2 \\ &= -\theta(1-\theta)\|v - w\|_V^2, \end{aligned}$$

d'où l'(in)égalité (3.13) avec $\alpha = 2$. Un autre exemple qui nous sera utile est la forte convexité de la fonctionnelle $J(v) = \|\nabla v\|_{L^2}^2$ dans $H_0^1(\Omega)$ pour Ω ouvert borné de $\mathbb{R}^d$. En effet,

$$J(\theta v + (1-\theta)w) - \theta J(v) - (1-\theta)J(w) = -\theta(1-\theta)\|\nabla(v - w)\|_{L^2}^2 \leqslant -\frac{\theta(1-\theta)}{1 + c_\Omega^2}\|v - w\|_{H^1}^2,$$

grâce à l'inégalité de Poincaré vue en cours d'Analyse et rappelée ci-dessous. D'où (3.13) avec $\alpha = \frac{2}{1+c_\Omega^2}$.

Lemme 3.3 (Poincaré)**.** *Soit Ω un ouvert borné de $\mathbb{R}^d$. Alors, il existe une constante strictement positive c_Ω ne dépendant que de Ω et telle que*

$$\forall v \in H_0^1(\Omega), \qquad \|v\|_{L^2} \leqslant c_\Omega \|\nabla v\|_{L^2}. \tag{3.14}$$

Remarque (Attention à la dimension infinie !)**.** La notion de forte convexité comporte des subtilités en dimension infinie, du fait que toutes les normes ne sont pas équivalentes (les notions de convexité et de stricte convexité ne faisant pas appel à la topologie). Par exemple, la fonctionnelle $J(v) = \int_\Omega v^2$ est fortement convexe dans $L^2(\Omega)$, mais elle n'est pas fortement convexe dans $H_0^1(\Omega)$ (en effet, dans le cas contraire, on pourrait trouver une constante C telle que, pour tout $v \in H_0^1(\Omega)$, $\|\nabla v\|_{L^2} \leqslant C\|v\|_{L^2}$ ...) □

Examinons maintenant la question de l'existence d'un minimiseur en dimension infinie. Nous admettons le résultat suivant.

Théorème 3.4 (Existence, dimension infinie)**.** *Soit V un* espace de Hilbert*. On suppose que*
(i) *J est* continue *dans V,*
(ii) *J est* coercive *dans V,*
(iii) *J est* convexe *dans V.*
Alors, J admet au moins un minimiseur global dans V.

Passons maintenant à la question de l'unicité du minimiseur.

Théorème 3.5 (Unicité)**.** *Soit V un espace vectoriel et J une fonctionnelle* strictement convexe *dans V. Alors, J admet au plus un minimiseur global dans V.*

Preuve. Si u_1 et u_2 sont deux minimiseurs distincts, on déduit de la stricte convexité de J que

$$J\left(\frac{u_1+u_2}{2}\right) < \frac{1}{2}J(u_1) + \frac{1}{2}J(u_2) = \inf_{v\in V} J(v),$$

ce qui fournit une contradiction. □

En dimension finie, toute fonctionnelle convexe est continue, mais ce n'est plus nécessairement le cas en dimension infinie (ne serait-ce que parce qu'il existe des applications linéaires de V dans $\mathbb{R}$ qui ne sont pas continues). L'hypothèse (i) n'est donc pas redondante avec l'hypothèse (iii). Par ailleurs, dans un espace de Hilbert, toute fonctionnelle convexe et continue peut être minorée par une fonctionnelle affine ; plus précisément, on peut montrer qu'il existe $p \in V$ et $\delta \in \mathbb{R}$ tels que

$$\forall v \in V, \qquad J(v) \geqslant (p, v)_V + \delta. \tag{3.15}$$

Toutefois, cette propriété n'est pas suffisante pour garantir la coercivité de J. L'hypothèse (ii) n'est donc pas inutile. Par contre, si J est fortement convexe, on peut montrer, même en dimension infinie, qu'il existe $\gamma > 0$ et $\delta \in \mathbb{R}$ tels que

$$\forall v \in V, \qquad J(v) \geqslant \gamma\|v\|_V^2 + \delta. \tag{3.16}$$

Par conséquent, si la fonctionnelle J est fortement convexe et continue dans un espace de Hilbert V, toutes les hypothèses des théorèmes 3.4 et 3.5 sont satisfaites, ce qui fournit l'existence et l'unicité du minimiseur.

Corollaire 3.6 (Existence et unicité)**.** *Soient V un* espace de Hilbert *et J une fonctionnelle* fortement convexe *dans V. Si l'espace V est de dimension infinie, on suppose de plus que J est* continue *dans V. Alors, J admet un et un seul minimiseur global dans V.*

Remarque (Preuve d'existence)**.** L'hypothèse de forte convexité permet de montrer l'existence du minimiseur en utilisant uniquement des notions vues en cours d'Analyse. Comme J est fortement convexe dans V, la propriété (3.16) montre que la fonctionnelle J est minorée dans V. Il est donc loisible de considérer une suite minimisante, c'est-à-dire une suite $(v_n)_{n\in\mathbb{N}}$ de V telle que

$J(v_n) \to \inf_{v\in V} J(v)$ quand $n \to +\infty$. Soit α le paramètre de forte convexité de J dans V. Il vient pour tout $m, n \geqslant 0$,

$$\frac{\alpha}{8}\|v_n - v_m\|_V^2 + \underbrace{J\left(\frac{v_n+v_m}{2}\right) - \inf_{v\in V} J(v)}_{\geqslant 0} \leqslant \frac{1}{2}\left(J(v_n) - \inf_{v\in V} J(v)\right) + \frac{1}{2}\left(J(v_m) - \inf_{v\in V} J(v)\right),$$

ce qui montre que la suite $(v_n)_{n\in\mathbb{N}}$ est *de Cauchy* dans V, donc converge vers une limite $u \in V$. La fonctionnelle J étant continue, $J(u) = \inf_{v\in V} J(v)$. D'où l'existence du minimiseur global de J dans V. □

3.2.2 Caractérisation

Il est bien connu qu'une *condition nécessaire* pour qu'un point $x_0 \in \mathbb{R}$ soit un minimiseur local d'une fonction dérivable $f : \mathbb{R} \to \mathbb{R}$ est que la dérivée f' s'annule en x_0. Ce critère peut se généraliser aux fonctionnelles *différentiables* de V dans $\mathbb{R}$.

Définition 3.7 (Différentielle)**.** *Soient V un* espace de Hilbert*, $v \in V$ et $J : V \to \mathbb{R}$.*

– *On dit que J est différentiable en v s'il existe une application linéaire continue $J'(v) \in V'$ telle que*

$$\forall w \in V, \qquad J(v+w) = J(v) + \langle J'(v), w\rangle_{V',V} + \mathrm{o}(\|w\|_V), \tag{3.17}$$

avec $\lim_{w\to 0} \mathrm{o}(\|w\|_V)/\|w\|_V = 0$. On dit que $J'(v)$ est la différentielle de J en v (observer que $J'(v) : V \to \mathbb{R}$). De par le théorème de représentation de Riesz–Fréchet vu en cours d'Analyse, la différentiabilité équivaut à l'existence d'un vecteur $\nabla J(v) \in V$, appelé gradient *de J en v, tel que*

$$\forall w \in V, \qquad J(v+w) = J(v) + (\nabla J(v), w)_V + \mathrm{o}(\|w\|_V), \tag{3.18}$$

c'est-à-dire que le gradient de J en v est le représentant dans V de la forme linéaire continue $J'(v) \in V'$,

$$\forall w \in V, \qquad \langle J'(v), w\rangle_{V',V} = (\nabla J(v), w)_V. \tag{3.19}$$

– *On dit que J est différentiable dans V si pour tout $v \in V$, J est différentiable en v.*

Une observation importante est qu'une fonctionnelle différentiable en $v \in V$ y est a fortiori continue.

En pratique, pour montrer qu'une fonctionnelle J est différentiable en $v \in V$, on se donne $w \in V$ et on développe $J(v+w)$ sous la forme

$$J(v+w) = J(v) + T_1(v,w) + T_2(v,w), \tag{3.20}$$

où $T_1(v,w)$ regroupe tous les termes d'ordre 1 en w et T_2 regroupe tous les termes d'ordre supérieur ou égal à 2 en w. On vérifie ensuite que (i) $T_1(v,w)$ est linéaire et continue en w (la continuité signifiant qu'il existe C tel que pour tout $w \in V$, $|T_1(v,w)| \leqslant C\|w\|_V$, la constante C pouvant dépendre de v) et (ii) $\lim_{w\to 0} T_2(v,w)/\|w\|_V = 0$.

En dimension finie avec $V = \mathbb{R}^N$ muni d'une base cartésienne et du produit scalaire usuel, la différentielle et le gradient de J en $v = (v_1, \ldots, v_N)^t$ s'expriment en fonction des dérivées partielles de J par rapport à chacune des composantes v_i sous la forme

$$J'(v) = \left(\tfrac{\partial J}{\partial v_1}(v), \ldots, \tfrac{\partial J}{\partial v_N}(v)\right), \qquad \nabla J(v) = \begin{pmatrix} \frac{\partial J}{\partial v_1}(v) \\ \vdots \\ \frac{\partial J}{\partial v_N}(v) \end{pmatrix}.$$

On a en effet en notant $(w_1, \ldots, w_N)$ les composantes d'un vecteur $w \in \mathbb{R}^N$,

$$J(v+w) = J(v) + \sum_{i=1}^{N} \frac{\partial J}{\partial v_i}(v) w_i + \mathrm{o}(\|w\|_{\mathbb{R}^N}).$$

En dimension finie, le lien entre différentielle et gradient étant immédiat, nous manipulerons plutôt le gradient.

La situation est plus subtile en dimension infinie. On notera la différence de nature entre la différentielle et le gradient : $J'(v) \in V'$ et $\nabla J(v) \in V$. Par la suite, nous manipulerons plutôt la différentielle $J'(v)$ en dimension infinie car il est relativement simple de spécifier l'action de $J'(v)$ sur un vecteur $w \in V$ alors que l'obtention de l'expression de $\nabla J(v) \in V$ n'est pas toujours immédiate. Pour s'en persuader, considérons l'exemple suivant : $V = H_0^1(\Omega)$ et $J(v) = \frac{1}{2}\int_\Omega v^2$. Il vient pour tout $w \in H_0^1(\Omega)$,

$$J(v+w) = \frac{1}{2}\int_\Omega v^2 + \int_\Omega vw + \frac{1}{2}\int_\Omega w^2.$$

Il est clair que $\int_\Omega w^2 = \mathrm{o}(\|w\|_{H^1})$ (noter la norme utilisée) et que l'application linéaire

$$H_0^1(\Omega) \ni w \longmapsto \int_\Omega vw \in \mathbb{R}$$

est continue dans V puisque $|\int_\Omega vw| \leqslant \|v\|_{L^2}\|w\|_{L^2} \leqslant \|v\|_{L^2}\|w\|_{H^1} = C\|w\|_{H^1}$ avec $C = \|v\|_{L^2}$ (noter à nouveau la norme utilisée pour w). On en déduit que

$$\langle J'(v), w\rangle_{V',V} = \int_\Omega vw.$$

En revanche, trouver l'expression de $\nabla J(v)$ revient à chercher la fonction $G := \nabla J(v) \in H_0^1(\Omega)$ telle que pour tout $w \in H_0^1(\Omega)$,

$$(G, w)_{H^1} = \langle J'(v), w\rangle_{V',V} = \int_\Omega vw.$$

En développant le produit scalaire du membre de gauche, il vient

$$\int_\Omega Gw + \int_\Omega \nabla G\cdot\nabla w = \int_\Omega vw.$$

Ce problème admet bien sûr une et une seule solution, mais on ne dispose pas d'une expression explicite pour sa solution $G \in H_0^1(\Omega)$: il faudrait en effet résoudre le problème $G - \Delta G = v$ dans Ω avec conditions aux limites de Dirichlet homogènes. Pour terminer sur une note plus positive, signalons un cas particulier (important !) où le calcul du gradient est possible (et très simple) même en dimension infinie, c'est celui où $J(v) = \frac{1}{2}\|v\|_V^2$. En effet, on vérifie aisément que $\nabla J(v) = v$.

Remarque (Différentielle au sens de Gateaux)**.** La définition 3.7 correspond à la notion de différentiabilité *au sens de Fréchet*. On dit que la fonctionnelle J est différentiable *au sens de Gateaux* en v s'il existe une application linéaire continue $J'(v) \in V'$, ou de manière équivalente un vecteur $\nabla J(v) \in V$, tels que

$$\begin{aligned}\forall w \in V, \qquad J(v+tw) &= J(v) + t\langle J'(v), w\rangle_{V',V} + \mathrm{o}(t)\\ &= J(v) + t(\nabla J(v), w)_V + \mathrm{o}(t),\end{aligned}$$

avec $\lim_{t\to 0} \mathrm{o}(t)/t = 0$ à w fixé. Toute fonctionnelle différentiable au sens de Fréchet l'est évidemment au sens de Gateaux. La réciproque est fausse, même en dimension finie. Une fonctionnelle peut même être différentiable au sens de Gateaux sans être continue. À titre d'illustration, on pourra considérer l'exemple suivant dans $V = \mathbb{R}^2$ avec $v = (x, y)$:

$$J(x,y) = \frac{x^6}{(y-x^2)^2 + x^8},\ (x,y) \neq (0,0) \qquad \text{et} \qquad J(0,0) = 0.$$

J est différentiable au sens de Gateaux en $(0,0)$ mais pas continue (s'en persuader en approchant $(0,0)$ par exemple le long de la parabole $\{y = x^2\}$). □

Théorème 3.8 (Condition d'Euler, point critique)**.** *Soient V un espace de Hilbert et J une fonctionnelle de V dans $\mathbb{R}$. On suppose que la fonctionnelle J admet un minimum local en $u \in V$ et que J est différentiable en u. Alors,*

$$J'(u) = 0 \ (\in V'). \tag{3.21}$$

Un vecteur $u \in V$ vérifiant (3.21) *est appelé* point critique *de J.*

Preuve. Soit $v \in V$. Pour tout $t \in \mathbb{R}$ suffisamment petit,

$$J(u) \leqslant J(u + tv) = J(u) + t\langle J'(u), v\rangle_{V',V} + \mathrm{o}(t).$$

En faisant $t \to 0$ d'abord avec $t > 0$ puis $t < 0$, on obtient $\langle J'(u), v\rangle_{V',V} = 0$ et comme v est arbitraire dans V, il vient $J'(u) = 0$. □

On peut facilement caractériser les fonctions convexes différentiables comme le montrent les résultats suivants (leur preuve est laissée en exercice).

Proposition 3.9 (Caractérisation de la convexité)**.** *Soit $J : V \to \mathbb{R}$ une fonctionnelle différentiable dans V. Les assertions suivantes sont équivalentes :*
(i) *J est* convexe *dans V ;*
(ii) *pour tout $(v, w) \in V \times V$, $J(w) \geqslant J(v) + \langle J'(v), w - v\rangle_{V',V}$;*
(iii) *pour tout $(v, w) \in V \times V$, $\langle J'(w) - J'(v), w - v\rangle_{V',V} \geqslant 0$.*

Proposition 3.10 (Caractérisation de la forte convexité)**.** *Soit $J : V \to \mathbb{R}$ une fonctionnelle différentiable dans V. Les assertions suivantes sont équivalentes :*
(i) *J est* fortement convexe *dans V (de paramètre α) ;*
(ii) *pour tout $(v, w) \in V \times V$, $J(w) \geqslant J(v) + \langle J'(v), w - v\rangle_{V',V} + \frac{\alpha}{2}\|w - v\|_V^2$;*
(iii) *pour tout $(v, w) \in V \times V$, $\langle J'(w) - J'(v), w - v\rangle_{V',V} \geqslant \alpha\|w - v\|_V^2$.*

Nous sommes maintenant en mesure de formuler une condition nécessaire et suffisante pour qu'un vecteur $u \in V$ soit un minimiseur d'une fonctionnelle convexe et différentiable.

Corollaire 3.11 (Condition nécessaire et suffisante)**.** *Soient V un* espace de Hilbert *et J une fonctionnelle* convexe et différentiable *de V dans $\mathbb{R}$. Alors, $u \in V$ est un minimiseur global de J dans V si et seulement si u est point critique de J.*

Preuve. L'implication résulte du théorème 3.8. La réciproque est une conséquence du point (ii) de la proposition 3.9 puisque $J'(u) = 0$ implique que $J(v) \geqslant J(u)$ pour tout $v \in V$. □

3.2.3 Formulation variationnelle et principe de moindre énergie

Soient V un espace de Hilbert, b une forme linéaire dans V et a une forme bilinéaire dans $V \times V$. On suppose que la forme a est *symétrique* :

$$\forall (v, w) \in V \times V, \qquad a(v, w) = a(w, v).$$

Introduisons la *fonctionnelle d'énergie*

$$J : V \ni v \longmapsto J(v) = \frac{1}{2}a(v, v) - b(v) \in \mathbb{R}, \tag{3.22}$$

et considérons le problème d'optimisation sans contrainte (3.1).

Lemme 3.12 (Différentielle de J)**.** *On suppose que a est symétrique et continue et que b est continue. Alors, la fonctionnelle J est différentiable dans V et on a*

$$\forall u \in V, \qquad \langle J'(u), v\rangle_{V',V} = a(u, v) - b(v), \ \forall v \in V. \tag{3.23}$$

Preuve. La preuve repose sur l'identité suivante :

$$\begin{aligned} J(u+v) &= \tfrac{1}{2}a(u+v,u+v) - b(v) \\ &= J(u) + [a(u,v) - b(v)] + \tfrac{1}{2}a(v,v), \end{aligned}$$

où nous avons utilisé la bilinéarité et la symétrie de la forme a et la linéarité de la forme b. De par la continuité de la forme a, $a(v,v) = \mathrm{o}(\|v\|_V)$ et de par la continuité des formes a et b, l'application linéaire $V \ni v \mapsto a(u,v) - b(v)$ est continue. □

Proposition 3.13 (Caractérisation du minimiseur)**.** *On suppose de plus que a est coercive. Alors, u est un minimiseur global de J dans V si et seulement si*

$$\forall v \in V, \qquad a(u,v) = b(v). \tag{3.24}$$

De plus, ce minimiseur existe et est unique.

Preuve. La coercivité de la forme a implique la forte convexité de la fonctionnelle J. En effet, en utilisant (3.23) et le critère (iii) de la proposition 3.10, il vient

$$\langle J'(w) - J'(v), w - v\rangle_{V',V} = a(w-v, w-v) \geqslant \alpha\|w-v\|_V^2.$$

La caractérisation du minimiseur résulte du corollaire 3.11 (la convexité de J suffit). De par la forte convexité de J, ce minimiseur existe et est unique (corollaire 3.6 ; rappelons que J est continue car différentiable). □

Ainsi, la formulation faible des problèmes elliptiques vue en cours d'Analyse s'interprète, pourvu que la forme bilinéaire a soit bien symétrique, comme la condition d'Euler associée à la recherche du point critique de la fonctionnelle J définie par (3.22) : (3.24) est une récriture de

$$J'(u) = 0 \ (\in V').$$

Observons pour conclure que dans le cadre de la méthode de Galerkine que nous étudierons au chapitre 4 conscaré à l'approximation par éléments finis des problèmes elliptiques, on se donne un espace d'approximation $V_h \subset V$ et on cherche $u_h \in V_h$ tel que

$$\forall v_h \in V_h, \qquad a(u_h, v_h) = b(v_h).$$

Ce problème est équivalent à chercher un minimiseur global de la fonctionnelle d'énergie J dans l'espace vectoriel V_h. Il s'agit cette fois d'un problème d'optimisation sans contrainte posé en dimension *finie*. La méthode de Galerkine consiste donc à remplacer la recherche d'un minimiseur dans un espace de dimension infinie par celle d'un minimiseur dans un espace de dimension finie. Comme $V_h \subset V$, on a

$$J(u_h) \geqslant J(u). \tag{3.25}$$

Ainsi, l'énergie de la solution approchée par éléments finis est toujours supérieure à l'énergie de la solution exacte. Insistons sur le fait que cette propriété est vraie si la forme bilinéaire avec laquelle on travaille est *symétrique*.

3.3 Optimisation numérique sans contrainte

L'objectif de cette section est l'étude de quelques algorithmes d'optimisation numérique sans contrainte afin d'approcher une solution du problème (3.1). Nous présentons le principe général de ces algorithmes et formulons des conditions suffisantes pour qu'ils convergent. Comme les algorithmes ci-dessous sont destinés à être mis en œuvre sur ordinateur, nous nous plaçons en dimension finie. Nous supposons que la fonctionnelle J est différentiable (et utilisons son gradient plutôt que sa différentielle).

3.3.1 Méthodes de descente

Définition 3.14 (Direction de descente)**.** *Soient V un espace vectoriel, $v \in V$ et $J : V \to \mathbb{R}$. On dit que $d \in V$ est une* direction de descente *de J en v si*

$$\exists \delta > 0, \quad \forall t \in [0, \delta], \quad J(v + td) \leqslant J(v). \tag{3.26}$$

On dit que la direction de descente est stricte *si l'inégalité* (3.26) *est stricte pour $t > 0$.*

On observera que la notion de direction de descente est locale : rien ne garantit que $J(v+td) \leqslant J(v)$ si le réel t est choisi trop grand.

Proposition 3.15 (Gradient et direction de descente)**.** *On suppose que J est différentiable en v. Alors, si d est direction de descente de J en v, on a*

$$(\nabla J(v), d)_V \leqslant 0. \tag{3.27}$$

De plus, si $\nabla J(v) \neq 0$, alors $d = -\nabla J(v)$ est une direction de descente stricte de J en v.

Preuve. Comme J est différentiable en v, on a en prenant $w = tv$ dans la définition 3.7,

$$J(v + td) = J(v) + t(\nabla J(v), d)_V + \mathrm{o}(t),$$

d'où les résultats énoncés en prenant t suffisamment petit. □

3.3.2 Algorithmes de gradient

Les algorithmes de gradient sont des méthodes itératives dont le but est de construire une suite $(v^k)_{k\in\mathbb{N}}$ de V qui converge vers un point critique de la fonctionnelle J. Le principe, basé sur la proposition 3.15, consiste à utiliser la direction de descente $-\nabla J(v^k)$ pour passer de v^k à v^{k+1}. Ces algorithmes nécessitent donc le calcul du gradient de la fonctionnelle J à chaque itération. Ils sont applicables lorsque J est différentiable, ce que nous supposerons dans toute cette section. On parle *d'algorithmes d'ordre 1.*

La version la plus simple de l'algorithme de gradient est celle à *pas fixe* (voir l'exercice 5 pour la version à pas optimal). L'algorithme de gradient à pas fixe est le suivant :

1. Initialisation : choisir $v^0 \in V$, poser $k = 0$; fixer le pas $\lambda > 0$ et le seuil de convergence $\varepsilon > 0$;
2. Boucle en k : pour $k \geqslant 0$, effectuer les opérations suivantes :
   (2.a) calculer le gradient de la fonctionnelle J en v^k, $\nabla J(v^k)$, et choisir pour direction de descente
   $$d^k = -\nabla J(v^k).$$
   Si $d^k = 0$, les itérations s'arrêtent car v^k est point critique de J : l'algorithme a convergé.
   (2.b) déterminer v^{k+1} selon la formule
   $$v^{k+1} = v^k + \lambda d^k.$$
   (2.c) tant que $\|v^{k+1} - v^k\|_V > \varepsilon$, poser $k \leftarrow k + 1$ et revenir à l'étape (2.a).

Remarque (Critère de convergence)**.** Le critère de convergence retenu à l'étape (2.c) n'est pas le seul possible. On peut le remplacer par un critère normalisé de la forme $\|v^{k+1} - v^k\|_V > \varepsilon\|v^0\|_V$ ou par un critère normalisé portant sur les variations de $J(v^k)$, par exemple $|J(v^{k+1}) - J(v^k)| > \varepsilon|J(v^0)|$. La différence entre un critère (normalisé) portant sur les variations de v^k et un critère portant sur celles de $J(v^k)$ sont faciles à appréhender sur un exemple simple où $V = \mathbb{R}$ et où $J(v) = \alpha v^2$ avec $\alpha > 0$. Lorsque $\alpha \gg 1$, le minimiseur ($v = 0$) se trouve au fond d'une cuvette très étroite. Dans ce cas, des petites variations de v^k peuvent induire des fortes variations de $J(v^k)$; aussi est-il plus judicieux de prendre en compte les variations de $J(v^k)$ afin de décider si la méthode itérative a bien convergé. Dans le cas $\alpha \ll 1$, $J(v^k)$ varie très peu même si v^k est encore relativement loin du minimiseur. Dans ce cas, il est plus judicieux de prendre en compte les variations (normalisées) de v^k dans le critère de convergence. □

L'algorithme de gradient à pas fixe peut s'interpréter comme un algorithme de point fixe pour la fonctionnelle

$$J_\lambda : V \ni v \longmapsto v - \lambda \nabla J(v) \in V. \tag{3.28}$$

Il est clair que trouver un point fixe de la fonctionnelle J_λ revient à trouver un point critique de la fonctionnelle J. Cette observation va nous permettre d'analyser la convergence de l'algorithme de gradient à pas fixe dans le cas particulier où la fonctionnelle J jouit de bonnes propriétés (notamment forte convexité).

Lemme 3.16 (Convergence)**.** *On suppose que la fonctionnelle J est fortement convexe de paramètre α et que l'application $\nabla J : V \to V$ est Lipschitzienne dans V, à savoir,*

$$\exists L > 0, \quad \forall (v, w) \in V \times V, \quad \|\nabla J(w) - \nabla J(v)\|_V \leqslant L \|w - v\|_V. \tag{3.29}$$

Alors, sous l'hypothèse

$$0 < \lambda < \frac{2\alpha}{L^2}, \tag{3.30}$$

la suite $(v^k)_{k\in\mathbb{N}}$ engendrée par l'algorithme de gradient à pas fixe converge, pour tout $v^0 \in V$, vers l'unique solution u du problème (3.1)*. Plus précisément, il existe $\rho \in]0, 1[$ tel que pour tout $k \geqslant 0$,*

$$\|u - v^{k+1}\|_V \leqslant \rho \|u - v^k\|_V. \tag{3.31}$$

On dit que la convergence est d'ordre un*, en référence à l'exposant du facteur $\|u - v^k\|_V$ dans le membre de droite.*

Preuve. Commençons par montrer que l'application J_λ est contractante sous l'hypothèse (3.30). Soit $(v, w) \in V \times V$. On observe que

$$\begin{aligned} \|J_\lambda(w) - J_\lambda(v)\|_V^2 &= \|(w - v) - \lambda(\nabla J(w) - \nabla J(v))\|_V^2 \\ &= \|w - v\|_V^2 - 2\lambda(\nabla J(w) - \nabla J(v), w - v)_V + \lambda^2 \|\nabla J(w) - \nabla J(v)\|_V^2 \\ &\leqslant (1 - 2\lambda\alpha + \lambda^2 L^2)\|w - v\|_V^2, \end{aligned}$$

grâce à la forte convexité de J et au caractère Lipschitzien de ∇J. L'hypothèse (3.30) implique que $(1 - 2\lambda\alpha + \lambda^2 L^2) \in]0, 1[$. Par suite, pour tout $(v, w) \in V \times V$,

$$\|J_\lambda(w) - J_\lambda(v)\|_V \leqslant \rho \|w - v\|_V,$$

avec $\rho = (1 - 2\lambda\alpha + \lambda^2 L^2)^{1/2} \in]0, 1[$. L'inégalité (3.31) est une conséquence immédiate du fait que J_λ est contractante, l'algorithme de gradient à pas fixe n'étant rien d'autre qu'une méthode itérative de point fixe pour la fonctionnelle J_λ. Il suffit en effet d'observer que

$$u - v^{k+1} = J_\lambda(u) - J_\lambda(v^k),$$

et d'appliquer la propriété de contraction de J_λ avec $w = u$ et $v = v^k$. Il est alors classique de montrer par récurrence sur $k \geqslant 0$ que $\|u - v^k\|_V \leqslant \rho^k \|u - v^0\|_V$; d'où la convergence. □

Remarque (Vitesse de convergence)**.** On vérifie facilement que la valeur optimale (minimale) de ρ est obtenue avec le choix $\lambda_{\text{opt}} = \frac{\alpha}{L^2}$, ce qui donne $\rho_{\text{opt}} = 1 - (\frac{\alpha}{L})^2$. Lorsque $\alpha \ll L$ (ce qui arrive souvent en pratique), cette valeur (et donc *a fortiori* toute valeur de ρ obtenue par un choix de λ vérifiant (3.30)) sera très proche de 1, ce qui implique que la convergence de l'algorithme de gradient à pas fixe sera lente. □

Remarque (Extensions)**.** Les algorithmes de gradient (donc d'ordre un) ne sont pas toujours les plus efficaces. Une des difficultés dans la conception des algorithmes d'optimisation est de trouver un compromis adéquat entre d'une part les *capacités d'exploration* de l'algorithme (v^{k+1} doit pouvoir être loin de v^k, notamment afin de pouvoir sortir des puits liés à des minima locaux) et d'autre part la *vitesse de convergence asymptotique* (si v^k est proche du minimiseur, l'algorithme

doit converger très vite). On peut d'une part considérer un algorithme *d'ordre deux*, le plus connu étant la méthode de Newton. Celle-ci nécessite le calcul de (l'inverse de) la matrice hessienne de la fonctionnelle J ce qui peut s'avérer relativement coûteux. On peut alors avoir recours à des méthodes de Newton inexactes où cette matrice hessienne (ou son inverse) n'est évaluée que de façon approchée. On retiendra, d'une manière générale, que les algorithmes d'ordre deux convergent plus vite que les algorithmes d'ordre un, mais que leur mise en œuvre est souvent plus onéreuse et que leur convergence est parfois erratique. On peut d'autre part considérer des algorithmes *d'ordre zéro*, c'est-à-dire qui ne nécessitent que l'évaluation de la fonctionnelle J. De tels algorithmes utilisent souvent des techniques stochastiques en introduisant de l'aléa dans le passage de v^k à v^{k+1}. Dans cette famille figurent l'algorithme du recuit simulé et les algorithmes génétiques (ou évolutionnaires). De tels algorithmes ont en général des vitesses de convergence asymptotique assez lentes, mais en revanche leur capacité d'exploration est importante. □

3.3.3 Systèmes linéaires et fonctionnelles quadratiques

On considère le système linéaire $Au = b$ où A est une matrice d'ordre N et b un vecteur de $\mathbb{R}^N$. Dans cette section, on suppose que la matrice A est *symétrique définie positive*. L'idée que nous allons développer consiste à voir la solution u du système linéaire comme le minimiseur d'une fonctionnelle quadratique, ce qui nous permettra de concevoir des méthodes itératives afin d'approcher cette solution.

Proposition 3.17 (Équivalence)**.** *Soit $A \in \mathbb{R}^{N,N}$ une matrice symétrique définie positive et $b \in \mathbb{R}^N$. On considère la* fonctionnelle quadratique

$$J : \mathbb{R}^N \ni v \longmapsto \frac{1}{2}(Av, v)_{\mathbb{R}^N} - (b, v)_{\mathbb{R}^N} \in \mathbb{R}. \tag{3.32}$$

On a l'équivalence suivante :

$$\Big(J(u) = \inf_{v \in \mathbb{R}^N} J(v)\Big) \Longleftrightarrow (Au = b). \tag{3.33}$$

Preuve. En procédant comme à la section 3.2.3, on montre que J est différentiable dans $\mathbb{R}^N$ avec

$$\forall v \in \mathbb{R}^N, \qquad \nabla J(v) = Av - b. \tag{3.34}$$

De par le critère (iii) de la proposition 3.10, la fonctionnelle J est fortement convexe dans $\mathbb{R}^N$, le paramètre α étant donné par la plus petite valeur propre de la matrice A. Du corollaire 3.6, nous déduisons que la fonctionnelle J admet un et un seul minimiseur dans $\mathbb{R}^N$ et de par le corollaire 3.11 nous pouvons le caractériser comme l'unique point critique de J, c'est-à-dire comme la solution du système linéaire $Au = b$. □

Pour tout vecteur $v \in \mathbb{R}^N$, la vecteur $r(v) = b - Av \in \mathbb{R}^N$ s'appelle le *résidu* de v. La formule (3.34) se récrit sous la forme $\nabla J(v) = -r(v)$ pour tout $v \in \mathbb{R}^N$. En appliquant l'algorithme de gradient à pas fixe à la fonctionnelle quadratique J, il vient pour tout $k \geqslant 0$,

$$d^k = -\nabla J(v^k) = r(v^k). \tag{3.35}$$

Le résidu $r(v^k)$ sert donc de direction de descente à partir de v^k.

Une notion importante lors de la résolution numérique du système linéaire $Au = b$ est celle de conditionnement de la matrice A.

Définition 3.18 (Conditionnement)**.** *Soit A une matrice symétrique définie positive. On définit le* conditionnement *de A, que l'on note $\kappa(A)$, comme le rapport entre sa plus grande et sa plus petite valeur propre. On dit que la matrice A est mal conditionnée lorsque $\kappa(A) \gg 1$.*

Lorsque la matrice A est mal conditionnée, la convergence de l'algorithme de gradient à pas fixe est très lente. En effet, le facteur ρ obtenu au Lemme 3.16 est au mieux de l'ordre de $1 - \kappa(A)^{-1}$

puisque, pour une fonctionnelle quadratique, α correspond à la plus petite valeur propre de A et L à sa plus grande. On a donc $\rho \approx 1^-$ lorsque $\kappa(A) \gg 1$. L'origine du problème peut se comprendre dans le cas très simple de la dimension deux ($V = \mathbb{R}^2$). La figure 3.1 présente un exemple d'isovaleurs d'une fonctionnelle quadratique J dont la matrice associée est mal conditionnée. Les isovaleurs de J sont des ellipses dont les demi-axes sont reliés aux deux valeurs propres de A ; ces ellipses sont donc très aplaties si A est mal conditionnée. Dans ces conditions, on observe qu'en se déplaçant à partir de v selon la direction de descente $-\nabla J(v) = r(v)$, la fonctionnelle J ne décroît que dans un très petit voisinage de v.

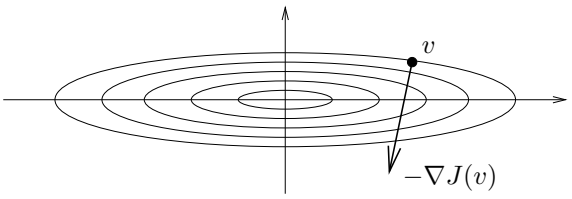


FIGURE 3.1 – Isovaleurs d'une fonctionnelle quadratique dont la matrice associée est mal conditionnée.

Remarque (Approximation par éléments finis)**.** Lorsque la matrice A est issue de l'approximation d'un problème elliptique par éléments finis (cf. chapitre 4), on montre sous des hypothèses assez générales que

$$\kappa(A) \approx h^{-2},$$

où h est le pas du maillage utilisé. Ainsi, l'utilisation de maillages fins, souhaitable pour la précision du calcul éléments finis, s'accompagne d'une dégradation du conditionnement de la matrice A. □

3.3.4 L'algorithme du gradient conjugué (complément)

L'algorithme du gradient conjugué, découvert en 1952 par Hestenes et Stiefel, est une méthode très efficace pour la minimisation de fonctionnelles quadratiques de la forme (3.32) avec une matrice A symétrique définie positive. Dans cet algorithme, à la k-ième itération, la fonctionnelle J est minimisée sur un sous-espace affine de dimension k. Le point remarquable (exploitant judicieusement la symétrie de A) est qu'il est possible d'obtenir cette propriété grâce à une modification très simple de l'algorithme de gradient à pas fixe, en modifiant la direction de descente $r(v^k)$ par une combinaison linéaire entre celle-ci et $r(v^{k-1})$. L'algorithme du gradient conjugué est le suivant :

1. Initialisation : choisir $v^0 \in V$, calculer $r^0 = r(v^0) = b - Av^0$, poser $k = 0$, $\beta^0 = 0$, $p^{-1} = 0$ et fixer le seuil de convergence $\varepsilon > 0$;
2. Boucle en k : pour $k \geqslant 0$, effectuer les opérations suivantes :
   (2.a) choisir pour direction de descente
   $$p^k = r^k + \beta^k p^{k-1}.$$
   Si $p^k = 0$, les itérations s'arrêtent car (on peut montrer que) v^k est point critique de J : l'algorithme a convergé.
   (2.b) déterminer les vecteurs v^{k+1} et r^{k+1} selon les formules
   $$\begin{aligned} \alpha^k &= \frac{(r^k, r^k)_{\mathbb{R}^N}}{(Ap^k, p^k)_{\mathbb{R}^N}}, \\ v^{k+1} &= v^k + \alpha^k p^k, \\ r^{k+1} &= r^k - \alpha^k A p^k. \end{aligned}$$
   Il est facile de voir par récurrence que $r^k = r(v^k)$, c'est-à-dire que r^k est bien le résidu de v^k. De plus, on peut montrer (cf. exercice 10) que le choix ci-dessus pour α^k est tel que
   $$J(v^{k+1}) = \inf_{t \in \mathbb{R}} J(v^k + tp^k).$$

(2.c) tant que $\|v^{k+1} - v^k\|_V > \varepsilon$ (ou tout autre critère de convergence), poser

$$\beta^{k+1} = \frac{(r^{k+1}, r^{k+1})_{\mathbb{R}^N}}{(r^k, r^k)_{\mathbb{R}^N}},$$

puis $k \leftarrow k+1$ et revenir à l'étape (2.a). On peut montrer (cf. exercice 10) que ce choix pour β^k assure que $(p^{k+1}, r^{k+1})_{\mathbb{R}^N} = 0$.

Il est commode d'introduire l'espace de Krylov défini pour $k \geqslant 1$ par

$$K_k = \text{vect}\{r^0, Ar^0, \ldots, A^{k-1}r^0\}, \tag{3.36}$$

Il est assez facile de voir par récurrence que

$$K_k = \text{vect}\{p^0, \ldots, p^{k-1}\}, \tag{3.37}$$

et que $v^k \in v^0 + K_k$. Le résultat suivant, que nous admettons et qui synthétise les propriétés de l'algorithme du gradient conjugué, est le fruit de manipulations algébriques reposant sur d'habiles récurrences (utilisant notamment la symétrie de la matrice A).

Lemme 3.19 (Propriétés)**.** *On suppose que la matrice A est symétrique définie positive. Alors, pour tout $k \geqslant 1$ et tant que l'algorithme du gradient conjugué n'a pas convergé,*

(i) *l'espace de Krylov K_k est de dimension k, la famille $\{p^0, \ldots, p^{k-1}\}$ en constitue une base A-orthogonale et la famille $\{r^0, \ldots, r^{k-1}\}$ une base orthogonale, c'est-à-dire pour tout $0 \leqslant m \leqslant k-1$ et $0 \leqslant n \leqslant k-1$ avec $m \neq n$,*

$$(Ap^m, p^n)_{\mathbb{R}^N} = 0 \qquad \text{et} \qquad (r^m, r^n)_{\mathbb{R}^N} = 0; \tag{3.38}$$

(ii) *v^k réalise le minimum de la fonctionnelle quadratique J dans l'espace affine $v^0 + K_k$.*

Une conséquence du lemme 3.19 est que l'algorithme du gradient conjugué converge en au plus N itérations. Toutefois, ce résultat n'est pas directement exploité en pratique car le numéricien a l'ambition de résoudre des systèmes linéaires de grande taille et espère donc atteindre le seuil de convergence qu'il s'est fixé en un nombre d'itérations bien inférieur à N. Cela nous amène à considérer la vitesse de convergence de l'algorithme. Nous admettons le résultat suivant.

Proposition 3.20 (Convergence)**.** *On pose $\|v\|_A = (Av, v)_{\mathbb{R}^N}^{1/2}$ pour tout $v \in \mathbb{R}^N$. On a*

$$\|u - v^k\|_A \leqslant 2\left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}\right)^k \|u - v^0\|_A, \tag{3.39}$$

où $\kappa(A)$ désigne le conditionnement de la matrice A.

Lorsque la matrice A est mal conditionnée, on a $\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1} \approx 1 - 2\kappa(A)^{-1/2} \approx 1^-$. Un remède efficace au mauvais conditionnement de la matrice A est l'utilisation d'un *préconditionneur*. Le principe consiste à se donner une matrice symétrique définie positive P et à appliquer l'algorithme du gradient conjugué au système préconditionné

$$\widetilde{A}\widetilde{u} = \widetilde{b},$$

avec

$$\widetilde{A} = P^{-1/2}AP^{-1/2} \qquad \text{et} \qquad \widetilde{b} = P^{-1/2}b.$$

La solution recherchée u est alors donnée par $u = P^{-1/2}\widetilde{u}$. Tout l'art du numéricien consiste à effectuer un choix judicieux de la matrice de préconditionnement P. Elle doit d'une part être relativement facile à inverser et d'autre part approcher convenablement la matrice A afin que le nombre de conditionnement de la nouvelle matrice $\widetilde{A}$ soit suffisamment petit (on notera que dans le cas limite où $P = A$, $\widetilde{A}$ est la matrice identité dont le nombre de conditionnement est égal à

un). On peut concevoir soit des préconditionneurs dédiés qui cherchent à prendre en compte une physique ou une discrétisation simplifiée (un exemple important est la méthode multigrille) soit des préconditionneurs algébriques qui ne prennent pas en compte directement l'origine physique et numérique de la matrice A (un exemple important est la décomposition LU ou de Choleski incomplète).

Remarque (Coût d'une itération)**.** Le coût d'une itération de l'algorithme du gradient conjugué peut s'évaluer en termes d'opérations élémentaires (additions, multiplications, etc.) effectuées. Dans la limite (pratique) où $N \gg 1$, ce coût est dominé par le coût du produit matrice-vecteur qui requiert de l'ordre de N^2 opérations (un seul produit matrice-vecteur est effectué par itération pour le calcul du vecteur Ap^k). Par ailleurs, une certain nombre de produits scalaires doivent également être effectués ainsi que des sommes de vecteurs. Le coût de telles opérations est proportionnel à N ; il est donc négligeable. On notera au passage qu'un produit matrice-vecteur est d'autant plus efficace que la matrice A est *creuse*, ce qui est le cas des matrices de rigidité issues de l'approximation de problèmes elliptiques par des méthodes d'éléments finis (cf. définition 4.25 au chapitre 4). □

Remarque (Extensions)**.** La littérature spécialisée sur l'algorithme du gradient conjugué est abondante, le but étant d'étendre le champ d'application de cet algorithme au-delà du cadre restreint de la minimisation de fonctionnelles quadratiques avec matrice symétrique définie positive. Dans le cadre de la résolution des systèmes linéaires, il existe des extensions de l'algorithme du gradient conjugué au cas de matrices non-symétriques. Ces extensions peuvent être classifiées en deux grandes familles.

- D'une part, les algorithmes qui jouissent d'une propriété d'optimalité sur un espace affine de dimension croissante avec les itérations. Dans ce cas, il est nécessaire de garder en mémoire une base complète de l'espace de Krylov, d'où un coût de calcul sensiblement plus élevé. Un exemple d'algorithme relevant de cette classe est GMRes (de l'acronyme anglais *Generalized Minimal Residual*).
- D'autre part, des algorithmes qui ne conservent pas en mémoire toutes les directions de descente parcourues aux itérations précédentes. Leur coût par itération est moindre que pour GMRes, mais ces algorithmes peuvent ne pas converger. Un exemple d'algorithme relevant de cette classe est BiCGStab (de l'acronyme anglais *Bi-Conjugate Gradient Stabilized*).

On notera enfin que dans le cadre plus général de l'optimisation, il existe des extensions non-linéaires de l'algorithme du gradient conjugué. □

3.4 Optimisation sous contraintes

Cette section est consacrée à l'étude de l'existence et l'unicité d'un minimiseur pour le problème d'optimisation sous contraintes (3.3) (en dimension finie ou infinie) et à sa caractérisation.

3.4.1 Existence et unicité

Commençons par le cas de la dimension finie.

Théorème 3.21 (Existence, dimension finie)**.** *On suppose que l'espace vectoriel V est de dimension* finie *et que l'ensemble K est* fermé *dans V. On suppose que la fonctionnelle J est* continue *et* coercive *dans K (si K est borné, il suffit de supposer que J est continue dans K). Alors, J admet au moins un minimiseur global dans K.*

La preuve, qui suit les mêmes arguments que dans le cas sans contrainte, est laissée au lecteur. Pour se convaincre de l'importance de l'hypothèse K fermé, on pourra se placer dans $V = \mathbb{R}$ avec $K =]0, 1[$ et $J(v) = v$; on a $\inf_{v \in K} J(v) = 0$ mais ce minimum n'est pas atteint dans K.

Afin d'obtenir des conditions suffisantes pour l'existence d'un minimiseur en dimension *infinie*, nous faisons non seulement une hypothèse de convexité sur la fonctionnelle J comme dans le cas

sans contrainte, mais également une hypothèse de convexité sur l'ensemble K. Rappelons à toutes fins utiles que l'ensemble K est dit *convexe* si

$$\forall (x,y) \in K \times K, \quad \forall \theta \in [0,1], \quad \theta x + (1-\theta)y \in K. \tag{3.40}$$

Nous admettons le résultat suivant.

Théorème 3.22 (Existence, dimension infinie, K convexe)**.** *Soit V un* espace de Hilbert*. On suppose que l'ensemble K est* fermé *dans V et* convexe*. On suppose que la fonctionnelle J est* continue, coercive et convexe *dans K. Alors, J admet au moins un minimiseur global dans K.*

Passons maintenant à des conditions suffisantes afin d'assurer l'unicité du minimiseur.

Théorème 3.23 (Unicité)**.** *Soient V un espace vectoriel, K un ensemble* convexe *et J une fonctionnelle* strictement convexe *dans K. Alors, J admet au plus un minimiseur dans K.*

Preuve. On procède comme dans la preuve du théorème 3.5. Si u_1 et u_2 sont deux minimiseurs distincts dans K, on observe que $\frac{1}{2}(u_1+u_2)$ est dans K car cet ensemble est convexe pour déduire de la stricte convexité de J que

$$J\left(\frac{u_1+u_2}{2}\right) < \frac{1}{2}J(u_1) + \frac{1}{2}J(u_2) = \inf_{v\in V} J(v),$$

ce qui fournit une contradiction. □

Enfin, un résultat d'existence et d'unicité peut être énoncé sous hypothèses de forte convexité et continuité de la fonctionnelle J et de convexité de l'ensemble K. On observera qu'en l'absence d'hypothèse de convexité sur K, le minimiseur n'est pas forcément unique, même si la fonctionnelle J est fortement convexe (se placer par exemple dans $V=\mathbb{R}$ avec $K=[-2,-1]\cup[1,2]$ et $J(v)=v^2$; J admet deux minimiseurs dans K qui sont ± 1).

Corollaire 3.24 (Existence et unicité)**.** *Soient V un* espace de Hilbert*, K un ensemble* convexe et fermé *de V et J une fonctionnelle* fortement convexe *dans K. Si l'espace V est de dimension infinie, on suppose de plus que J est* continue*. Alors, J admet un et un seul minimiseur global dans K.*

3.4.2 Caractérisation

Par la suite, nous supposons J différentiable dans V.

Théorème 3.25. *(Condition d'Euler–Lagrange) Soient V un* espace de Hilbert*, K un ensemble* convexe fermé *de V et J une fonctionnelle de V dans $\mathbb{R}$. On suppose que la fonctionnelle J admet un minimum local en $u \in K$ et que J est* différentiable *en u. Alors,*

$$\forall v \in K, \qquad \langle J'(u), v-u\rangle_{V',V} \geqslant 0. \tag{3.41}$$

De plus, si u est un point intérieur à K, alors u est un point critique de J, c'est-à-dire qu'on retrouve la condition d'Euler

$$J'(u) = 0 \ (\in V'). \tag{3.42}$$

Réciproquement, si u vérifie (3.41) *ou* (3.42) *et que J est* convexe*, alors u est un minimiseur global de J dans K.*

Preuve. Par hypothèse, pour tout $v \in K$ et pour tout $t \in [0,1]$ suffisamment petit,

$$J((1-t)u+tv) \geqslant J(u),$$

car $(1-t)u+tv \in K$ par convexité de K. Or,

$$J((1-t)u+tv) = J(u) + t\langle J'(u), v-u\rangle_{V',V} + \mathrm{o}(t),$$

car J est différentiable en u. En passant à la limite $t \to 0$ avec $t>0$, on obtient (3.41). Dans le cas où u est un point intérieur à K, il suffit de remarquer que $v-u$ décrit l'ensemble des directions dans V afin d'obtenir (3.42). Enfin, l'assertion réciproque résulte du point (ii) de la proposition 3.9, qui reste valable en se restreignant à une fonctionnelle convexe dans un ensemble convexe K. □

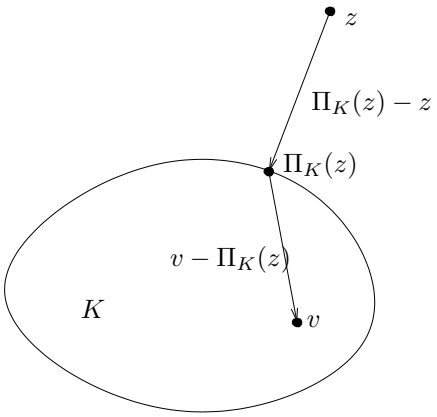


FIGURE 3.2 – Projection orthogonale sur un convexe fermé.

Une application utile de la condition d'Euler–Lagrange est la projection orthogonale sur un ensemble convexe fermé K dans un espace de Hilbert V. On désigne par $\Pi_K : V \to K$ l'application qui à $z \in V$ associe l'élément $\Pi_K(z) \in K$ tel que

$$\|z - \Pi_K(z)\|_V = \inf_{v \in K} \|z - v\|_V. \tag{3.43}$$

L'analyse mathématique de ce problème est très simple. Pour tout $z \in V$ fixé, la fonctionnelle $J_z : V \ni v \mapsto J_z(v) = \|z - v\|_V^2$ est fortement convexe si bien que le problème (3.43) admet une et une seule solution $\Pi_K(z) \in K$. On dit que $\Pi_K(z)$ est la *projection orthogonale* de z sur K. De plus, comme $\nabla J_z(v) = 2(v - z)$ (noter l'utilisation du gradient même en dimension infinie), la condition d'Euler–Lagrange (3.41) donne

$$\forall v \in K, \qquad (\Pi_K(z) - z, v - \Pi_K(z))_V \geqslant 0. \tag{3.44}$$

L'interprétation géométrique de cette condition est que l'angle entre les vecteurs $\Pi_K(z) - z$ et $v - \Pi_K(z)$ est aigu pour tout $v \in K$ (voir figure 3.2). Dans le cas particulier où K est un *pavé* de $\mathbb{R}^N$ de la forme

$$K = \prod_{i=1}^{N} [a_i, b_i], \tag{3.45}$$

pour des réels a_i et b_i, on pourra vérifier en exercice à partir de la caractérisation (3.44), que pour un vecteur $z \in \mathbb{R}^N$ de composantes $(z_1, \ldots, z_N)$, le vecteur $\Pi_K(z) \in \mathbb{R}^N$ a pour composantes

$$\Pi_K(z)_i = \max(a_i, \min(z_i, b_i)), \qquad \forall 1 \leqslant i \leqslant N. \tag{3.46}$$

Ces formules s'étendent au cas non-borné avec les conventions usuelles : pour $x \in \mathbb{R}$, $\min(x, +\infty) = x$ et $\max(-\infty, x) = x$.

Passons maintenant au cas où l'ensemble des états admissibles K n'est pas convexe.

Définition 3.26 (Directions admissibles)**.** *Soit $u \in K$. On définit $K(u)$ comme l'ensemble des vecteurs de V qui sont tangents à une courbe de K passant par u. On a donc $z \in K(u)$ si et seulement si il existe $t_0 > 0$ et une application $\varphi : [0, t_0] \ni t \mapsto \varphi(t) \in K$ telle que $\varphi(0) = u$ et $\varphi'(0) = z$ au sens où*

$$\varphi(t) - \varphi(0) = tz + \mathrm{o}(t).$$

On dit que $K(u)$ est le cône des directions admissibles *en v.*

La figure 3.3 illustre cette définition. On observera que l'ensemble $K(u)$ est non-vide (car il contient toujours 0) et que $K(u)$ est un cône (si $z \in K(u)$, alors $\lambda z \in K(u)$ pour tout $\lambda \in \mathbb{R}_+$). De plus, si u est intérieur à K, on a $K(u) = V$ car dans ces conditions, l'ensemble K contient une boule de centre u et tous les rayons de cette boule peuvent être utilisés comme courbes de K passant par u. Enfin, on peut montrer que l'ensemble $K(u)$ est fermé dans V.

En reprenant la preuve du théorème 3.25, on prouve sans peine le résultat suivant.

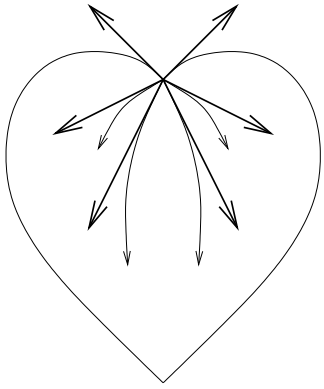


FIGURE 3.3 – Cône des directions admissibles en un point u de K situé sur la frontière ; l'ensemble K a la forme d'un cœur.

Théorème 3.27 (Caractérisation)**.** *Soient V un* espace de Hilbert*, K un sous-ensemble* fermé *de V et J une fonctionnelle de V dans $\mathbb{R}$. On suppose que la fonctionnelle J admet un minimum local en $u \in K$ et que J est* différentiable *en u. Alors,*

$$\forall z \in K(u), \qquad \langle J'(u), z\rangle_{V',V} \geqslant 0. \tag{3.47}$$

Nous allons maintenant étudier deux cas particuliers où nous pourrons mieux caractériser les directions admissibles et ainsi reformuler (3.47) sous une forme plus facilement utilisable en pratique.

Cas des contraintes égalité

Nous considérons le cas où K est défini de la manière suivante :

$$K = \{v \in V;\ \Phi(v) = 0\}, \tag{3.48}$$

où $\Phi : V \to \mathbb{R}^m$ est une application différentiable, les applications composantes étant notées $\Phi_i : V \to \mathbb{R}$, $1 \leqslant i \leqslant m$. L'application Φ étant différentiable, elle est continue si bien que l'ensemble K est fermé dans V. En revanche, nous ne supposons pas que l'ensemble K est nécessairement convexe, une condition suffisante pour cela étant que l'application Φ soit affine.

Définition 3.28 (Qualification des contraintes)**.** *On dit que les contraintes égalité sont* qualifiées *en un point $u \in K$ si la famille $\{\Phi_i'(u)\}_{1\leqslant i\leqslant m}$ est libre (dans V').*

Théorème 3.29 (Multiplicateurs de Lagrange)**.** *Soit $u \in K$. On suppose que les contraintes égalité sont qualifiées en u. Alors, une* condition nécessaire *pour que u soit un minimiseur local de J dans K est qu'il existe m réels $(p_1, \ldots, p_m)$, appelés* multiplicateurs de Lagrange*, tels que*

$$J'(u) + \sum_{j=1}^{m} p_j \Phi_j'(u) = 0 \ (\in V'). \tag{3.49}$$

Les multiplicateurs de Lagrange, s'ils existent, sont uniques.

Preuve. Nous esquissons la preuve. En utilisant la qualification des contraintes et le théorème des fonctions implicites, on montre que

$$K(u) = \{z \in V;\ \forall 1 \leqslant i \leqslant m,\ \langle \Phi_i'(u), z\rangle_{V',V} = 0\},$$

ce qui fait de $K(u)$ un espace vectoriel (en fait, $K(u)$ est l'espace tangent à la variété K au point u). On peut donc tester avec $\pm z \in K(u)$ si bien que

$$\forall z \in K(u), \qquad \langle J'(u), z\rangle_{V',V} = 0.$$

Pour une forme linéaire $\psi \in V'$, notons $\psi^\perp$ son noyau. La caractérisation de $K(u)$ s'écrit

$$K(u) = \bigcap_{i=1}^{m} (\Phi_i'(u))^\perp,$$

et la condition d'optimalité

$$\bigcap_{i=1}^{m} (\Phi_i'(u))^\perp \subset (J'(u))^\perp.$$

Par suite, grâce à des résultats classiques sur l'opérateur $^\perp$ dans les espaces de Hilbert,

$$J'(u) \in \sum_{i=1}^{m} \Phi_i'(u),$$

ce qui n'est rien d'autre que la condition (3.49). L'unicité des multiplicateurs de Lagrange résulte de la qualification des contraintes. □

Un exemple important est celui où $m = 1$ avec $\Phi(v) = \|v\|_V^2 - 1$ si bien que K est la sphère unité de V. Dans ce cas, il vient $\nabla\Phi(v) = 2v$ (noter l'utilisation du gradient même en dimension infinie). La contrainte est qualifiée : sinon, $\nabla\Phi(v) = 0$, v serait donc nul ce qui est impossible puisque $\|v\|_V = 1$. Enfin, le théorème 3.29 donne

$$\nabla J(v) + 2pv = 0 \ (\in V).$$

Ainsi, le réel $-2p$ s'interprète comme une valeur propre de l'opérateur $\nabla J : V \to V$.

Remarque (Interprétation géométrique)**.** Considérons le cas simple où $V = \mathbb{R}^2$ et $m = 1$ (si bien que $\Phi : \mathbb{R}^2 \to \mathbb{R}$). Traçons dans le repère cartésien de coordonnées (x, y) la courbe d'équation $\Phi(x, y) = 0$ représentant l'ensemble des états admissibles K et les courbes de niveau de J, c'est-à-dire les courbes d'équation $J(x, y) = \lambda$ pour $\lambda \in \mathbb{R}$ (voir la figure 3.4 où les courbes de niveau de J sont représentées par des ellipses). En parcourant la courbe d'équation $\Phi(x, y) = 0$, on s'aperçoit que la valeur minimale de J est atteinte lorsque cette courbe est tangente à une courbe de niveau de J. On utilise alors le fait que $\nabla\Phi(x, y)$ est orthogonal au point (x, y) à la courbe d'équation $\Phi(x, y) = 0$. De même, $\nabla J(x, y)$ est orthogonal aux courbes de niveau de J. Ces deux vecteurs sont donc colinéaires au point $u \in K$ où J atteint son minimum. □

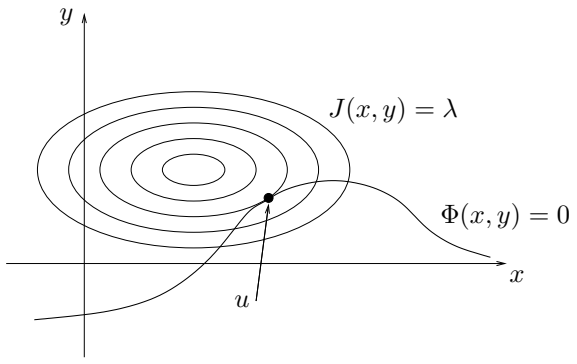


FIGURE 3.4 – Interprétation géométrique des multiplicateurs de Lagrange.

Remarque (Interprétation marginaliste des multiplicateurs de Lagrange)**.** Plaçons-nous dans le cas $m = 1$ et avec V de dimension finie pour simplifier. En remplaçant 0 par ϵ dans la contrainte

$\Phi(v) = 0$, l'ensemble K devient $K_\epsilon = \{v \in V;\ \Phi(v) = \epsilon\}$. Notons u_ϵ un minimiseur (en supposant qu'il existe) de J dans K_ϵ et posons $J_\epsilon = J(u_\epsilon)$. Alors,

$$\frac{d}{d\epsilon} J_\epsilon = \left(\nabla J(u_\epsilon), \frac{d}{d\epsilon} u_\epsilon \right)_V .$$

Par ailleurs, en dérivant la relation $\Phi(u_\epsilon) = \epsilon$ par rapport à ϵ, il vient

$$(\nabla \Phi(u_\epsilon), \frac{d}{d\epsilon} u_\epsilon)_V = 1.$$

Par conséquent, en utilisant le fait que $\nabla J(u_\epsilon) + p_\epsilon \nabla \Phi(u_\epsilon) = 0$ où p_ϵ est le multiplicateur de Lagrange pour le problème de minimisation dans K_ϵ, il vient

$$\frac{d}{d\epsilon} J_\epsilon = -p_\epsilon,$$

et en supposant que $\lim_{\epsilon \to 0} p_\epsilon = p$, on en déduit que

$$\frac{d}{d\epsilon} J_\epsilon \bigg|_{\epsilon=0} = -p.$$

En d'autre termes, le multiplicateur de Lagrange p s'interprète formellement comme la sensibilité (négative) de la fonction de coût à une variation infinitésimale de la contrainte Φ. □

Cas des contraintes inégalité

Nous considérons maintenant le cas particulier où

$$K = \{v \in V;\ \Phi(v) \leqslant 0\}, \tag{3.50}$$

l'application $\Phi : V \to \mathbb{R}^m$ étant à nouveau supposée différentiable (K est donc fermé dans V).

Définition 3.30 (Contrainte active)**.** *On dit que la contrainte Φ_i est* active *(ou* saturée*) en un point $u \in K$ si $\Phi_i(u) = 0$. On note*

$$\mathcal{A}(u) = \{i \in \{1, \dots, m\};\ \Phi_i(u) = 0\}. \tag{3.51}$$

Par exemple, dans $V = \mathbb{R}^2$ avec $K = \{(x, y) \in V; x \leqslant 0, y \leqslant 0\}$, il y a deux contraintes inégalité, à savoir $\Phi_1(x, y) := x \leqslant 0$ et $\Phi_2(x, y) := y \leqslant 0$. Au point $(0, 0)$, les deux contraintes sont actives, en tout point $(x, 0)$ avec $x < 0$ seule la deuxième contrainte est active, en tout point $(0, y)$ avec $y < 0$ seule la première contrainte est active, et en tout point (x, y) avec $x < 0$ et $y < 0$ aucune contrainte n'est active.

Dans le cas des contraintes inégalité, l'identification des directions admissibles en un point $u \in K$ est plus délicate qu'auparavant. Nous allons nous restreindre à des situations relativement simples par le biais d'une hypothèse de qualification des contraintes. L'intuition derrière cette hypothèse est de permettre d'effectuer des petites variations autour du point $u \in K$ afin de tester son optimalité.

Définition 3.31 (Qualification des contraintes)**.** *On dit que les contraintes inégalité sont* qualifiées *en un point $u \in K$ si toutes les applications composantes sont* affines *ou si la famille des différentielles des contraintes actives en u, $\{\Phi_i'(u)\}_{i \in \mathcal{A}(u)}$, est libre (dans V').*

Nous admettons le résultat suivant.

Théorème 3.32 (Multiplicateurs de Lagrange)**.** *Soit $u \in K$. On suppose que les contraintes inégalité sont qualifiées en u. Alors, une* condition nécessaire *pour que u soit un minimiseur local*

de J dans K est qu'il existe un vecteur $p \in \mathbb{R}^m_+$ de composantes positives $(p_1, \dots, p_m)$ *appelées* multiplicateurs de Lagrange*, tel que*

$$J'(u) + \sum_{j=1}^{m} p_j \Phi'_j(u) = 0 \ (\in V'), \tag{3.52}$$

$$p \cdot \Phi(u) = 0 \ (\in \mathbb{R}). \tag{3.53}$$

Ici et par la suite, le symbole $\cdot$ fait référence au produit scalaire usuel dans $\mathbb{R}^m$. Dans le cas où la qualification des contraintes résulte de la liberté de la famille $\{\Phi'_i(u)\}_{i\in\mathcal{A}(u)}$, les multiplicateurs de Lagrange, s'ils existent, sont uniques.

La condition (3.53) est appelée *condition des écarts complémentaires* ou *relations d'exclusion*. Puisque toutes les composantes p_i sont positives ($p \in \mathbb{R}^m_+$) et que tous les réels $\Phi_i(u)$ sont négatifs ($u \in K$), on observe que tous les termes dans le produit scalaire $p \cdot \Phi(u)$ ont le même signe. Chacun de ces termes doit donc être nul d'après (3.53). Cette condition s'écrit donc de façon plus explicite sous la forme

$$\forall 1 \leqslant i \leqslant m, \qquad p_i \Phi_i(u) = 0. \tag{3.54}$$

Ainsi, $(p_i > 0) \Longrightarrow (\Phi_i(u) = 0)$ et $(\Phi_i(u) < 0) \Longrightarrow (p_i = 0)$.

3.4.3 Algorithme de gradient (à pas fixe) avec projection

L'algorithme de gradient à pas fixe vu à la section 3.3.2 pour les problèmes d'optimisation numérique sans contrainte s'étend au cas sous contraintes en projetant les itérations sur l'ensemble convexe fermé K des états admissibles. L'algorithme est le suivant :

1. Initialisation : choisir $v^0 \in K$, poser $k = 0$; fixer le pas $\lambda > 0$ et le seuil de convergence $\varepsilon > 0$;
2. Boucle en k : pour $k \geqslant 0$, effectuer les opérations suivantes :
   (2.a) calculer le gradient de la fonctionnelle J en v^k, $\nabla J(v^k)$, et choisir pour direction de descente
   $$d^k = -\nabla J(v^k).$$
   Si $d^k = 0$, les itérations s'arrêtent car v^k est point critique de J : l'algorithme a convergé.
   (2.b) déterminer v^{k+1} selon la formule
   $$v^{k+1} = \Pi_K(v^k + \lambda d^k).$$
   (2.c) tant que $\|v^{k+1} - v^k\|_V > \varepsilon$ (par exemple), poser $k \leftarrow k + 1$ et revenir à l'étape (2.a).

L'algorithme de gradient à pas fixe avec projection est un algorithme de point fixe pour la fonctionnelle

$$\widehat{J}_\lambda : V \ni v \longmapsto \Pi_K(v - \lambda \nabla J(v)) \in V. \tag{3.55}$$

Un tel point fixe vérifie

$$u = \Pi_K(u - \lambda \nabla J(u)).$$

Par suite, $u \in K$ et en utilisant (3.44) avec $z = u - \lambda \nabla J(u)$, $\Pi_K(z) = u$ et $y = v$ arbitraire dans K, on obtient

$$(u - (u - \lambda \nabla J(u)), v - u)_V = \lambda (\nabla J(u), v - u))_V \geqslant 0.$$

Comme $\lambda > 0$, on retrouve bien la condition d'Euler–Lagrange (3.41).

Proposition 3.33 (Convergence)**.** *On suppose que la fonctionnelle J est fortement convexe de paramètre α et que l'application $\nabla J : V \to V$ est Lipschitzienne dans V de constante L. Alors, sous l'hypothèse* (3.30)*, la suite $(v^k)_{k\in\mathbb{N}}$ engendrée par l'algorithme de gradient à pas fixe avec projection converge pour tout $v^0 \in V$ à l'ordre un vers l'unique solution u du problème d'optimisation sous contraintes* (3.3)*. Plus précisément, il existe $\rho \in]0, 1[$ tel que pour tout $k \geqslant 0$,*

$$\|u - v^{k+1}\|_V \leqslant \rho \|u - v^k\|_V. \tag{3.56}$$

Preuve. Puisque $v^{k+1} = \widehat{J}_\lambda(v^k)$, nous allons montrer la convergence de la suite $(v^k)_{k\in\mathbb{N}}$ en montrant que l'application $\widehat{J}_\lambda$ est contractante. Montrons tout d'abord la propriété suivante :

$$\forall (v,w) \in V \times V, \qquad \|\Pi_K(v) - \Pi_K(w)\|_V \leqslant \|v - w\|_V. \tag{3.57}$$

Posons $\delta = \Pi_K(v) - \Pi_K(w)$ et observons que

$$\|\delta\|_V^2 = (\Pi_K(v) - v, \delta)_V + (v - w, \delta)_V + (w - \Pi_K(w), \delta)_V.$$

D'après (3.44), le premier et le troisième terme du membre de droite sont négatifs. Par suite,

$$\|\delta\|_V^2 \leqslant (v - w, \delta)_V,$$

d'où nous déduisons (3.57) en utilisant l'inégalité de Cauchy–Schwarz. Nous concluons en observant que $\widehat{J}_\lambda$ est la composée de l'application J_λ définie par (3.28) et de la projection Π_K et que l'application J_λ est contractante sous la condition (3.30) de par le lemme 3.16. □

Ayant prouvé la convergence de l'algorithme de gradient à pas fixe avec projection, il nous reste à en examiner sa mise en œuvre pratique. Hormis, quelques cas particuliers (par exemple, celui où l'ensemble K est un pavé (voir (3.46)) ou un sphère pour la norme de V (ou une norme plus faible)), déterminer la projection $\Pi_K(v)$ d'un point $v \in V$ sur K est une opération de complexité souvent équivalente à la résolution du problème de départ (3.3). L'algorithme de gradient à pas fixe avec projection n'est donc viable en pratique que pour ces cas particuliers auxquels nous nous limiterons. Dans le cas général, on peut avoir recours à des méthodes de dualité, dont l'exemple le plus important est l'algorithme d'Uzawa étudié dans la section suivante en guise de complément.

3.5 Méthodes de dualité (complément)

L'objectif de cette section est de présenter les bases théoriques et la réalisation numérique de méthodes de dualité pour traiter des problèmes d'optimisation sous contraintes.

3.5.1 Lagrangien et point selle

Une façon élégante et utile en pratique de reformuler le problème d'optimisation sous contraintes inégalité consiste à introduire le *Lagrangien* $\mathcal{L}$ défini comme suit :

$$\mathcal{L} : V \times \mathbb{R}_+^m \ni (v,q) \longmapsto J(v) + q \cdot \Phi(v) \in \mathbb{R}. \tag{3.58}$$

Nous considérons des contraintes sous la forme (3.50) (il est également possible d'introduire le Lagrangien, sur $V \times \mathbb{R}^m$, pour traiter m contraintes égalité). Nous supposons que les fonctionnelles J et Φ sont différentiables dans V.

Définition 3.34 (Point selle)**.** *On dit que le couple (u,p) est un* point selle *du Lagrangien $\mathcal{L}$ dans $V \times \mathbb{R}_+^m$ si on a*

$$\forall (v,q) \in V \times \mathbb{R}_+^m, \qquad \mathcal{L}(u,q) \leqslant \mathcal{L}(u,p) \leqslant \mathcal{L}(v,p). \tag{3.59}$$

En d'autres termes,

$$\sup_{q\in\mathbb{R}_+^m} \mathcal{L}(u,q) = \mathcal{L}(u,p) = \inf_{v\in V} \mathcal{L}(v,p). \tag{3.60}$$

La figure 3.5 illustre la notion de point selle. Un point selle satisfait une condition d'optimalité plus générale que (3.60). Celle-ci est précisée dans la proposition suivante.

Proposition 3.35 (Propriété avec point selle)**.** *Soit (u,p) un point selle du Lagrangien $\mathcal{L}$ dans $V \times \mathbb{R}_+^m$. On a*

$$\sup_{q\in\mathbb{R}_+^m} \inf_{v\in V} \mathcal{L}(v,q) = \mathcal{L}(u,p) = \inf_{v\in V} \sup_{q\in\mathbb{R}_+^m} \mathcal{L}(v,q). \tag{3.61}$$

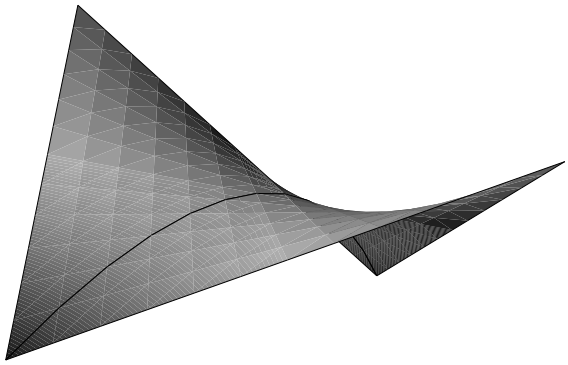


FIGURE 3.5 – Graphe d'un Lagrangien présentant un point selle.

Preuve. Montrons la première égalité. On introduit la fonctionnelle

$$G : \mathbb{R}^m_+ \ni q \longmapsto G(q) := \inf_{v\in V} \mathcal{L}(v,q) \in \mathbb{R} \cup \{-\infty\}. \tag{3.62}$$

Puisque (u,p) est point selle de $\mathcal{L}$, on a pour tout $q \in \mathbb{R}^m_+$,

$$G(q) = \inf_{v\in V} \mathcal{L}(v,q) \leqslant \mathcal{L}(u,q) \leqslant \mathcal{L}(u,p).$$

De plus, $G(p) = \mathcal{L}(u,p)$ toujours par la propriété du point selle. D'où $\sup_{q\in\mathbb{R}^m_+} G(q) = G(p) = \mathcal{L}(u,p)$.
La preuve de la deuxième inégalité est analogue. On introduit la fonctionnelle

$$H : V \ni v \longmapsto H(v) := \sup_{q\in\mathbb{R}^m_+} \mathcal{L}(v,q) \in \mathbb{R} \cup \{+\infty\}, \tag{3.63}$$

et on vérifie que

$$H(v) \geqslant \mathcal{L}(v,p) \geqslant \mathcal{L}(u,p) = H(u),$$

si bien que $\inf_{v\in V} H(v) = H(u) = \mathcal{L}(u,p)$. □

Remarque ($\sup\inf \leqslant \inf\sup$)**.** Soit $(u,p) \in V \times \mathbb{R}^m_+$. On ne suppose pas que (u,p) soit un point selle de $\mathcal{L}$. On constate que

$$\mathcal{L}(u,p) \leqslant \sup_{q\in\mathbb{R}^m_+} \mathcal{L}(u,q),$$

si bien que

$$\inf_{v\in V} \mathcal{L}(v,p) \leqslant \inf_{v\in V} \sup_{q\in\mathbb{R}^m_+} \mathcal{L}(v,q).$$

Par suite,

$$\sup_{q\in\mathbb{R}^m_+} \inf_{v\in V} \mathcal{L}(v,q) \leqslant \inf_{v\in V} \sup_{q\in\mathbb{R}^m_+} \mathcal{L}(v,q). \tag{3.64}$$

L'existence d'un point selle assure donc l'égalité dans (3.64). En l'absence de point selle, il se peut que l'inégalité (3.64) soit stricte. On parle de gap de dualité. □

Voyons maintenant quels sont les liens entre la recherche d'un minimiseur du problème d'optimisation sous contraintes et la recherche d'un point selle du Lagrangien.

Proposition 3.36 (Propriétés d'un point selle)**.** *Si le couple $(u,p) \in V \times \mathbb{R}^m_+$ est un point selle du Lagrangien $\mathcal{L}$ dans $V \times \mathbb{R}^m_+$, alors $u \in K$, u est un minimiseur global de J dans K et le couple (u,p) vérifie la condition des écarts complémentaires $p \cdot \Phi(u) = 0$.*

Preuve. Considérons tout d'abord le fait que $\mathcal{L}(u,p) = \sup_{q\in\mathbb{R}^m_+} \mathcal{L}(u,q)$ si bien que

$$\forall q \in \mathbb{R}^m_+, \qquad q \cdot \Phi(u) \leqslant p \cdot \Phi(u).$$

S'il existait $i \in \{1,\dots,m\}$ tel que $\Phi_i(u) > 0$, on pourrait faire exploser le minorant en faisant tendre $q_i \to +\infty$, ce qui est absurde. On a donc $\Phi(u) \leqslant 0$, c'est-à-dire $u \in K$. De plus, pour tout $i \notin \mathcal{A}(u)$, on a $\Phi_i(u) < 0$ et dans ce cas, nécessairement $p_i = 0$ car si $p_i > 0$, on pourrait prendre $q_j = p_j$ pour $j \neq i$ et $q_i = \frac{1}{2}p_i$ et obtenir $q \cdot \Phi(u) > p \cdot \Phi(u)$. Par suite, la condition des écarts complémentaires $p\cdot\Phi(u) = 0$ est satisfaite. Enfin, en considérant le fait que $\mathcal{L}(u,p) = \inf_{v\in V} \mathcal{L}(v,p)$, il vient pour tout $v \in K$,

$$J(u) = \mathcal{L}(u,p) \leqslant \mathcal{L}(v,p) = J(v) + p \cdot \Phi(v) \leqslant J(v),$$

puisque $p \cdot \Phi(u) = 0$ et $p \cdot \Phi(v) \leqslant 0$. Par suite, u est un minimiseur global de J dans K. □

Lorsque les fonctionnelles J et $\{\Phi\}_{1\leqslant i\leqslant m}$ sont *convexes* (si bien que K est convexe), nous disposons d'un résultat *d'équivalence* entre la résolution du problème d'optimisation sous contraintes et la recherche d'un point selle pour le Lagrangien. Le résultat ci-dessous est connu sous le nom de théorème de Kuhn et Tucker (ou de Karush, Kuhn et Tucker).

Théorème 3.37 (Kuhn–Tucker)**.** *On suppose que les fonctionnelles J et $\{\Phi\}_{1\leqslant i\leqslant m}$ sont* convexes*. Soit $u \in K$. On suppose que les contraintes inégalité sont qualifiées en u. Alors, les assertions suivantes sont équivalentes :*

(i) *u est un minimiseur global de J dans K ;*
(ii) *il existe $p \in \mathbb{R}^m_+$ tel que le couple (u,p) est un point selle du Lagrangien $\mathcal{L}$ dans $V \times \mathbb{R}^m_+$;*
(iii) *il existe $p \in \mathbb{R}^m_+$ tel que les relations* (3.52) *et* (3.53) *sont satisfaites.*

Preuve. L'implication (i) ⇒ (iii) est l'assertion du théorème 3.32. L'implication (ii) ⇒ (i) résulte de la proposition 3.36. Considérons enfin l'implication (iii) ⇒ (ii). Pour tout $q \in \mathbb{R}^m_+$, il est clair que $q \cdot \Phi(u) \leqslant 0 = p \cdot \Phi(u)$ de par la relation des écarts complémentaires (3.53). D'où

$$\mathcal{L}(u,p) = \sup_{q\in\mathbb{R}^m_+} \mathcal{L}(u,q).$$

Par ailleurs, à $p \in \mathbb{R}^m_+$ fixé, la fonctionnelle $J_p : V \ni v \mapsto J(v) + p\cdot\Phi(v)$ est par hypothèse convexe. La relation (3.52) implique que le point u est un point critique de J_p, et donc un minimiseur global. Cela implique que pour tout $v \in V$, $J_p(u) \leqslant J_p(v)$, ou encore

$$\mathcal{L}(u,p) = \inf_{v\in V} \mathcal{L}(v,p),$$

ce qui complète la preuve. □

On vérifie facilement en considérant la fonctionnelle H définie par (3.63) que

$$H(v) = \sup_{q\in\mathbb{R}^m_+} \mathcal{L}(v,q) = \begin{cases} J(v) & \text{si } v \in K, \\ +\infty & \text{sinon.} \end{cases}$$

Ainsi, minimiser la fonctionnelle H dans l'ensemble V revient à chercher le minimiseur u de J dans K. Par ailleurs, maximiser la fonctionnelle G définie par (3.62) fournit un multiplicateur de Lagrange $p \in \mathbb{R}^m_+$. L'équivalence entre les deux approches est exprimée par le biais de la proposition 3.35. Lorsque des méthodes d'optimisation numérique sont conçues à partir de cette équivalence, on parle de *méthodes de dualité*.

3.5.2 Algorithme d'Uzawa

Nous considérons un problème d'optimisation sous contraintes inégalité avec

$$K = \{v \in V;\ \Phi(v) \leqslant 0\}, \tag{3.65}$$

ensemble qui est supposé convexe et fermé (ce qui est le cas, par exemple, si les applications composantes Φ_i, $1 \leqslant i \leqslant m$, sont convexes et continues). Nous supposons également que la fonctionnelle J est *fortement convexe* dans V si bien qu'elle admet un unique minimiseur global u dans K. De plus, de par le théorème 3.37, nous savons qu'il existe un vecteur $p \in \mathbb{R}^m_+$ tel que le couple (u, p) est un point selle du lagrangien $\mathcal{L}$ défini par (3.58). L'idée principale dans l'algorithme d'Uzawa consiste à observer que puisque $\mathcal{L}(u,p) = \sup_{q\in\mathbb{R}^m_+} \mathcal{L}(u,q)$, on a

$$\forall q \in \mathbb{R}^m_+, \qquad q \cdot \Phi(u) \leqslant p \cdot \Phi(u).$$

Par conséquent, pour tout réel $\lambda > 0$, il vient

$$\forall q \in \mathbb{R}^m_+, \qquad (p-q)\cdot(p - \lambda\Phi(u) - p) \leqslant 0,$$

ce qui montre d'après (3.44) que

$$p = \Pi_{\mathbb{R}^m_+}(p + \lambda\Phi(u)). \tag{3.66}$$

Dans l'algorithme d'Uzawa, deux suites sont générées, une suite $(v^k)_{k\in\mathbb{N}}$ d'éléments de V et une suite $(q^k)_{k\in\mathbb{N}}$ d'éléments de $\mathbb{R}^m_+$, selon le principe suivant :

1. Initialisation : choisir $q^0 \in \mathbb{R}^m_+$, poser $k = 0$ et fixer un seuil de convergence $\varepsilon > 0$ ainsi que la valeur du paramètre $\lambda > 0$;
2. Boucle en k : pour $k \geqslant 0$, effectuer les opérations suivantes :
   (2.a) déterminer v^k en résolvant le problème d'optimisation *sans contrainte*

$$\mathcal{L}(v^k, q^k) = G(q^k) = \inf_{v\in V} \mathcal{L}(v, q^k); \tag{3.67}$$

   Comme la fonctionnelle $\mathcal{L}(v, q^k)$ est fortement convexe en v à q^k fixé, ce problème admet une et une seule solution.
   (2.b) si $k \geqslant 1$ et si $\|v^k - v^{k-1}\|_V \leqslant \varepsilon$ (par exemple), l'algorithme a convergé ;
   (2.c) déterminer q^{k+1} par projection sur $\mathbb{R}^m_+$

$$q^{k+1} = \Pi_{\mathbb{R}^m_+}(q^k + \lambda\Phi(v^k)), \tag{3.68}$$

   c'est-à-dire, en composantes dans la base cartésienne de $\mathbb{R}^m$,

$$q_i^{k+1} = \max(0, q_i^k + \lambda\Phi_i(v^k)), \qquad \forall 1 \leqslant i \leqslant m.$$

   Cette projection est particulièrement simple à mettre en œuvre puisqu'il suffit d'appliquer les formules (3.46) avec $a_i = 0$ et $b_i = +\infty$ pour tout $1 \leqslant i \leqslant m$.
   (2.d) poser $k \leftarrow k+1$ et revenir à l'étape (2.a).

Proposition 3.38 (Convergence)**.** *On suppose que la fonctionnelle J est fortement convexe de paramètre α et que l'application Φ est Lipschitzienne dans V de constante L. Alors, sous la condition*

$$0 < \lambda < \frac{2\alpha}{L^2}, \tag{3.69}$$

la suite $(v^k)_{k\in\mathbb{N}}$ engendrée par l'algorithme d'Uzawa converge vers l'unique minimiseur global de J dans K.

Remarque (Convergence des multiplicateurs de Lagrange)**.** La convergence de la suite $(q^k)_{k\in\mathbb{N}}$ est plus délicate à analyser. Elle est possible sous des hypothèses supplémentaires. □

3.6 Exercices

Exercice 1. *(Contrôle optimal)* On reprend le problème de contrôle optimal présenté à la section 3.1.4. On pose $V = H_0^1(\Omega)$ et $Y = L^2(\Omega)$. Soit $f \in L^2(\Omega)$. Pour $y \in Y$, on note $\Psi_f(y)$ l'unique solution dans V du problème

$$-\Delta v = f + y \quad \text{dans } \Omega.$$

Le problème consiste à chercher $y \in Y$ minimisant le critère

$$J(y) = \int_\Omega |\Psi_f(y) - v_0|^2 + \int_\Omega |y|^2,$$

pour $v_0 \in Y$ donné.

1. Montrer que pour tout $(y, z) \in Y \times Y$, $\Psi_f(y) - \Psi_f(z) = \Psi_0(y - z)$.
2. Montrer que l'application Ψ_0 est linéaire continue de Y dans Y.
3. Montrer que J est différentiable dans Y et que pour tout $(y, z) \in Y \times Y$, on a

$$\langle J(y), z\rangle_{Y',Y} = 2\int_\Omega (\Psi_f(y) - v_0)\Psi_0(z) + 2\int_\Omega yz.$$

4. Montrer que J est fortement convexe dans Y. (Indication : utiliser le critère (iii) de la proposition 3.10.)
5. En déduire l'existence et l'unicité du minimiseur global de J dans Y.
6. Pour $y \in Y$, on note $\theta(y)$ l'unique solution dans V du problème

$$-\Delta v = \Psi_f(y) - v_0 \quad \text{dans } \Omega.$$

Montrer que $\nabla J(y) = 2(\theta(y) + y)$. On dit que $\theta(y)$ est *l'état adjoint* de $\Psi_f(y)$.

Exercice 2. *(Constante de Poincaré et multiplicateur de Lagrange)* On rappelle l'inégalité de Poincaré sur un ouvert borné Ω de $\mathbb{R}^d$: il existe une constante c_Ω telle que

$$\forall v \in H_0^1(\Omega), \qquad \|v\|_{L^2} \leqslant c_\Omega \|\nabla v\|_{L^2}.$$

On appelle meilleure constante de Poincaré la plus petite valeur que peut prendre la constante c_Ω pour que cette inégalité reste vraie. On pose $V = H_0^1(\Omega)$, $K = \{v \in V;\ \|v\|_{L^2}^2 = 1\}$ et on introduit la fonctionnelle de V dans $\mathbb{R}$ définie par $J(v) = \|\nabla v\|_{L^2}^2$. On admet que J admet un et un seul minimiseur global dans K. On le notera u.

1. Montrer que la contrainte est qualifiée en u.
2. Exprimer la meilleure constante de Poincaré en fonction du multiplicateur de Lagrange associé à la contrainte $\|v\|_{L^2}^2 = 1$.
3. Application : déterminer la meilleure constante de Poincaré pour $\Omega =]0, 1[$.

Exercice 3. *(Optimisation d'une fonctionnelle quadratique sous contrainte linéaire)* On pose $V = \mathbb{R}^N$ et on considère la fonctionnelle quadratique

$$J(v) = \frac{1}{2}(Av, v)_{\mathbb{R}^N} - (b, v)_{\mathbb{R}^N},$$

avec une matrice A symétrique définie positive et $b \in \mathbb{R}^N$. On considère une contrainte égalité de la forme

$$\Phi(v) = (d, v)_{\mathbb{R}^N} - c,$$

avec $d \in \mathbb{R}^N$, $d \neq 0$, et $c \in \mathbb{R}$.

1. Montrer que J admet un et un seul minimiseur u dans $K = \{v \in V;\ \Phi(v) = 0\}$.

2. Montrer que si u est le minimiseur de J dans K, alors il existe $p \in \mathbb{R}$ tel que

$$Z \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix} \qquad \text{avec} \qquad Z = \left[\begin{array}{c:c} A & d \\ \hdashline d^t & 0 \end{array}\right] \in \mathbb{R}^{N+1,N+1}.$$

3. Montrer que la matrice Z est inversible. Est-elle symétrique ? Est-elle positive ?

Exercice 4. *(Critères de forte convexité)* Le but de cet exercice est de prouver la proposition 3.10.

1. Prouver l'implication (i) $\Rightarrow$ (ii).
2. Prouver l'implication (ii) $\Rightarrow$ (iii). (Indication : écrire (ii) deux fois en échangeant les rôles de v et de w.)
3. Prouver l'implication (iii) $\Rightarrow$ (i). (Indication : fixer $(v, w) \in V \times V$ et considérer l'application

$$\varphi : \mathbb{R} \ni t \longmapsto \varphi(t) = J(v + t(w - v)) \in \mathbb{R}.$$

Montrer dans un premier temps que pour $t \geqslant s$,

$$\varphi'(t) - \varphi'(s) \geqslant \alpha \|w - v\|_V^2 (t - s),$$

puis pour tout $\theta \in]0, 1[$, intégrer cette inégalité en t entre θ et 1 et en s entre 0 et θ.)

Exercice 5. *(Fonctionnelle non-convexe)* Sur l'espace $V = H_0^1(\Omega)$ avec $\Omega =]0, 1[$, on définit la fonctionnelle

$$J(v) = \int_\Omega (|v'(x)| - 1)^2 dx + \int_\Omega |v(x)|^2 dx, \qquad \forall v \in V.$$

1. Montrer que J est coercive. (Indication : on développera le carré et on utilisera le fait que $|a| \leqslant \frac{1}{4}a^2 + 1$ pour tout $a \in \mathbb{R}$.)
2. Montrer que J est continue.
3. Montrer que $\inf_{v \in V} J(v) = 0$. (Indication : construire une suite minimisante en « dents de scie ».)
4. Montrer que J n'admet pas de minimiseur global dans V.
5. Quelle est l'hypothèse qui fait défaut pour pouvoir appliquer le théorème 3.4 ?

Exercice 6. *(Trois visions de la moyenne)* On considère une suite ordonnée de N réels $x_1 < x_2 < \ldots < x_N$.

1. Déterminer le minimiseur dans $\mathbb{R}$ de la fonctionnelle $J_2(y) = \sum_{i=1}^N (x_i - y)^2$.
2. Même question pour la fonctionnelle $J_\infty(y) = \max_{1 \leqslant i \leqslant N} |x_i - y|$.
3. Même question pour la fonctionnelle $J_1(y) = \sum_{i=1}^N |x_i - y|$. (Indication : montrer que pour tout $1 \leqslant i \leqslant N - 1$ que $J_1(x_i) = J_1(x_{i+1}) + (N - 2i)(x_{i+1} - x_i)$.)

Exercice 7. *(Algorithme du gradient conjugué)*

1. Montrer par récurrence que $r^k = r(v^k) = b - Av^k$ pour tout $k \geqslant 0$. Expliquer pourquoi, d'un point de vue coût de calcul, il est toutefois plus intéressant d'évaluer r^{k+1} par la formule donnée à l'étape (2.b) de l'algorithme.
2. Montrer que pour tout $k \geqslant 0$, $p^k = 0$ implique $r^k = 0$. Justifier le critère de convergence de l'étape (2.a).
3. Montrer que $(r^k, p^l)_{\mathbb{R}^N} = 0$ pour tout $k \geqslant 1$ et pour tout $l < k$.
4. On fixe $k \geqslant 1$. Pour tout $s = (s^0, \ldots, s^{k-1}) \in \mathbb{R}^k$, on pose $v(s) = v^0 + \sum_{i=0}^{k-1} s^i p^i$ et on introduit la fonctionnelle

$$\Psi : \mathbb{R}^k \ni s \longmapsto \Psi(s) = J(v(s)) \in \mathbb{R}.$$

Vérifier que pour tout $0 \leqslant i \leqslant k-1$,

$$\frac{\partial \Psi}{\partial s_i}(s) = (\nabla J(v(s)), p^i)_{\mathbb{R}^N},$$

et en déduire que le vecteur v^k réalise le minimum de la fonctionnelle J dans l'espace affine $v^0 + K_k$.

Exercice 8. *(Algorithme d'Uzawa et méthode de dualité)* On se place dans le cadre de la section 3.5.2 : on travaille avec une fonctionnelle J fortement convexe dans V et un ensemble convexe et fermé d'états admissibles $K = \{v \in V;\ \Phi(v) \leqslant 0\}$. On désigne par $u \in K$ l'unique minimiseur global de J dans K et par p un élément de $\mathbb{R}^m_+$ tel que le couple (u, p) est point selle du Lagrangien $\mathcal{L}$ défini par (3.58). On introduit la fonctionnelle

$$G : \mathbb{R}^m_+ \ni q \longmapsto G(q) := \inf_{v \in V} \mathcal{L}(v, q) \in \mathbb{R}.$$

On rappelle que $G(p) = \sup_{q \in \mathbb{R}^m_+} G(q) = \mathcal{L}(u, p)$ (voir la proposition 3.35).

1. Vérifier que la fonctionnelle G est concave.
2. On désigne par v_q l'unique élément de V tel que $G(q) = \mathcal{L}(v_q, q)$. Montrer que pour tout $q \in \mathbb{R}^m_+$ et pour tout $\delta \in \mathbb{R}^m$ tel que $q + \delta \in \mathbb{R}^m_+$, on a
$$\delta \cdot \Phi(v_{q+\delta}) \leqslant G(q+\delta) - G(q) \leqslant \delta \cdot \Phi(v_q).$$
3. En supposant que v_q dépend continûment de q, montrer que $\nabla G(q) = \Phi(v_q)$ $(\in \mathbb{R}^m)$.
4. Interpréter l'algorithme d'Uzawa comme un algorithme de gradient à pas fixe avec projection appliqué à un problème que l'on précisera.

Corrigés

Exercice 1. *(Contrôle optimal)*

1. Soit $(y, z) \in Y \times Y$. Il est clair que $\Psi_f(y) - \Psi_f(z) \in V$. De plus,
$$-\Delta(\Psi_f(y) - \Psi_f(z)) = (f + y) - (f + z) = y - z.$$
D'où $\Psi_f(y) - \Psi_f(z) = \Psi_0(y - z)$.
2. La linéarité de Ψ_0 est évidente. Par ailleurs, pour tout $z \in Y$, il vient
$$\|\Psi_0(z)\|_Y = \|\Psi_0(z)\|_{L^2} \leqslant \|\Psi_0(z)\|_{H^1} \leqslant (1 + c_\Omega^2)\|z\|_{L^2} = (1 + c_\Omega^2)\|z\|_Y,$$
car si $u \in V$ satisfait $-\Delta u = z$, on a en utilisant l'inégalité de Poincaré (3.14),
$$(1 + c_\Omega^2)^{-1}\|u\|^2_{H^1} = \int_\Omega \nabla u \cdot \nabla u = \int_\Omega zu \leqslant \|z\|_{L^2}\|u\|_{L^2} \leqslant \|z\|_{L^2}\|u\|_{H^1}.$$
Ceci montre la continuité de Ψ_0.
3. Pour tout $(y, z) \in Y \times Y$, il vient
$$\begin{aligned} J(y+z) &= \int_\Omega |\Psi_f(y+z) - v_0|^2 + \int_\Omega |y+z|^2 \\ &= \int_\Omega |\Psi_f(y) + \Psi_0(z) - v_0|^2 + \int_\Omega |y+z|^2 \\ &= J(y) + 2\left(\int_\Omega (\Psi_f(y) - v_0)\Psi_0(z) + \int_\Omega yz\right) + o(\|z\|_Y). \end{aligned}$$
Il reste à vérifier que le terme entre parenthèses définit bien une application linéaire *continue* dans Y, ce qui est le cas car en utilisant la question précédente, il vient pour tout $z \in Y$,
$$2\int_\Omega (\Psi_f(y) - v_0)\Psi_0(z) + 2\int_\Omega yz \leqslant 2((1 + c_\Omega^2)\|\Psi_f(y) - v_0\|_{L^2} + \|y\|_{L^2})\|z\|_Y.$$

4. Pour tout $(y,z) \in Y \times Y$, on observe que

$$\begin{aligned}\langle J'(y) - J'(z), y-z\rangle_{Y',Y} &= \int_\Omega (\Psi_f(y) - \Psi_f(z))\Psi_0(y-z) + \int_\Omega |y-z|^2 \\ &= \int_\Omega |\Psi_0(y-z)|^2 + \int_\Omega |y-z|^2 \\ &\geqslant \|y-z\|_Y^2,\end{aligned}$$

   d'où la forte convexité de J.
5. Conséquence immédiate du corollaire 3.6, la continuité de J dans Y résultant de sa différentiabilité.
6. Par construction, il vient pour tout $z \in Y$,

$$\begin{aligned}\langle J'(y), z\rangle_{Y',Y} &= 2\int_\Omega (\Psi_f(y) - v_0)\Psi_0(z) + 2\int_\Omega yz \\ &= 2\int_\Omega (-\Delta\theta(y))\Psi_0(z) + 2\int_\Omega yz \\ &= 2\int_\Omega \nabla\theta(y)\cdot\nabla\Psi_0(z) + 2\int_\Omega yz \\ &= 2\int_\Omega \theta(y)(-\Delta\Psi_0(z)) + 2\int_\Omega yz \\ &= 2\int_\Omega \theta(y)z + 2\int_\Omega yz = 2\int_\Omega (\theta(y)+y)z.\end{aligned}$$

   D'où le résultat annoncé.

Exercice 2. *(Constante de Poincaré et multiplicateur de Lagrange)*

1. On constate que pour tout $v \in H_0^1(\Omega)$, $\langle \Phi'(u), v\rangle_{V',V} = 2\int_\Omega uv$. En particulier, en prenant $v = u$, il vient $\langle \Phi(u), u\rangle_{V',V} = 2\|u\|_{L^2}^2 = 2 \neq 0$, ce qui montre que $\Phi'(u) \neq 0$ et donc que la contrainte est qualifiée.
2. On applique le théorème 3.29 avec $m = 1$. On calcule pour tout $v \in V$,

$$\langle J'(u), v\rangle_{V',V} = 2\int_\Omega \nabla u\cdot\nabla v \qquad \text{et} \qquad \langle \Phi'(u), v\rangle_{V',V} = 2\int_\Omega uv.$$

   Par conséquent, il existe $p \in \mathbb{R}$ tel que le minimiseur u satisfait

$$\int_\Omega \nabla u\cdot\nabla v + p\int_\Omega uv = 0, \qquad \forall v \in V.$$

   En prenant $v = u$, il vient $p = -\|\nabla u\|_{L^2}^2 < 0$ ($p \neq 0$ car sinon $u = 0$). D'où, pour tout $v \in V$, $v \neq 0$,

$$J\left(\frac{v}{\|v\|_{L^2}}\right) \geqslant J(u) = -p,$$

   ou encore $\|\nabla v\|_{L^2}^2 \geqslant (-p)\|v\|_{L^2}^2$. On en déduit que $c_\Omega^2 = -\frac{1}{p} (> 0)$.
3. Sur $\Omega =]0,1[$, le minimiseur satisfait $-u'' + pu = 0$, $u(0) = u(1) = 0$ et $u \neq 0$. Ceci n'est possible que s'il existe un entier $n > 0$ tel que $p = -n^2\pi^2$ et $u(x) = \sin(n\pi x)$. La meilleure constante de Poincaré est obtenue pour $n = 1$. Il vient

$$c_\Omega = \frac{1}{\pi}.$$

   Pour cette valeur de la constante, l'inégalité de Poincaré devient une égalité pour la fonction $\sin(\pi x)$.

Exercice 3. *(Optimisation d'une fonctionnelle quadratique sous contrainte linéaire)*

1. L'ensemble K étant un convexe fermé et la fonctionnelle J étant fortement convexe, on peut appliquer le corollaire 3.24.
2. On applique le théorème 3.29 en observant que $\nabla J(u) = Au - b$ et $\nabla\Phi(u) = d$ (les contraintes sont qualifiées puisque $d \neq 0$).
3. Soit $(v, q)^t$ un vecteur dans le noyau de Z. On déduit de la première équation que $v = qA^{-1}d$ si bien qu'en reportant dans la deuxième équation, il vient $q(d, A^{-1}d)_{\mathbb{R}^N} = 0$. Or, $d \neq 0$ par hypothèse et A est définie positive ; d'où $q = 0$ et par suite $v = 0$. La matrice Z est donc inversible. Cette matrice est à l'évidence symétrique. Par contre, elle n'est pas positive puisque la quantité $(v, Av)_{\mathbb{R}^N} + 2q(d, v)_{\mathbb{R}^N}$ n'est pas positive pour tout $(v, q) \in \mathbb{R}^N \times \mathbb{R}$.

Exercice 4. *(Critères de forte convexité)*

1. Soit $(v, w) \in V \times V$ et $t \in [0, 1]$; on pose $\delta = w - v$. De par la forte convexité de J, il vient
$$J(v + t\delta) = J(tw + (1-t)v) \leqslant (1-t)J(v) + tJ(w) - \alpha\frac{t(1-t)}{2}\|\delta\|_V^2,$$
si bien que
$$J(v + t\delta) - J(v) - t\langle J'(v), \delta\rangle_{V',V} \leqslant t(J(w) - J(v)) - t\langle J'(v), \delta\rangle_{V',V} - \alpha\frac{t(1-t)}{2}\|\delta\|_V^2.$$
En faisant tendre $t \to 0^+$ et en utilisant le fait que le membre de gauche est un $o(t)$ car J est différentiable, il vient
$$0 \leqslant J(w) - J(v) - \langle J'(v), \delta\rangle_{V',V} - \frac{1}{2}\alpha\|\delta\|_V^2.$$
D'où le point (ii) en réarrangeant.
2. Écrivons le point (ii) deux fois en échangeant les rôles de v et de w. Il vient
$$J(w) \geqslant J(v) + \langle J'(v), \delta\rangle_{V',V} + \frac{1}{2}\alpha\|\delta\|_V^2,$$
$$J(v) \geqslant J(w) - \langle J'(w), \delta\rangle_{V',V} + \frac{1}{2}\alpha\|\delta\|_V^2.$$
D'où le point (iii) en sommant membre à membre.
3. Fixons $(v, w) \in V \times V$ et $t \in \mathbb{R}$, posons $\varphi(t) = J(v + t\delta)$ avec $\delta = w - v$. Il est clair que
$$\varphi'(t) = \langle J'(v + t\delta), \delta\rangle_{V',V},$$
de sorte que pour $t \geqslant s$, en écrivant le point (iii) pour les vecteurs $v + t\delta$ et $v + s\delta$, il vient
$$\varphi'(t) - \varphi'(s) \geqslant \alpha\|w - v\|_V^2(t - s).$$
Puis, on observe que
$$\begin{aligned}\int_\theta^1 \left(\int_0^\theta [\varphi'(t) - \varphi'(s)]\, ds\right) dt &= \int_\theta^1 [\theta\varphi'(t) - \varphi(\theta) + \varphi(0)]\, dt \\ &= \theta\varphi(1) + (1-\theta)\varphi(0) - \varphi(\theta),\end{aligned}$$
et que
$$\begin{aligned}\int_\theta^1 \left(\int_0^\theta [t - s]\, ds\right) dt &= \int_\theta^1 [t\theta - \tfrac{1}{2}\theta^2]\, dt \\ &= \tfrac{1}{2}\theta(1-\theta).\end{aligned}$$
On conclut en constatant que
$$\theta\varphi(1) + (1-\theta)\varphi(0) - \varphi(\theta) = \theta J(w) + (1-\theta)J(v) - J(\theta w + (1-\theta)v).$$

Exercice 5. *(Fonctionnelle non-convexe)*

1. On a
$$\begin{aligned} J(v) &= \int_\Omega (|v'(x)|^2 - 2|v'(x)| + 1 + |v(x)|^2)dx \\ &= \|v\|_{H^1}^2 + 1 - 2\int_\Omega |v'(x)|dx. \end{aligned}$$
Or, $|v'(x)| \leqslant \frac{1}{4}|v'(x)|^2 + 1$ si bien que
$$J(v) \geqslant \|v\|_{H^1}^2 + 1 - 2\int_\Omega \left(\frac{1}{4}|v'(x)|^2 + 1\right) dx \geqslant \frac{1}{2}\|v\|_{H^1}^2 - 1,$$
ce qui montre la coercivité.
2. Soit $v \in V$. Pour tout $\delta \in V$, on a
$$\begin{aligned} J(v+\delta) - J(v) = &\int_\Omega (|v'(x) + \delta'(x)|^2 - |v'(x)|^2)dx - 2\int_\Omega (|v'(x) + \delta'(x)| - |v'(x)|)dx \\ &+ \int_\Omega (|v(x) + \delta(x)|^2 - |v(x)|^2)dx. \end{aligned}$$
En utilisant l'inégalité triangulaire, on obtient
$$\begin{aligned} |J(v+\delta) - J(v)| \leqslant &\int_\Omega |\delta'(x)|(|v'(x) + \delta'(x)| + |v'(x)|)dx + 2\int_\Omega |\delta'(x)|dx \\ &+ \int_\Omega |\delta(x)|(|v(x) + \delta(x)| + |v(x)|)dx. \end{aligned}$$
En supposant que $\|\delta\|_{H^1} \leqslant 1$ et en utilisant l'inégalité de Cauchy–Schwarz, il vient
$$\begin{aligned} \int_\Omega |\delta'(x)|(|v'(x) + \delta'(x)| + |v'(x)|)dx &\leqslant \int_\Omega |\delta'(x)|(2|v'(x)| + |\delta'(x)|)dx \\ &\leqslant \|\delta'\|_{L^2}\|(2|v'| + |\delta'|)\|_{L^2} \\ &\leqslant \|\delta\|_{H^1}(2\|v\|_{H^1} + 1), \end{aligned}$$
et en traitant de manière analogue les deux autres termes du majorant de $|J(v+\delta) - J(v)|$, on obtient finalement
$$|J(v+\delta) - J(v)| \leqslant C(v)\|\delta\|_{H^1},$$
où la constante $C(v)$ dépend de v mais pas de δ. Ceci montre la continuité de J.
3. La suite minimisante est construite comme suit : pour tout $n \geqslant 0$, l'intervalle $]0,1[$ est découpé en 2^n mailles uniformes en utilisant les points $\{x_k = k2^{-n}\}_{0\leqslant k\leqslant 2^n}$ et on pose
$$v_n(x) = \sum_{k=1}^{2^n} (x - x_{k-1})1_{[x_{k-1}, x_{k-1/2}]}(x) + (x_k - x)1_{[x_{k-1/2}, x_k]}(x),$$
où $x_{k-1/2} = \frac{1}{2}(x_{k-1} + x_k)$ (voir la figure 3.6 pour une illustration). Il est clair que $|v_n'(x)| = 1$ (p.p.) et que $|v_n(x)| \leqslant 2^{-(n+1)}$ si bien que
$$J(v_n) = 0 + \|v_n\|_{L^2}^2 \leqslant 2^{-2(n+1)} \to 0 \quad \text{quand } n \to +\infty.$$
Par ailleurs, il est clair que $J(v) \geqslant 0$ pour tout $v \in V$. D'où $\inf_{v\in V} J(v) = 0$.
4. S'il existait $v \in V$ qui soit un minimiseur global de J dans V, on aurait d'après la question précédente $J(v) = 0$, ce qui implique $v = 0$ (p.p.) et $|v'| = 1$ (p.p.), ce qui est impossible.
5. L'hypothèse de convexité fait défaut.

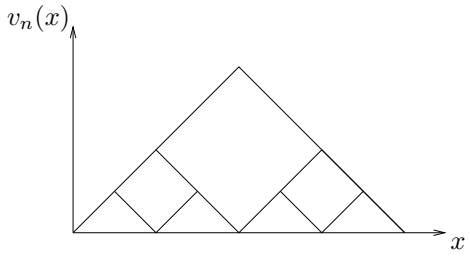


FIGURE 3.6 – Illustration de l'exercice 5 : fonctions v_0, v_1 et v_2.

Exercice 6. *(Trois visions de la moyenne)*

1. La fonctionnelle J_2 est différentiable et fortement convexe dans $\mathbb{R}$. Il vient $\nabla J(y) = 2\sum_{i=1}^{N}(x_i - y)$. Son minimiseur est donc
$$y_2 = \frac{1}{N}\sum_{i=1}^{N} x_i,$$
ce qui correspond à la moyenne arithmétique usuelle des x_i.

2. On observe que
$$J_\infty(y) = \max(|x_1 - y|, |x_N - y|) \geqslant \frac{1}{2}(x_N - x_1) = J\left(\frac{1}{2}(x_1 + x_N)\right).$$
Le minimiseur est donc
$$y_\infty = \frac{1}{2}(x_1 + x_N).$$

3. Pour tout $1 \leqslant i \leqslant N-1$ il vient
$$\begin{aligned} J_1(x_i) &= \sum_{j=1}^{i-1}(x_i - x_j) + \sum_{j=i+1}^{N}(x_j - x_i) \\ &= \sum_{j=1}^{i}(x_{i+1} - x_j) - i(x_{i+1} - x_i) + \sum_{j=i+2}^{N}(x_j - x_{i+1}) + (N-i)(x_{i+1} - x_i) \\ &= J_1(x_{i+1}) + (N - 2i)(x_{i+1} - x_i). \end{aligned}$$
Par ailleurs, puisque la fonctionnelle J_1 est affine par morceaux et convexe, un des points x_i est un minimiseur. On en déduit que si N est pair, J_1 atteint son minimum pour tout $y_1 \in [x_{N/2}, x_{N/2+1}]$ et que si N est impair, J_1 atteint son minimum en $y_1 = x_{(N+1)/2}$.

Exercice 7. *(Algorithme du gradient conjugué)*

1. On a $r^0 = b - Av^0 = r(v^0)$ par construction. De plus, en supposant que $r^k = r(v^k) = b - Av^k$, il vient
$$r^{k+1} = r^k - \alpha^k A p^k = b - Av^k - \alpha^k A p^k = b - (v^k + \alpha^k p^k) = b - Av^{k+1} = r(v^{k+1}).$$
Calculer r^{k+1} par cette formule requiert l'évaluation du produit matrice-vecteur Av^{k+1} dont le coût est de l'ordre de N^2 opérations, alors qu'avec la formule $r^{k+1} = r^k - \alpha^k A p^k$, le coût est de l'ordre de $2N$ opérations, le produit Ap^k ayant déjà été effectué pour le calcul du coefficient α^k.

2. Pour $k = 0$, on a $p^0 = r^0$; l'implication est donc bien vérifiée. Pour $k \geqslant 1$, $p^k = 0$ implique que $r^k \in K_{k-1}$. Or, r^k est orthogonal à cet espace ; on a donc $r^k = 0$. Par conséquent, si $p^k = 0$, v^k est bien point critique de J puisque d'après la question précédente $\nabla J(v^k) = Av^k - b = -r^k = 0$.

3. Soit $k \geqslant 1$ et $l < k$. Alors, $p^l \in K_k$. Or, r^k est orthogonal à K_k car la famille $\{r^0, \dots, r^k\}$ est orthogonale et $\{r^0, \dots, r^{k-1}\}$ constitue une base de K_k.
4. Pour $t \in \mathbb{R}$,
$$J(v(s) + tp^i) = J(v(s)) + t(\nabla J(v(s)), p^i)_{\mathbb{R}^N} + \mathrm{o}(t),$$
ce qui montre que
$$\frac{\partial \Psi}{\partial s_i}(s) = (\nabla J(v(s)), p^i)_{\mathbb{R}^N}.$$
Or, $\nabla J(v(s)) = Av(s) - b$. En posant $\alpha = (\alpha^0, \dots, \alpha^k) \in \mathbb{R}^k$, il vient
$$\frac{\partial \Psi}{\partial s_i}(\alpha) = (\nabla J(v^k), p^i)_{\mathbb{R}^N} = (Av^k - b, p^i)_{\mathbb{R}^N} = -(r^k, p^i)_{\mathbb{R}^N} = 0,$$
d'après les questions précédentes. Le vecteur α est donc point critique de la fonctionnelle Ψ. Or, celle-ci est fortement convexe, ce qui montre que le vecteur v^k réalise le minimum de la fonctionnelle J dans l'espace affine $v^0 + K_k$.

Exercice 8. *(Algorithme d'Uzawa et méthode de dualité)*

1. Soit $v \in V$. On a pour tout $(q, r) \in \mathbb{R}^m_+ \times \mathbb{R}^m_+$ et pour tout $\theta \in [0, 1]$,
$$\begin{aligned}\mathcal{L}(v, \theta q + (1-\theta)r) &= J(v) + \theta q \cdot \Phi(v) + (1-\theta) r \cdot \Phi(v) \\ &= \theta \mathcal{L}(v, q) + (1-\theta)\mathcal{L}(v, r) \\ &\geqslant \theta G(q) + (1-\theta) G(r).\end{aligned}$$
On conclut en prenant l'infimum sur $v \in V$ du membre de gauche.
2. On a
$$\begin{aligned}G(q+\delta) = \mathcal{L}(v_{q+\delta}, q+\delta) &\leqslant \mathcal{L}(v_q, q+\delta) \\ &= \mathcal{L}(v_q, q) + \delta \cdot \Phi(v_q) \\ &= G(q) + \delta \cdot \Phi(v_q).\end{aligned}$$
De même,
$$\begin{aligned}\delta \cdot \Phi(v_{q+\delta}) &= \mathcal{L}(v_{q+\delta}, q+\delta) - \mathcal{L}(v_{q+\delta}, q) \\ &= G(q+\delta) - \mathcal{L}(v_{q+\delta}, q) \\ &\leqslant G(q+\delta) - G(q).\end{aligned}$$
3. Conséquence immédiate de la question précédente.
4. L'algorithme d'Uzawa est un algorithme de gradient à pas fixe avec projection appliqué au problème de la minimisation de la fonctionnelle $-G$ dans le convexe $\mathbb{R}^m_+$.

Chapitre 4

Éléments finis

Ce chapitre est consacré à l'approximation numérique par la méthode des éléments finis de problèmes dits *elliptiques*. Nous nous concentrerons pour simplifier sur le problème de Dirichlet qui consiste à chercher une fonction $u : \overline{\Omega} \to \mathbb{R}$ telle que

$$\left\{ \begin{array}{rcll} -\Delta u(x) & = & f(x), & \forall x \in \Omega, \\ u(x) & = & 0, & \forall x \in \partial\Omega, \end{array} \right. \tag{4.1}$$

où Ω est ouvert borné de $\mathbb{R}^d$ de frontière $\partial\Omega$ et f une fonction donnée, supposée suffisamment régulière. Il s'agit d'un problème *stationnaire* et *linéaire*. À toutes fins utiles, on rappelle qu'en notant $(x_1, \ldots, x_d)$ les coordonnées cartésiennes de $\mathbb{R}^d$, on a

$$\Delta u(x) = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}(x).$$

La méthode des éléments finis constitue un outil très efficace afin d'approcher la solution du problème (4.1). Cette méthode s'applique plus généralement à une large gamme de modèles physiques dont les lois de conservation, mais ces développements sortent du cadre de ce cours. Ce chapitre est organisé comme suit. La section 4.1 contient deux exemples de modèles physiques conduisant à(4.1) puis en rappelle le cadre mathématique (vu en cours d'Analyse). La section 4.2

expose le principe général de la méthode des éléments finis. Puis, les sections 4.3 et 4.4 décrivent, respectivement, la construction d'éléments finis en dimension 1 et 2 et les appliquent à l'approximation de la solution du problème (4.1).

4.1 Motivations et rappel du cadre mathématique

Cette section donne deux exemples de modèles physiques conduisant à des problèmes elliptiques puis rappelle les principaux résultats mathématiques concernant ces problèmes.

4.1.1 Modèles issus de la thermique et de la mécanique

Le problème (4.1) intervient par exemple dans la modélisation de transferts thermiques et de l'équilibre d'une membrane élastique sous un chargement.

1. *Transferts thermiques.* Considérons un matériau occupant le volume Ω. Celui-ci reçoit en tout point $x \in \Omega$ une quantité d'énergie $\widehat{f}(x)$ par unité de volume. En notant $q(x) = (q_1(x), \ldots, q_d(x))$ le vecteur flux de chaleur au point x, l'équation de conservation de l'énergie s'écrit
$$\nabla{\cdot}q(x) = \sum_{i=1}^{d} \frac{\partial q_i}{\partial x_i}(x) = \widehat{f}(x), \qquad \forall x \in \Omega.$$
Soit $u(x)$ la température du matériau au point x. Le vecteur flux de chaleur peut être relié au gradient de température par la loi de Fourier
$$q(x) = -\kappa \nabla u(x), \qquad \forall x \in \Omega,$$
où κ est la conductivité thermique du matériau, supposée constante pour simplifier. En combinant les deux équations ci-dessus, en divisant par κ et en posant $f(x) := \frac{1}{\kappa}\widehat{f}(x)$, nous obtenons la première équation de (4.1) puisque $\nabla{\cdot}\nabla u = \Delta u$. Par ailleurs, en supposant la température fixée à une valeur constante u_0 sur $\partial\Omega$ et quitte à changer u en $u - u_0$, nous obtenons la condition aux limites dans (4.1). D'autres conditions aux limites peuvent être considérées, comme par exemple celle de prescrire la valeur de la composante normale du flux de chaleur q (et donc la dérivée normale de u) sur $\partial\Omega$. Dans ce cas, on parle de *conditions aux limites de Neumann.*
2. *Équilibre d'une membrane élastique.* Soit une membrane plane qui au repos occupe le domaine $\Omega \subset \mathbb{R}^2$. La membrane est tendue selon un champ de contraintes $\sigma \in \mathbb{R}^{2,2}$ donné. Appliquons maintenant un chargement $\widehat{f}$ (force par unité de surface) dans la direction orthogonale au plan de la membrane au repos ; $\widehat{f}$ est donc ici une fonction à valeurs scalaires. Si ce chargement est suffisamment petit, la déformation de la membrane à l'équilibre résulte en première approximation d'un champ de déplacement dans la direction orthogonale à ce même plan. Ce déplacement peut donc être décrit par une fonction à valeurs scalaires $u : \overline{\Omega} \to \mathbb{R}$ (voir la figure 4.1). Dans le cadre de l'élasticité linéaire (cf. cours de Mécanique), on montre que la fonction u est régie par l'EDP
$$-\nabla{\cdot}\big(\sigma{\cdot}\nabla u\big)(x) = \widehat{f}(x), \qquad \forall x \in \Omega. \tag{4.2}$$
Lorsque la membrane au repos est tendue de façon uniforme et isotrope, on a $\sigma = \tau I_2$ où $\tau \in \mathbb{R}$ est la tension de la membrane et I_2 la matrice identité de $\mathbb{R}^{2,2}$. En posant $f(x) := \frac{1}{\tau}\widehat{f}(x)$, nous obtenons la première équation de (4.1). Par ailleurs, la condition aux limites résulte du fait que les bords de la membrane sont maintenus fixes, si bien que le déplacement est nul à la frontière.

4.1.2 Rappel du cadre mathématique

L'objet de cette section est de rappeler deux notions importantes vues en cours d'Analyse : la formulation faible du problème (4.1) et le théorème de Lax–Milgram.

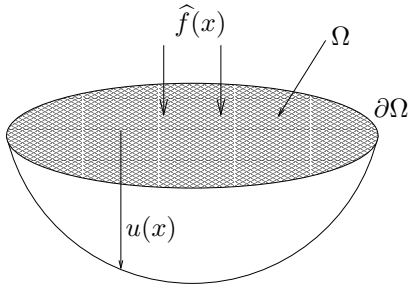


FIGURE 4.1 – Équilibre d'une membrane élastique.

Formulation faible

On suppose par la suite que $f \in L^2(\Omega)$. La norme canonique de cet espace sera notée $\|\cdot\|_{L^2}$. La *formulation faible* du problème (4.1) consiste à

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv, \quad \forall v \in V, \end{cases} \tag{4.3}$$

où

$$V = H_0^1(\Omega) = \{\, v \in H^1(\Omega);\ v = 0 \,\text{p.p. sur}\, \partial\Omega \,\},$$

avec $H^1(\Omega) = \{v \in L^2(\Omega);\ \nabla v \in L^2(\Omega)^d\}$, les dérivées étant entendues au sens des distributions. Si u est solution de (4.3), u satisfait $-\Delta u = f$ p.p. dans Ω et $u = 0$ p.p. sur $\partial\Omega$. Réciproquement, si $u \in H^1(\Omega)$ satisfait ces équations, u est solution de (4.3). Dans le cadre de l'élasticité linéaire, la formulation faible exprime le principe des *travaux virtuels*.

Théorème de Lax–Milgram

Introduisons la forme bilinéaire

$$a : V \times V \ni (u, v) \longmapsto a(u, v) = \int_\Omega \nabla u \cdot \nabla v \in \mathbb{R},$$

et la forme linéaire

$$b : V \ni v \longmapsto b(v) = \int_\Omega fv \in \mathbb{R}.$$

La formulation faible (4.3) s'écrit sous la forme abstraite suivante :

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ a(u, v) = b(v), \quad \forall v \in V. \end{cases} \tag{4.4}$$

Le théorème de Lax–Milgram fournit des conditions suffisantes pour que le problème (4.4) soit bien posé, c'est-à-dire qu'il admette une et une seule solution.

Théorème 4.1 (Lax–Milgram)**.** *Soit V un espace de Hilbert équipé de la norme $\|\cdot\|_V$. On suppose que :*

(i) *la forme linéaire b est* continue*,*

$$\exists \beta < +\infty, \quad \forall v \in V, \quad |b(v)| \leqslant \beta \|v\|_V;$$

(ii) *la forme bilinéaire a est* continue*,*

$$\exists \omega < +\infty, \quad \forall (u, v) \in V \times V, \quad |a(u, v)| \leqslant \omega \|u\|_V \|v\|_V;$$

(iii) *la forme bilinéaire a est* coercive *(on dit également V-*elliptique*)*

$$\exists \alpha > 0, \quad \forall u \in V, \quad a(u,u) \geqslant \alpha \|u\|_V^2. \tag{4.5}$$

Alors, le problème (4.4) *admet une et une seule solution. De plus, son unique solution satisfait l'estimation a priori*

$$\|u\|_V \leqslant \frac{\beta}{\alpha}. \tag{4.6}$$

Le théorème de Lax–Milgram permet de montrer que le problème (4.3) est bien posé. En effet, $H_0^1(\Omega)$ équipé de la norme

$$\|v\|_{H^1} = \left(\int_\Omega v^2 + \int_\Omega |\nabla v|^2 \right)^{1/2} = \left(\|v\|_{L^2}^2 + \|\nabla v\|_{L^2}^2 \right)^{1/2}, \tag{4.7}$$

est un espace de Hilbert. De plus, les formes a et b sont clairement continues (prendre $\beta = \|f\|_{L^2}$ et $\omega = 1$). Par ailleurs, la coercivité de la forme a résulte de l'inégalité de Poincaré (3.14), qui permet de montrer que la forme bilinéaire a est coercive sur $H_0^1(\Omega)$ avec la constante

$$\alpha = \frac{1}{1 + c_\Omega^2}. \tag{4.8}$$

Enfin, l'estimation *a priori* (4.6) s'obtient en prenant $v = u$ dans (4.4) et en utilisant la continuité de la forme linéaire b et la coercivité de la forme bilinéaire a :

$$\alpha \|u\|_V^2 \leqslant a(u,u) = b(u) \leqslant \beta \|u\|_V,$$

ce qui implique (4.6).

4.2 La méthode de Galerkine

La brique fondamentale sur laquelle est construite la méthode des éléments finis est la méthode de Galerkine.[1] Dans cette section, nous allons en étudier le principe, son caractère optimal en termes d'erreur d'approximation et sa reformulation par le biais d'un système linéaire.

4.2.1 Principe

La méthode de Galerkine fournit un moyen simple et élégant d'approcher la solution du problème (4.4). Par la suite, nous supposerons toujours que les hypothèses du théorème de Lax–Milgram sont satisfaites, si bien que ce problème est bien posé.

Le principe de la méthode de Galerkine consiste à remplacer l'espace de dimension infinie V (dans lequel vit la solution exacte u) par un sous-espace de dimension finie V_h (dans lequel sera cherchée la solution approchée u_h). L'espace V_h est appelé *espace d'approximation*. L'indice h fait référence à la finesse du maillage qui a servi à construire l'espace V_h ; son rôle sera précisé dans les sections suivantes. Par la suite, nous ferons l'hypothèse

$$V_h \subset V.$$

On parle *d'approximation conforme*. On peut également concevoir des méthodes de Galerkine dans un cadre non-conforme, mais cela sort du cadre de ce cours.

La version approchée du problème (4.4) consiste à

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ a(u_h, v_h) = b(v_h), \quad \forall v_h \in V_h. \end{cases} \tag{4.9}$$

Proposition 4.2. *Le problème approché* (4.9) *admet une et une seule solution.*

Preuve. a étant coercive sur V, elle l'est *a fortiori* sur V_h puisque $V_h \subset V$. On conclut grâce au théorème de Lax–Milgram. □

1. Le nom de Boris Grigorievich Galerkine (1871–1945) est parfois écrit Galerkin comme en anglais ; nous suivrons ici l'usage général du français pour les noms russes avec une désinence en -ine.

4.2.2 Estimation d'erreur

Notre objectif est maintenant d'estimer *l'erreur d'approximation* $e = u - u_h$ dans la norme $\|\cdot\|_V$. Commençons par une propriété importante de la méthode de Galerkine.

Lemme 4.3 (Orthogonalité de Galerkine)**.** *L'erreur $e = u - u_h$ est telle que*

$$a(e, v_h) = 0, \qquad \forall v_h \in V_h. \tag{4.10}$$

Preuve. Vérification immédiate puisque, de par la propriété de conformité $V_h \subset V$, on a, pour tout $v_h \in V_h$, $a(u, v_h) = b(v_h)$, d'où par linéarité de a, $a(u - u_h, v_h) = a(u, v_h) - a(u_h, v_h) = 0$. □

Lemme 4.4 (Céa)**.** *On a l'estimation d'erreur*

$$\|u - u_h\|_V \leqslant \frac{\omega}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{4.11}$$

Preuve. On a, pour tout $v_h \in V_h$,

$$a(u - u_h, u - u_h) = a(u - u_h, u) = a(u - u_h, u - v_h).$$

En utilisant la coercivité et la continuité de la forme a, il vient

$$\alpha\|u - u_h\|_V^2 \leqslant \omega\|u - u_h\|_V\|u - v_h\|_V$$

d'où l'estimation. □

Il est essentiel d'observer que l'estimation (4.11) est *optimale*. En effet, puisque la solution approchée est cherchée dans V_h, on a nécessairement

$$\inf_{v_h \in V_h} \|u - v_h\|_V \leqslant \|u - u_h\|_V.$$

L'estimation (4.11) montre que l'erreur $\|u - u_h\|_V$ est également majorée par cette borne inférieure multipliée par un facteur qui ne dépend que de la forme bilinéaire a (et donc du modèle physique), indépendamment du choix de l'espace d'approximation V_h. Une autre conséquence importante de l'estimation (4.11) est que la méthode de Galerkine fournit la solution exacte lorsque celle-ci se trouve appartenir à l'espace d'approximation V_h (de tels miracles peuvent (très rarement) se produire !).

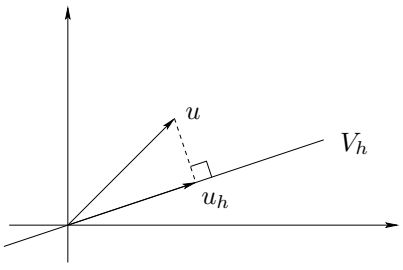


FIGURE 4.2 – Interprétation géométrique de la méthode de Galerkine dans $V = \mathbb{R}^2$.

Remarque (Interprétation géométrique de l'orthogonalité de Galerkine)**.** Lorsque la forme bilinéaire a est symétrique, celle-ci définit un produit scalaire $a(\cdot, \cdot)$ sur V. Les hypothèses de coercivité et continuité sur a signifient que la norme induite par ce produit scalaire est équivalente à la norme $\|\cdot\|_V$ puisque

$$\alpha\|v\|_V^2 \leqslant a(v, v) \leqslant \omega\|v\|_V^2, \qquad \forall v \in V.$$

La relation d'orthogonalité (4.10) admet une interprétation géométrique simple : u_h est la projection orthogonale sur V_h de la solution exacte u par rapport au produit scalaire $a(\cdot, \cdot)$ (voir

figure 4.2). Observons également que la propriété de symétrie de la forme bilinéaire a permet d'améliorer l'estimation d'erreur (4.11). En effet, en utilisant la relation de Pythagore, il vient, pour tout $v_h \in V_h$,

$$a(u-u_h, u-u_h) \leqslant a(u-v_h, u-v_h),$$

ce qui, combiné avec les propriétés de coercivité et continuité de a, fournit

$$\|u-u_h\|_V \leqslant \left(\frac{\omega}{\alpha}\right)^{1/2} \inf_{v_h \in V_h} \|u-v_h\|_V.$$

On observera que cette estimation est meilleure que (4.11) puisque $\frac{\omega}{\alpha} \geqslant 1$. □

4.2.3 Le système linéaire

L'espace V_h étant de dimension finie, le problème approché (4.9) se ramène à la résolution d'un système linéaire. En effet, posons $N = \dim(V_h)$ et soit $(\varphi_1, \ldots, \varphi_N)$ une base de V_h. Posons

$$u_h = \sum_{i=1}^{N} u_i \varphi_i.$$

Le problème (4.9) est équivalent à chercher $U = (u_1, \ldots, u_N) \in \mathbb{R}^N$ tel que

$$\sum_{j=1}^{N} a(\varphi_j, \varphi_i)\, u_j = b(\varphi_i), \qquad \forall 1 \leqslant i \leqslant N.$$

En posant

$$A = (A_{ij})_{1 \leqslant i,j \leqslant N} \in \mathbb{R}^{N,N}, \qquad A_{ij} = a(\varphi_j, \varphi_i), \tag{4.12}$$

et

$$B = (B_i)_{1 \leqslant i \leqslant N} \in \mathbb{R}^N, \qquad B_i = b(\varphi_i), \tag{4.13}$$

nous obtenons le système linéaire

$$AU = B. \tag{4.14}$$

La matrice A s'appelle la *matrice de rigidité* en référence aux problèmes de mécanique où elle a été introduite pour la première fois.

Les propriétés de la matrice A sont directement héritées de celles de la forme bilinéaire a. Nous avons le résultat suivant.

Proposition 4.5 (Propriétés de la matrice de rigidité)**.** *Si la forme bilinéaire a est symétrique, la matrice A est symétrique. Par ailleurs, si la forme bilinéaire a est coercive, la matrice A est définie positive.*

Preuve. La propriété sur la symétrie de A est évidente. Montrons celle sur la définie positivité. Soit $\xi = (\xi_1, \ldots, \xi_N) \in \mathbb{R}^N$ et posons $x = \sum_{i=1}^N \xi_i \varphi_i$ (noter que $x \in V_h \subset V$). Un calcul direct montre que

$$\begin{aligned}
(\xi, A\xi)_{\mathbb{R}^N} &= \sum_{i,j=1}^{N} \xi_i A_{ij} \xi_j \\
&= \sum_{i,j=1}^{N} \xi_i \xi_j a(\varphi_j, \varphi_i) \\
&= a\left(\sum_{j=1}^{N} \xi_j \varphi_j, \sum_{i=1}^{N} \xi_i \varphi_i\right) \\
&= a(x,x),
\end{aligned}$$

si bien que $(\xi, A\xi)_{\mathbb{R}^N} = 0$ implique par coercivité $x = 0$, c'est-à-dire $\xi = 0$. □

4.3 Éléments finis en dimension 1

Dans cette section, on suppose que $\Omega =]a, b[$. En une dimension d'espace, nous notons $\mathbb{P}_k$ l'espace des polynômes de degré inférieur ou égal à k :

$$\mathbb{P}_k = \{p : \mathbb{R} \to \mathbb{R};\ p(x) = \sum_{l=0}^{k} \alpha_l x^l;\ (\alpha_0, \ldots, \alpha_k) \in \mathbb{R}^{k+1}\}.$$

$\mathbb{P}_k$ est un espace vectoriel de dimension $(k+1)$.

4.3.1 Maillages

Un maillage de Ω est de la forme

$$a = x_0 < x_1 < \ldots < x_n < x_{n+1} = b, \tag{4.15}$$

avec un pas de discrétisation local

$$h_i = x_{i+1} - x_i, \qquad \forall 0 \leqslant i \leqslant n.$$

Nous posons $h = \max_{0 \leqslant i \leqslant n} h_i$. Les points $\{x_i\}_{0 \leqslant i \leqslant n+1}$ constituent les *sommets* du maillage et les points $\{x_i\}_{1 \leqslant i \leqslant n}$ les *sommets intérieurs*. On pose

$$K_i := [x_i, x_{i+1}], \qquad \forall 0 \leqslant i \leqslant n.$$

Les segments $\{K_i\}_{0 \leqslant i \leqslant n}$ sont appelés les *éléments* du maillage ou plus simplement les *mailles*. En une dimension d'espace, la relation entre le nombre de sommets N_s, le nombre de sommets intérieurs $N_{s,i}$ et le nombre de mailles N_e est très simple :

$$N_s = N_{s,i} + 2 = N_e + 1 = n + 2. \tag{4.16}$$

Pour des raisons de simplicité, nous supposerons souvent que les sommets du maillage sont régulièrement espacés si bien que pour tout $0 \leqslant i \leqslant n+1$, on a $x_i = ih$ avec

$$h = \frac{b-a}{n+1}.$$

De tels maillages sont dits *uniformes*.

4.3.2 Élément fini de Lagrange $\mathbb{P}_1$

Considérons l'espace fonctionnel de dimension finie constitué des fonctions continues et affines par morceaux sur le maillage (4.15) :

$$V_h^{(1)} = \{v_h \in C^0(\overline{\Omega});\ \forall 0 \leqslant i \leqslant n,\ v_h|_{[x_i, x_{i+1}]} \in \mathbb{P}_1;\ v_h(a) = v_h(b) = 0\}. \tag{4.17}$$

Les fonctions de $V_h^{(1)}$ sont de classe C^1 par morceaux et sont globalement de classe C^0 ; la formule des sauts vue en cours d'Analyse nous permet d'affirmer que la dérivée au sens des distributions d'une fonction de $V_h^{(1)}$ s'évalue simplement en considérant la dérivée usuelle sur chaque maille. Noter que ces dérivées sont constantes maille par maille ; elles sont donc de carré sommable sur Ω. Par conséquent, $V_h^{(1)} \subset H^1(\Omega)$. De plus, les fonctions de $V_h^{(1)}$ sont nulles au bord. Il en résulte le résultat de conformité suivant.

Proposition 4.6 (Conformité)**.** $V_h^{(1)} \subset H_0^1(\Omega)$.

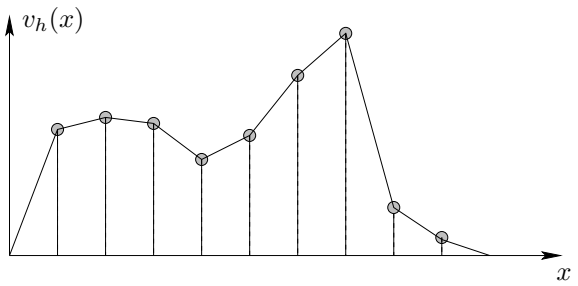


FIGURE 4.3 – Fonction dans l'espace d'approximation $V_h^{(1)}$.

Une fonction dans l'espace d'approximation $V_h^{(1)}$ est illustrée sur la figure 4.3. Introduisons les fonctions suivantes :

$$\forall 1 \leqslant i \leqslant n, \qquad \varphi_i(x) = \begin{cases} \dfrac{x_{i+1} - x}{x_{i+1} - x_i} & \text{si } x \in K_i, \\ \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & \text{si } x \in K_{i-1}, \\ 0 & \text{sinon.} \end{cases}$$

Ces fonctions sont illustrées sur la figure 4.4. Elles sont appelées *fonctions chapeau* en référence à la forme de leur graphe. On observera que, par construction, $\varphi_i \in V_h^{(1)}$ pour tout $1 \leqslant i \leqslant n$ et que

$$\varphi_i(x_j) = \delta_{ij}, \qquad \forall 1 \leqslant i, j \leqslant n, \tag{4.18}$$

où δ_{ij} est le symbole de Kronecker.

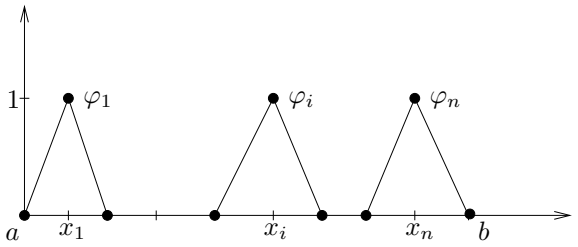


FIGURE 4.4 – Fonctions chapeau.

Proposition 4.7 (Base de $V_h^{(1)}$)**.** *La famille $\{\varphi_1, \ldots, \varphi_n\}$ constitue une base de l'espace $V_h^{(1)}$.*

Preuve. La famille est libre car si on a

$$\sum_{i=1}^{n} \alpha_i \varphi_i \equiv 0 \qquad \text{sur } \Omega,$$

en évaluant cette expression en un sommet intérieur quelconque du maillage x_j et en utilisant le fait que $\varphi_i(x_j) = \delta_{ij}$, il vient

$$\alpha_j = 0.$$

Par ailleurs, la famille $\{\varphi_1, \ldots, \varphi_n\}$ est génératrice de $V_h^{(1)}$. En effet, soit $v_h \in V_h^{(1)}$ et considérons la fonction $w_h : \overline{\Omega} \to \mathbb{R}$ définie par

$$w_h(x) = \sum_{i=1}^{n} v_h(x_i) \varphi_i(x).$$

Sur chaque maille, les fonctions v_h et w_h sont affines et coïncident en deux points distincts (les extrémités de la maille). Elles coïncident donc sur toute la maille. Par suite, $v_h = w_h$ sur $\overline{\Omega}$, ce qui montre que toute fonction de $V_h^{(1)}$ peut se décomposer comme somme des fonctions chapeau. □

Corollaire 4.8 (Dimension de $V_h^{(1)}$)**.** $\dim(V_h^{(1)}) = n$.

Nous introduisons l'opérateur d'interpolation

$$\mathcal{I}_h^{(1)} : C^0(\overline{\Omega}) \ni v \longmapsto \sum_{i=1}^{n} v(x_i)\varphi_i \in V_h^1. \tag{4.19}$$

$\mathcal{I}_h^{(1)} v$ est l'unique fonction de $V_h^{(1)}$ prenant la même valeur que la fonction v en tous les sommets intérieurs du maillage ; voir la figure 4.5. Nous admettons le résultat suivant.

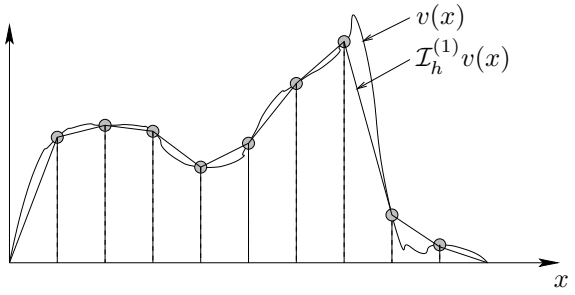


FIGURE 4.5 – Exemple d'interpolation d'une fonction $v \in H_0^1(\Omega)$.

Théorème 4.9 (Interpolation)**.** *Il existe une constante $c_{\mathcal{I}^{(1)}}$, indépendante de h, telle que pour toute fonction $v \in H^2(\Omega) \cap H_0^1(\Omega)$,*

$$\|v - \mathcal{I}_h^{(1)} v\|_{H^1} \leqslant c_{\mathcal{I}^{(1)}} h |v|_{H^2} \qquad \text{et} \qquad \|v - \mathcal{I}_h^{(1)} v\|_{L^2} \leqslant c_{\mathcal{I}^{(1)}} h^2 |v|_{H^2}. \tag{4.20}$$

On rappelle que $H^2(\Omega) = \{v \in L^2(\Omega);\ v' \in L^2(\Omega);\ v'' \in L^2(\Omega)\}$ et que $|v|_{H^2} = (\int_\Omega |v''|^2)^{1/2}$.

Ce résultat montre qu'en raffinant le maillage (c'est-à-dire en prenant un pas de discrétisation h de plus en plus petit), nous pouvons approcher toute fonction de $H^2(\Omega) \cap H_0^1(\Omega)$ par une fonction de $V_h^{(1)}$ avec une plus grande précision et que *l'erreur d'interpolation* $v - \mathcal{I}_h^{(1)} v$ tend vers 0 lorsque $h \to 0$. Enfin, cette erreur tend vers zéro à l'ordre 1 en h pour la norme H^1 et à l'ordre 2 en h pour la norme L^2.

4.3.3 Application au problème de Dirichlet

Soit $f \in L^2(\Omega)$. Nous souhaitons approcher l'unique fonction $u \in V := H_0^1(\Omega)$ telle que

$$\int_a^b u'v' = \int_a^b fv, \qquad \forall v \in V, \tag{4.21}$$

en utilisant la méthode de Galerkine et l'espace d'approximation $V_h^{(1)}$ construit ci-dessus.

Problème discret et analyse d'erreur

Le problème discret consiste à

$$\begin{cases} \text{Chercher } u_h \in V_h^{(1)} \text{ tel que} \\ \displaystyle\int_a^b u_h' v_h' = \int_a^b f v_h, \quad \forall v_h \in V_h^{(1)}. \end{cases} \tag{4.22}$$

Nous avons vu que ce problème revient à la résolution d'un système linéaire de la forme $AU = B$. La matrice de rigidité A est de taille $N := N_{s,i} = n$, le nombre de sommets intérieurs du maillage, et son terme générique est donné par

$$A_{ij} = \int_a^b \varphi_i' \varphi_j', \qquad \forall 1 \leqslant i, j \leqslant N. \tag{4.23}$$

Le membre de droite a pour composantes

$$B_i = \int_a^b f\varphi_i, \qquad \forall 1 \leqslant i \leqslant N. \tag{4.24}$$

On rappelle que de par la proposition 4.5, la matrice A est symétrique définie positive, si bien que le système linéaire $AU = B$ admet une et une seule solution U.

Théorème 4.10 (Estimation d'erreur)**.** *Soit U l'unique solution du système linéaire $AU = B$. On pose $u_h = \sum_{i=1}^N U_i \varphi_i$. Alors, on a l'estimation d'erreur suivante : il existe une constante c, pouvant dépendre de Ω et de f mais pas de h, telle que*

$$\|u - u_h\|_{H^1} \leqslant ch. \tag{4.25}$$

Preuve. Il s'agit d'une conséquence immédiate du Lemme de Céa 4.4 et du théorème d'interpolation 4.9. En utilisant (4.8) et le fait que $-u'' = f \in L^2(\Omega)$ implique $u \in H^2(\Omega)$, il vient

$$\begin{aligned}
\|u - u_h\|_{H^1} &\leqslant (1 + c_\Omega^2) \inf_{v_h \in V_h^{(1)}} \|u - v_h\|_{H^1} \\
&\leqslant (1 + c_\Omega^2) \|u - \mathcal{I}_h^{(1)} u\|_{H^1} \\
&\leqslant (1 + c_\Omega^2) c_{\mathcal{I}^{(1)}} h |u|_{H^2} \\
&\leqslant \{(1 + c_\Omega^2) c_{\mathcal{I}^{(1)}} \|f\|_{L^2}\} h.
\end{aligned}$$

D'où l'estimation (4.25) avec $c = (1 + c_\Omega^2) c_{\mathcal{I}^{(1)}} \|f\|_{L^2}$. □

L'estimation (4.25) signifie que la convergence de l'approximation fournie par la méthode des éléments finis de Lagrange $\mathbb{P}_1$ est d'ordre 1 en norme H^1. En particulier, si le pas du maillage est divisé par deux, il en va de même de l'erreur en norme H^1.

Remarque (Autres normes)**.** Attention, le lemme de Céa ne permet pas d'obtenir directement une estimation d'erreur en norme L^2. En effet, ce lemme est formulé en utilisant la norme $\|\cdot\|_V$ pour laquelle la forme bilinéaire a est coercive et continue, et qui ici coïncide avec la norme H^1. L'exercice 7 montre comment une estimation d'erreur en norme L^2 peut être obtenue. □

Assemblage de la matrice de rigidité

Examinons maintenant d'un peu plus près la structure de la matrice de rigidité A. Pour deux indices $1 \leqslant i, j \leqslant N$ tels que $|i - j| > 1$, les supports des fonctions chapeau φ_i et φ_j sont *disjoints*. Il en résulte que

$$A_{ij} = 0, \qquad |i - j| > 1,$$

si bien que la matrice A est *tridiagonale* (le symbole $\bullet$ indique un coefficient *a priori* non-nul) :

$$A = \begin{pmatrix} \bullet & \bullet & 0 & \dots & 0 \\ \bullet & \bullet & \bullet & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \bullet & \bullet & \bullet \\ 0 & \dots & 0 & \bullet & \bullet \end{pmatrix}.$$

Cette propriété remarquable découle simplement du choix qui a été fait pour le support des fonctions de base $\{\varphi_1, \ldots, \varphi_n\}$. On retiendra qu'en choisissant des fonctions de base ayant un « petit » support, on obtient une matrice de rigidité comprenant « beaucoup » d'éléments nuls. On dit que la matrice de rigidité ainsi obtenue est *creuse* (une définition plus précise sera donnée par la suite ; voir la définition 4.25).

Terminons l'évaluation des coefficients de la matrice A dans le cas d'un maillage uniforme de pas h. Un calcul direct montre que pour tout $1 \leqslant i \leqslant N$,

$$\int_a^b \varphi_i' \, \varphi_i' = \int_{K_{i-1} \cup K_i} \varphi_i' \, \varphi_i' = \frac{2}{h},$$

et pour tout $1 \leqslant i \leqslant N-1$,

$$\int_a^b \varphi_i' \, \varphi_{i+1}' = \int_{K_i} \varphi_i' \, \varphi_{i+1}' = -\frac{1}{h}.$$

En conclusion,

$$A = \frac{1}{h} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \ldots & 0 & -1 & 2 \end{pmatrix}, \tag{4.26}$$

ce que nous noterons sous forme compacte

$$A = \frac{1}{h} \operatorname{tridiag}(-1, 2, -1).$$

Le caractère défini positif de la matrice A peut se vérifier directement de manière relativement simple puisque, pour tout $\xi = (\xi_1, \ldots, \xi_N) \in \mathbb{R}^N$, il vient

$$h(\xi, A\xi)_{\mathbb{R}^N} = \xi_1^2 + \sum_{i=2}^{N} (\xi_i - \xi_{i-1})^2 + \xi_N^2.$$

Fonctions de forme locales

Afin de généraliser ce qui a été fait à des éléments finis de Lagrange utilisant des polynômes de degré élevé, il est utile d'introduire la notion de fonction de forme locale. Considérons la maille de référence $K^* = [0, 1]$ sur laquelle nous définissons les fonctions de forme locales $\{\theta_0^{(1)}, \theta_1^{(1)}\}$ (l'indice supérieur fait ici référence au degré des polynômes) par

$$\theta_0^{(1)}(t) = 1 - t, \qquad \theta_1^{(1)}(t) = t. \tag{4.27}$$

En introduisant les nœuds de K^* définis par $a_0^{(1)} = 0$ et $a_1^{(1)} = 1$, il vient

$$\theta_m^{(1)}(a_n^{(1)}) = \delta_{mn}, \qquad \forall 0 \leqslant m, n \leqslant 1. \tag{4.28}$$

Les fonctions de forme locales $\{\theta_0^{(1)}, \theta_1^{(1)}\}$ sont illustrées sur la figure 4.6

Par ailleurs, toute maille $K_i = [x_i, x_{i+1}]$, $0 \leqslant i \leqslant n$, est l'image de la maille de référence K^* par la transformation affine

$$T_i : K^* \ni t \longmapsto x = x_i + t h_i \in K_i, \tag{4.29}$$

de transformation inverse

$$T_i^{-1} : K_i \ni x \longmapsto t = \frac{x - x_i}{h_i} \in K^*.$$

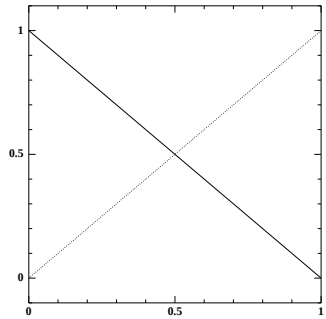


FIGURE 4.6 – Fonctions de forme locales pour l'élément fini de Lagrange $\mathbb{P}_1$.

On vérifie facilement que les fonctions chapeau peuvent être exprimées en fonction des deux fonctions de forme locales $\{\theta_0^{(1)}, \theta_1^{(1)}\}$ et des transformations affines ci-dessus sous la forme

$$\forall 1 \leqslant i \leqslant n, \qquad \varphi_i(x) = \begin{cases} \theta_0^{(1)}(T_i^{-1}(x)) & \text{si } x \in K_i, \\ \theta_1^{(1)}(T_{i-1}^{-1}(x)) & \text{si } x \in K_{i-1}, \\ 0 & \text{sinon,} \end{cases}$$

si bien qu'en utilisant la règle de la dérivation composée, il vient

$$\forall 1 \leqslant i \leqslant n, \qquad \varphi_i'(x) = \begin{cases} \dfrac{1}{h_i}(\theta_0^{(1)})'(T_i^{-1}(x)) & \text{si } x \in K_i, \\ \dfrac{1}{h_{i-1}}(\theta_1^{(1)})'(T_{i-1}^{-1}(x)) & \text{si } x \in K_{i-1}, \\ 0 & \text{sinon.} \end{cases}$$

Introduisons la matrice de rigidité élémentaire d'ordre deux donnée par

$$\mathcal{E} = \left(\int_{K^*} (\theta_m^{(1)})'(t)(\theta_n^{(1)})'(t)dt \right)_{0 \leqslant m,n \leqslant 1} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Les termes non-nuls dans la matrice de rigidité A peuvent s'évaluer en utilisant cette matrice. En effet, il vient pour tout $1 \leqslant i \leqslant n$,

$$\begin{aligned} A_{ii} &= \int_a^b |\varphi_i'(x)|^2 dx \\ &= \sum_{j=0}^{n} \int_{K_j} |\varphi_i'(x)|^2 dx \\ &= \int_{K_{i-1}} |\varphi_i'(x)|^2 dx + \int_{K_i} |\varphi_i'(x)|^2 dx \\ &= \int_{K_{i-1}} \frac{1}{h_{i-1}^2} |(\theta_1^{(1)})'(T_{i-1}^{-1}(x))|^2 dx + \int_{K_i} \frac{1}{h_i^2} |(\theta_0^{(1)})'(T_i^{-1}(x))|^2 dx \\ &= \frac{1}{h_{i-1}} \int_{K^*} |(\theta_1^{(1)})'(t)|^2 dt + \frac{1}{h_i} \int_{K^*} |(\theta_0^{(1)})'(t)|^2 dt \\ &= \frac{1}{h_{i-1}} \mathcal{E}_{22} + \frac{1}{h_i} \mathcal{E}_{11}. \end{aligned}$$

De même, pour tout $1 \leqslant i \leqslant n-1$,

$$A_{i,i+1} = \frac{1}{h_i} \mathcal{E}_{12}.$$

Lorsque le maillage est uniforme, on retrouve bien la matrice A donnée par (4.26).

Quadratures

L'évaluation du membre de droite dans le système linéaire (4.14) nécessite d'évaluer les intégrales

$$\int_a^b f(x)\varphi_i(x)dx, \qquad \forall 1 \leqslant i \leqslant N.$$

Pour une fonction $f \in L^2(\Omega)$ quelconque, ces intégrales ne peuvent pas s'évaluer de façon analytique. On utilise alors une formule d'intégration numérique (cf. section 2.3) : on décompose tout d'abord l'intégrale sur $\Omega =]a, b[$ en une somme d'intégrales sur les mailles intersectant le support de la fonction φ_i,

$$\int_a^b f(x)\varphi_i(x)dx = \sum_{j=0}^{n} \int_{K_j} f(x)\varphi_i(x)dx = \int_{K_{i-1}} f(x)\varphi_i(x)dx + \int_{K_i} f(x)\varphi_i(x)dx.$$

Afin de calculer les deux intégrales du membre de droite, on utilise des *quadratures* de la forme

$$\int_{K_i} \chi(x)dx \simeq \sum_{l=1}^{l_g} \omega_l \chi(\xi_l), \tag{4.30}$$

avec l_g réels $(\omega_1, \ldots, \omega_{l_g})$ appelés *poids de la quadrature* et l_g points $(\xi_1, \ldots, \xi_{l_g})$ de K_i appelés *nœuds de la quadrature.*[2] Rappelons que le plus grand entier k tel que

$$\forall p \in \mathbb{P}_k, \qquad \int_{K_i} p(x)dx = \sum_{l=1}^{l_g} \omega_l p(\xi_l), \tag{4.31}$$

est appelé *degré* de la quadrature (voir definition 2.14) ; on le note k_g.

Il est commode de définir les formules de quadrature sur une maille K_i en spécifiant les poids et les nœuds sur la maille de référence K^* puis en utilisant la transformation T_i pour obtenir les poids et les nœuds de la quadrature sur K_i. Ainsi, si on a choisi des poids $(\omega_1, \ldots, \omega_{l_g})$ et des nœuds $(\xi_1, \ldots, \xi_{l_g})$ sur K^*, il vient

$$\int_{K_i} \chi(x)dx = h_i \int_{K^*} \chi(T_i(t))dt \simeq h_i \sum_{l=1}^{l_g} \omega_l \chi(T_i(\xi_l)).$$

On obtient donc une quadrature sur K_i de poids $(h_i\omega_1, \ldots, h_i\omega_{l_g})$ et de nœuds $(T_i(\xi_1), \ldots, T_i(\xi_{l_g}))$. De plus, cette quadrature est de degré k_g puisque pour tout $p \in \mathbb{P}_k$, $p \circ T_i \in \mathbb{P}_k$. Le tableau 4.1 fournit quelques exemples de quadratures sur K^*.

Une quadrature est d'autant plus précise que son degré est élevé. La proposition suivante précise ce point.

Proposition 4.11 (Erreur de quadrature)**.** *On considère une quadrature de degré k_g sur K_i définie par des poids $(\omega_1, \ldots, \omega_{l_g})$ et des nœuds $(\xi_1, \ldots, \xi_{l_g})$ sur K^*. Alors, il existe une constante c_q, indépendante de la maille K_i, telle que*

$$\forall \chi \in C^{k_g+1}(K_i), \qquad \left| \int_{K_i} \chi(x)dx - \sum_{l=1}^{l_g} h_i\omega_l \chi(T_i(\xi_l)) \right| \leqslant c_q h_i^{k_g+1} \|\chi\|_{C^{k_g+1}(K_i)}. \tag{4.32}$$

Voici un exemple : en utilisant la quadrature du point milieu (tableau 4.1), il vient

$$\int_a^b f(x)\varphi_i(x)dx = \int_{K_{i-1}\cup K_i} f(x)\varphi_i(x)dx = \frac{h_{i-1}}{2} f\left(\frac{x_{i-1}+x_i}{2}\right) + \frac{h_i}{2} f\left(\frac{x_i+x_{i+1}}{2}\right) + \mathrm{O}(h^2),$$

2. L'indice g fait référence à Gauss car les nœuds de la quadrature sont également appelés points de Gauss.

nom	l_g	poids	nœuds	k_g
point milieu	1	(1)	$(\frac{1}{2})$	1
trapèze	2	$(\frac{1}{2}, \frac{1}{2})$	$(0, 1)$	1
Cavalieri–Simpson	3	$(\frac{1}{6}, \frac{2}{3}, \frac{1}{6})$	$(0, \frac{1}{2}, 1)$	2
Gauss 2 points	2	$(\frac{1}{2}, \frac{1}{2})$	$(\frac{1}{2} - \frac{\sqrt{3}}{6}, \frac{1}{2} + \frac{\sqrt{3}}{6})$	3
Gauss 3 points	3	$(\frac{5}{18}, \frac{8}{18}, \frac{5}{18})$	$(\frac{1}{2} - \sqrt{\frac{3}{20}}, \frac{1}{2}, \frac{1}{2} + \sqrt{\frac{3}{20}})$	5

TABLE 4.1 – Quadratures sur $K^* = [0, 1]$.

pourvu que f soit de classe C^2 sur $K_{i-1} \cup K_i$. Sous la même hypothèse de régularité, la quadrature du trapèze donne

$$\int_a^b f(x)\varphi_i(x)dx = \frac{h_{i-1} + h_i}{2} f(x_i) + \mathrm{O}(h^2).$$

Utiliser une quadrature de degré élevé est coûteux puisque cela demande d'évaluer la fonction f en un nombre important de points. Par exemple, si le maillage contient N_e mailles et si la quadrature utilise l_g points dans chaque maille, la fonction f devra être évaluée $N_e l_g$ fois. Par ailleurs, la quadrature doit être suffisamment précise pour préserver la convergence de la solution approchée vers la solution exacte, telle que donnée par le théorème 4.10. Nous admettons le résultat suivant (on rappelle que $N = N_{s,i} = n$).

Théorème 4.12 (Estimation d'erreur avec quadrature)**.** *Soit U_q l'unique solution du système linéaire $AU_q = B_q$ avec A donnée par* (4.26) *et B_q de composantes*

$$B_{q,i} = \sum_{j \in \{i-1, i\}} \sum_{l=1}^{l_g} h_j \omega_l f(T_j(\xi_l))\varphi_i(T_j(\xi_l)), \qquad \forall 1 \leqslant i \leqslant N. \tag{4.33}$$

On pose $u_{q,h} = \sum_{i=1}^N U_{q,i}\varphi_i$. On suppose que $k_g \geqslant 0$. Alors, on a l'estimation d'erreur suivante : il existe une constante c, pouvant dépendre de Ω, du choix de la quadrature et de f mais pas de h, telle que

$$\|u - u_{q,h}\|_{H^1} \leqslant ch. \tag{4.34}$$

4.3.4 Élément fini de Lagrange $\mathbb{P}_2$

Considérons maintenant l'espace fonctionnel de dimension finie constitué des fonctions continues et paraboliques par morceaux sur le maillage (4.15) :

$$V_h^{(2)} = \{v_h \in C^0(\overline{\Omega});\ v_h|_{[x_i, x_{i+1}]} \in \mathbb{P}_2,\ \forall 0 \leqslant i \leqslant n;\ v_h(a) = v_h(b) = 0\}.$$

Une fonction de cet espace est illustrée sur la figure 4.7. Noter que $V_h^{(2)} \subset H_0^1(\Omega)$.

Introduisons les fonctions de forme locales $\{\theta_0^{(2)}, \theta_1^{(2)}, \theta_2^{(2)}\}$ définies sur $K^* = [0, 1]$, la maille de référence, par

$$\theta_0^{(2)}(t) = (2t - 1)(t - 1), \qquad \theta_1^{(2)}(t) = 4t(1 - t), \qquad \theta_2^{(2)}(t) = t(2t - 1). \tag{4.35}$$

En introduisant les nœuds de K^* définis par $a_0^{(2)} = 0$, $a_1^{(2)} = \frac{1}{2}$ et $a_2^{(2)} = 1$, il vient

$$\theta_m^{(2)}(a_n^{(2)}) = \delta_{mn}, \qquad \forall 0 \leqslant m, n \leqslant 2. \tag{4.36}$$

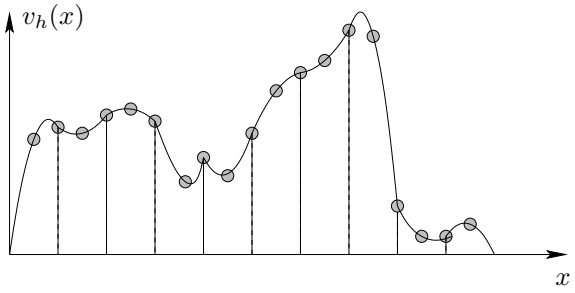


FIGURE 4.7 – Fonction dans l'espace d'approximation $V_h^{(2)}$.

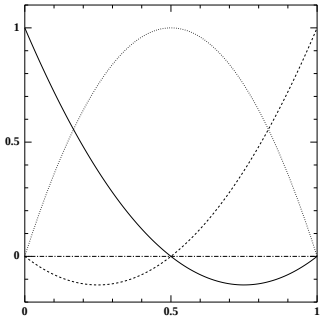


FIGURE 4.8 – Fonctions de forme locales pour l'élément fini de Lagrange $\mathbb{P}_2$.

Les fonctions de forme locales $\{\theta_0^{(2)}, \theta_1^{(2)}, \theta_2^{(2)}\}$ sont illustrées sur la figure 4.8.

Introduisons les fonctions $\{\varphi_1, \ldots, \varphi_n\}$ telles que

$$\forall 1 \leqslant i \leqslant n, \qquad \varphi_i(x) = \begin{cases} \theta_0^{(2)}(T_i^{-1}(x)) & \text{si } x \in K_i, \\ \theta_2^{(2)}(T_{i-1}^{-1}(x)) & \text{si } x \in K_{i-1}, \\ 0 & \text{sinon,} \end{cases}$$

et les fonctions $\{\psi_0, \ldots, \psi_n\}$ telles que

$$\forall 0 \leqslant i \leqslant n, \qquad \psi_i(x) = \begin{cases} \theta_1^{(2)}(T_i^{-1}(x)) & \text{si } x \in K_i, \\ 0 & \text{sinon.} \end{cases}$$

On rappelle que les transformations T_i sont définies par (4.29). Les fonctions ci-dessus sont illustrées sur la figure 4.9. On observera la différence entre le support des fonctions φ_i, qui englobe deux mailles, et celui des fonctions ψ_i, qui est réduit à une maille. Pour cette raison, les fonctions ψ_i sont souvent appelées *fonctions bulles*. Il sera utile d'introduire les milieux des mailles

$$x_{i+\frac{1}{2}} = \frac{1}{2}(x_i + x_{i+1}), \qquad \forall 0 \leqslant i \leqslant n.$$

Nous constatons que pour tout $1 \leqslant i \leqslant n$,

$$\varphi_i \in V_h^{(2)}, \qquad \varphi_i(x_j) = \delta_{ij}, \qquad \varphi_i(x_{j+\frac{1}{2}}) = 0, \tag{4.37}$$

et pour tout $0 \leqslant i \leqslant n$,

$$\psi_i \in V_h^{(2)}, \qquad \psi_i(x_j) = 0, \qquad \psi_i(x_{j+\frac{1}{2}}) = \delta_{ij}. \tag{4.38}$$

Ainsi, les fonctions φ_i sont naturellement associées aux sommets du maillage et les fonctions ψ_i aux milieux des mailles.

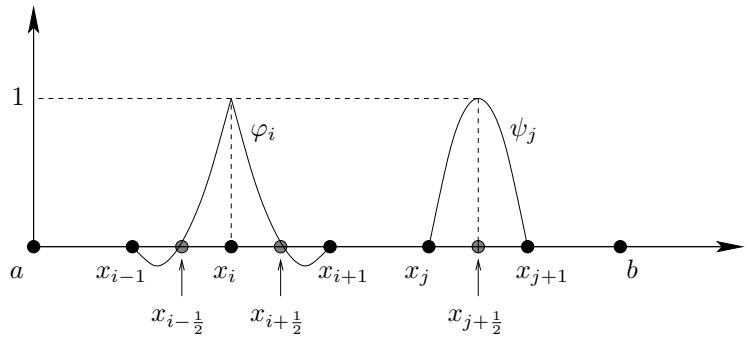


FIGURE 4.9 – Fonctions φ_i et ψ_j ; les sommets du maillage sont indiqués par des cercles noirs et les milieux des mailles par des cercles grisés.

Proposition 4.13 (Base de $V_h^{(2)}$)**.** *La famille* $\{\varphi_1, \ldots, \varphi_n, \psi_0, \ldots, \psi_n\}$ *constitue une base de l'espace* $V_h^{(2)}$.

Preuve. Elle est analogue à celle de la proposition 4.7. En particulier, pour montrer que la famille est génératrice, on considère une fonction $v_h \in V_h^{(2)}$ et on introduit la fonction w_h définie par

$$w_h(x) = \sum_{i=1}^{n} v_h(x_i)\varphi_i(x) + \sum_{i=0}^{n} v_h(x_{i+\frac{1}{2}})\psi_i(x).$$

Les fonctions v_h et w_h sont, par construction, paraboliques par morceaux et sur chaque maille, elles coïncident en trois points distincts : les deux extrémités de la maille et le point milieu. Ces deux fonctions sont donc égales. □

Corollaire 4.14 (Dimension de $V_h^{(2)}$)**.** $\dim(V_h^{(2)}) = 2n + 1$.

Nous introduisons l'opérateur d'interpolation

$$\mathcal{I}_h^{(2)} : C^0(\overline{\Omega}) \ni v \longmapsto \sum_{i=1}^{n} v(x_i)\varphi_i + \sum_{i=0}^{n} v(x_{i+\frac{1}{2}})\psi_i \in V_h^{(2)}. \tag{4.39}$$

$\mathcal{I}_h^{(2)} v$ est l'unique fonction de $V_h^{(2)}$ prenant la même valeur que la fonction v en tous les sommets du maillage et en tous les milieux des mailles. Nous admettons le résultat suivant.

Théorème 4.15 (Interpolation)**.** *Il existe une constante* $c_{\mathcal{I}^{(2)}}$*, indépendante de* h*, telle que pour toute fonction* $v \in H^3(\Omega) \cap H_0^1(\Omega)$*,*

$$\|v - \mathcal{I}_h^{(2)} v\|_{H^1} \leqslant c_{\mathcal{I}^{(2)}} h^2 |v|_{H^3} \qquad \text{et} \qquad \|v - \mathcal{I}_h^{(2)} v\|_{L^2} \leqslant c_{\mathcal{I}^{(2)}} h^3 |v|_{H^3}. \tag{4.40}$$

On rappelle que $H^3(\Omega) = \{v \in L^2(\Omega);\ v' \in L^2(\Omega);\ v'' \in L^2(\Omega);\ v''' \in L^2(\Omega)\}$ *et que* $|v|_{H^3} = (\int_\Omega |v'''|^2)^{1/2}$.

Une comparaison avec le théorème d'interpolation 4.9 est instructive : travailler avec des fonctions paraboliques par morceaux plutôt qu'avec des fonctions affines par morceaux fournit des propriétés d'interpolation plus précises (puisque les exposants de h sont plus élevés dans les estimations d'erreur d'interpolation) *pourvu que la fonction à interpoler soit suffisamment régulière*, à savoir $v \in H^3(\Omega)$ au lieu de $v \in H^2(\Omega)$.

Passons maintenant à l'approximation du problème modèle (4.21) par des éléments finis de Lagrange $\mathbb{P}_2$. Le problème discret consiste à

$$\begin{cases} \text{Chercher } u_h \in V_h^{(2)} \text{ tel que} \\ \displaystyle\int_a^b u_h' v_h' = \int_a^b f v_h, \quad \forall v_h \in V_h^{(2)}. \end{cases} \tag{4.41}$$

Ce problème revient à la résolution d'un système linéaire de la forme $AU = B$. La matrice de rigidité A est de taille $N := N_{s,i} + N_e = 2n + 1$ et son terme générique est donné par

$$A_{ij} = \int_a^b \Phi_i' \Phi_j', \qquad \forall 1 \leqslant i, j \leqslant N.$$

Le membre de droite a pour composantes

$$B_i = \int_a^b f\Phi_i, \qquad \forall 1 \leqslant i \leqslant N.$$

Ici, nous avons introduit pour des raisons de concision la famille de fonctions $\{\Phi_1, \ldots, \Phi_{2n+1}\}$ telle que $\Phi_i = \varphi_i$ pour tout $1 \leqslant i \leqslant n$ et $\Phi_i = \psi_{i-n-1}$ pour tout $n+1 \leqslant i \leqslant 2n+1$. À nouveau, de par la proposition 4.5, la matrice A est symétrique définie positive, si bien que le système linéaire $AU = B$ admet une et une seule solution U.

Théorème 4.16 (Estimation d'erreur)**.** *Soit U l'unique solution du système linéaire $AU = B$. On pose $u_h = \sum_{i=1}^N U_i \Phi_i$. On suppose que $f \in H^1(\Omega)$. Alors, on a l'estimation d'erreur suivante : il existe une constante c, pouvant dépendre de Ω et de f mais pas de h, telle que*

$$\|u - u_h\|_{H^1} \leqslant ch^2. \tag{4.42}$$

Preuve. On utilise à nouveau le Lemme de Céa 4.4, ce qui donne

$$\begin{aligned}\|u - u_h\|_{H^1} &\leqslant (1 + c_\Omega^2) \inf_{v_h \in V_h^{(2)}} \|u - v_h\|_{H^1} \\ &\leqslant (1 + c_\Omega^2) \|u - \mathcal{I}_h^{(2)} u\|_{H^1}.\end{aligned}$$

Afin d'appliquer le théorème d'interpolation 4.15, on a besoin de l'hypothèse de régularité $u \in H^3(\Omega)$. Puisque $-u'' = f$ et que $f \in H^1(\Omega)$ par hypothèse, on a bien $u \in H^3(\Omega)$ et $|u|_{H^3} = |f|_{H^1}$. On conclut comme dans la preuve du théorème 4.10. La constante c vaut $(1 + c_\Omega^2) c_{\mathcal{I}^{(2)}} |f|_{H^1}$. □

L'estimation (4.42) signifie que la convergence de l'approximation fournie par la méthode des éléments finis de Lagrange $\mathbb{P}_2$ est d'ordre 2 en norme H^1. En particulier, si le pas du maillage est divisé par deux, l'erreur en norme H^1 se trouvera divisée par quatre. Attention, ce taux de convergence est obtenu uniquement si la donnée f est dans $H^1(\Omega)$. Si $f \in L^2(\Omega)$ mais $f \notin H^1(\Omega)$, on obtient uniquement l'estimation d'erreur

$$\|u - u_h\|_{H^1} \leqslant ch.$$

Cette estimation est dite *sous-optimale*. Dans cette situation, il n'est pas intéressant d'utiliser l'élément fini de Lagrange $\mathbb{P}_2$, puisqu'un taux de convergence égal à 1 est déjà fourni par l'élément fini de Lagrange $\mathbb{P}_1$ à un coût de calcul moindre, le système linéaire à résoudre étant de taille n au lieu de $(2n + 1)$.

Examinons maintenant d'un peu plus près la structure de la matrice de rigidité A. Celle-ci admet une structure bloc

$$A = \left[\begin{array}{c:c} A^{\varphi\varphi} & A^{\varphi\psi} \\ \hdashline A^{\psi\varphi} & A^{\psi\psi} \end{array}\right] = \left[\begin{array}{c:c} \left(\displaystyle\int_a^b \varphi_i' \varphi_j'\right)_{1 \leqslant i,j \leqslant n} & \left(\displaystyle\int_a^b \varphi_i' \psi_j'\right)_{1 \leqslant i \leqslant n, 0 \leqslant j \leqslant n} \\ \hdashline \left(\displaystyle\int_a^b \psi_i' \varphi_j'\right)_{0 \leqslant i \leqslant n, 1 \leqslant j \leqslant n} & \left(\displaystyle\int_a^b \psi_i' \psi_j'\right)_{0 \leqslant i,j \leqslant n} \end{array}\right],$$

où $A^{\varphi\varphi} \in \mathbb{R}^{n,n}$ est carrée, $A^{\varphi\psi} \in \mathbb{R}^{n,n+1}$ est rectangulaire, $A^{\psi\varphi} \in \mathbb{R}^{n+1,n}$ également et $A^{\psi\psi} \in \mathbb{R}^{n+1,n+1}$ est carrée. La matrice A étant symétrique, on a

$$A^{\varphi\varphi} = (A^{\varphi\varphi})^t, \qquad A^{\varphi\psi} = (A^{\psi\varphi})^t, \qquad A^{\psi\psi} = (A^{\psi\psi})^t.$$

De plus, en examinant les supports des fonctions φ_i et ψ_j, on déduit que $A^{\varphi\varphi}$ est tridiagonale, $A^{\psi\psi}$ diagonale et $A^{\varphi\psi}$ bidiagonale avec la disposition suivante pour les éléments non-nuls :

$$A^{\varphi\psi} = \begin{pmatrix} \bullet & \bullet & 0 & \dots & 0 \\ 0 & \bullet & \bullet & \ddots & \vdots \\ \vdots & \ddots & \bullet & \bullet & 0 \\ 0 & \dots & 0 & \bullet & \bullet \end{pmatrix}.$$

Il reste à calculer les coefficients non-nuls dans les différents blocs. Pour cela, considérons la matrice d'ordre trois donnée par

$$\mathcal{E} = \left(\int_{K^*} (\theta_m^{(2)})'(t)(\theta_n^{(2)})'(t)dt \right)_{0 \leqslant m,n \leqslant 2} = \frac{1}{3} \begin{pmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{pmatrix}.$$

Alors, en procédant comme pour l'élément fini de Lagrange $\mathbb{P}_1$, nous obtenons sur un maillage uniforme

$$\begin{aligned} A_{ii}^{\varphi\varphi} &= \frac{1}{h}(\mathcal{E}_{11} + \mathcal{E}_{33}) = \frac{14}{3h}, \qquad A_{i,i+1}^{\varphi\varphi} = A_{i+1,i}^{\varphi\varphi} = \frac{1}{h}\mathcal{E}_{13} = \frac{1}{3h}, \\ A_{ii}^{\psi\psi} &= \frac{1}{h}\mathcal{E}_{22} = \frac{16}{3h}, \\ A_{ii}^{\varphi\psi} &= \frac{1}{h}\mathcal{E}_{23} = -\frac{8}{3h}, \\ A_{i,i+1}^{\varphi\psi} &= \frac{1}{h}\mathcal{E}_{12} = -\frac{8}{3h}. \end{aligned}$$

En conclusion, en posant $a = 14$, $b = 1$, $c = -8$ et $d = 16$, il vient

$$A = \frac{1}{3h} \left[\begin{array}{ccccc:ccccc} a & b & 0 & \dots & 0 & c & c & 0 & \dots & 0 \\ b & a & b & \ddots & \vdots & 0 & c & c & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & b & a & b & \vdots & & \ddots & c & c & 0 \\ 0 & \dots & 0 & b & a & 0 & \dots & \dots & 0 & c & c \\ \hdashline c & 0 & \dots & \dots & 0 & d & 0 & \dots & \dots & \dots & 0 \\ c & c & \ddots & & \vdots & 0 & d & 0 & & & \vdots \\ 0 & c & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & c & 0 & \vdots & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & c & c & \vdots & & & 0 & d & 0 \\ 0 & \dots & \dots & 0 & c & 0 & \dots & \dots & \dots & 0 & d \end{array}\right].$$

Il est intéressant d'observer que la structure de la matrice de rigidité peut être radicalement modifiée en changeant l'ordre des fonctions de base. Ainsi, en travaillant avec la famille

$$\{\psi_0, \varphi_1, \psi_1, \varphi_2, \dots, \psi_{n-1}, \varphi_n, \psi_n\},$$

on peut vérifier facilement, en raisonnant à nouveau sur les supports des différentes fonctions de

base, que la matrice A devient pentadiagonale (on a posé $e = 0$ afin de faciliter la lecture) :

$$A = \frac{1}{3h}\begin{pmatrix} d & c & e & 0 & \dots & \dots & \dots & 0 \\ c & a & c & b & \ddots & & & \vdots \\ e & c & d & c & e & \ddots & & \vdots \\ 0 & b & c & a & c & b & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & e \\ \vdots & & & \ddots & b & c & a & c \\ 0 & \dots & \dots & \dots & 0 & e & c & d \end{pmatrix}.$$

4.4 Élément fini de Lagrange $\mathbb{P}_1$ en dimension 2

L'objectif de cette section est d'étendre l'élément fini de Lagrange $\mathbb{P}_1$ à la dimension deux afin d'approcher la solution du problème modèle (4.3). Pour simplifier, nous supposons que le domaine Ω est un *polygone* et nous nous limitons aux maillages par des triangles.

4.4.1 Maillages (ou triangulations)

Définition 4.17 (Maillage)**.** *Un* maillage *ou* triangulation *est un recouvrement du polygone Ω par des triangles (par convention, ces triangles sont supposés fermés). En notant $\{K_1, \dots, K_{N_e}\}$ ces triangles, on a donc*

$$\overline{\Omega} = \bigcup_{i=1}^{N_e} K_i. \tag{4.43}$$

On dit que la triangulation est admissible *si pour tout $i \neq j$, l'ensemble $K_i \cap K_j$ est soit vide, soit réduit à un point qui est un sommet à la fois de K_i et de K_j, soit égal à un segment qui est une arête à la fois de K_i et de K_j.*

La figure 4.10 présente un exemple et un contre-exemple de triangulation admissible.

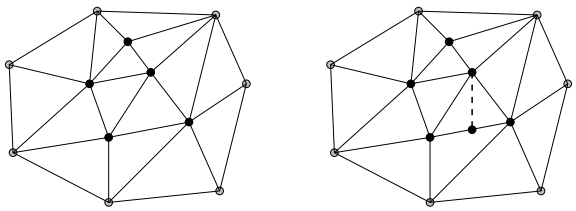


FIGURE 4.10 – Exemple (à gauche) et contre-exemple (à droite) de triangulation admissible.

Par la suite, les triangles K_i sont appelés *mailles* et les arêtes des triangles sont appelées *faces*. Pour tout $1 \leqslant i \leqslant N_e$, nous notons h_i le diamètre de la maille K_i (défini comme la plus grande des longueurs des arêtes de K_i) et nous posons

$$h = \max_{1 \leqslant i \leqslant N_e} h_i.$$

Ce paramètre caractérise la finesse globale du maillage. Nous notons $\{s_{i,1}, s_{i,2}, s_{i,3}\}$ les trois sommets de la maille K_i. En regroupant tous ces sommets, nous obtenons un ensemble de points de cardinal noté N_s et dont les éléments, appelés *sommets du maillage*, sont numérotés sous la forme $\{s_1, \dots, s_{N_s}\}$. Il sera utile de distinguer l'ensemble des sommets du maillage situés dans Ω, de cardinal noté $N_{s,i}$, de celui des sommets situés sur la frontière $\partial\Omega$, de cardinal noté $N_{s,\partial}$.

Par construction, $N_s = N_{s,i} + N_{s,\partial}$. De même, nous notons $\{f_{i,1}, f_{i,2}, f_{i,3}\}$ les trois faces de la maille K_i. En regroupant tous ces faces, nous obtenons un ensemble de segments, appelés *faces du maillage*, dont le cardinal est noté N_f. Quelle que soit la triangulation admissible considérée, les entiers N_e, N_f, N_s et $N_{s,\partial}$ satisfont les relations remarquables suivantes, qui portent le nom de *relations d'Euler* : Sous l'hypothèse que Ω est simplement connexe (Ω ne contient pas de trous), on a

$$\begin{cases} N_e - N_f + N_s = 1, \\ \quad 2N_f - N_{s,\partial} = 3N_e, \end{cases} \tag{4.44}$$

d'où l'on tire, en réarrangeant ces équations,

$$\begin{cases} N_e = 2N_s - N_{s,\partial} - 2, \\ N_f = 3N_s - N_{s,\partial} - 3. \end{cases} \tag{4.45}$$

Dans la limite pratique où $N_{s,\partial} \ll N_s$, il vient

$$\begin{cases} N_e \simeq 2N_s, \\ N_f \simeq 3N_s, \end{cases} \tag{4.46}$$

ce qui veut dire qu'un maillage relativement fin contient approximativement deux fois plus de mailles que de sommets et trois fois plus de faces que de sommets.

4.4.2 Espace polynomial $\mathbb{P}_1$

En deux dimensions d'espace, on pose

$$\mathbb{P}_1 = \{p : \mathbb{R}^2 \to \mathbb{R};\ p(x,y) = \alpha + \beta x + \gamma y;\ (\alpha,\beta,\gamma) \in \mathbb{R}^3\}. \tag{4.47}$$

$\mathbb{P}_1$ est un espace vectoriel de dimension 3. Le résultat suivant joue un rôle clé dans la construction de l'élément fini de Lagrange $\mathbb{P}_1$ en dimension 2.

Proposition 4.18 (Propriétés de $\mathbb{P}_1$)**.** *Un polynôme $p \in \mathbb{P}_1$ est déterminé de manière unique par la valeur qu'il prend en trois points non-alignés. De plus, sa restriction à un segment non réduit à un point est déterminée de manière unique par la valeur qu'elle prend aux deux extrémités de ce segment.*

Soit K^* un triangle fixé de $\mathbb{R}^2$, que l'on suppose non-dégénéré (c'est-à-dire que ses trois sommets ne sont pas alignés). Notons $\{a_1^*, a_2^*, a_3^*\}$ ses sommets. De par la proposition 4.18, il existe une unique fonction $\lambda_1^* \in \mathbb{P}_1$ telle que

$$\lambda_1^*(a_1^*) = 1, \qquad \lambda_1^*(a_2^*) = 0, \qquad \lambda_1^*(a_3^*) = 0.$$

De même, il existe une unique fonction $\lambda_2^* \in \mathbb{P}_1$ telle que $\lambda_2^*(a_1^*) = 0$, $\lambda_2^*(a_2^*) = 1$ et $\lambda_2^*(a_3^*) = 0$, et une unique fonction $\lambda_3^* \in \mathbb{P}_1$ telle que $\lambda_3^*(a_1^*) = 0$, $\lambda_3^*(a_2^*) = 0$ et $\lambda_3^*(a_3^*) = 1$. Les fonctions $\{\lambda_1^*, \lambda_2^*, \lambda_3^*\}$ s'appellent les *coordonnées barycentriques* du triangle K^*. Voici quelques propriétés essentielles des coordonnées barycentriques, toutes de vérification relativement immédiate :

(i) $\lambda_1^* + \lambda_2^* + \lambda_3^* \equiv 1$;
(ii) pour tout $i \in \{1,2,3\}$, λ_i^* est identiquement nulle sur l'arête de K^* opposée au sommet a_i^* ;
(iii) pour tout $x^* \in K^*$ et pour tout $i \in \{1,2,3\}$, $0 \leqslant \lambda_i^*(x^*) \leqslant 1$;
(iv) en notant G^* le barycentre de K^*, on a $\lambda_i^*(G^*) = \frac{1}{3}$ pour tout $i \in \{1,2,3\}$.

4.4.3 Espace d'approximation

Posons

$$V_h^{(1)} = \{v_h \in C^0(\overline{\Omega});\ \forall 1 \leqslant i \leqslant N_e,\ v_h|_{K_i} \in \mathbb{P}_1;\ v_h|_{\partial\Omega} = 0\}.$$

Les fonctions de $V_h^{(1)}$ sont de classe C^1 par morceaux et sont globalement de classe C^0 ; la formule des sauts nous permet d'affirmer que $V_h^{(1)} \subset H^1(\Omega)$, le gradient (au sens des distributions) d'une fonction de $V_h^{(1)}$ pouvant s'évaluer simplement en considérant les dérivées usuelles en x et en y localement sur chaque maille. Noter que ces dérivées sont *constantes* maille par maille. Par ailleurs, les fonctions de $V_h^{(1)}$ sont nulles au bord. Il en résulte le résultat de conformité suivant.

Proposition 4.19 (Conformité)**.** *$V_h^{(1)} \subset H_0^1(\Omega)$.*

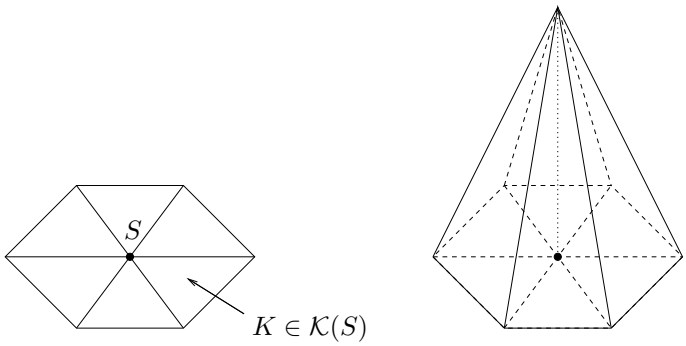


FIGURE 4.11 – Exemple de fonction φ_S en dimension 2.

Nous allons maintenant identifier une base naturelle de l'espace $V_h^{(1)}$ en procédant comme dans le cas unidimensionnel. Soit S un sommet intérieur du maillage. Notons $\mathcal{K}(S)$ l'ensemble des mailles dont S est un sommet. Pour $K \in \mathcal{K}(S)$, notons $\lambda_{K,S}$ la coordonnée barycentrique de K associée au sommet S et posons

$$\varphi_S(x,y) = \begin{cases} \lambda_{K,S}(x,y) & \text{si } (x,y) \in K \text{ pour } K \in \mathcal{K}(S), \\ 0 & \text{sinon.} \end{cases} \tag{4.48}$$

Le support et le graphe d'une fonction φ_S sont illustrés sur la figure 4.11 ; sur cet exemple, le support de φ_S est constitué de 6 triangles. Par construction, φ_S vaut 1 au sommet intérieur S et s'annule sur tous les autres sommets du maillage. De plus, la restriction d'un polynôme de $\mathbb{P}_1$ à une face séparant deux triangles adjacents étant déterminée de manière unique par la valeur que prend ce polynôme aux deux extrémités de cette face, il est clair que $\varphi_S \in C^0(\overline{\Omega})$. Par ailleurs, pour tout $1 \leqslant i \leqslant N_e$, $\varphi_S|_{K_i} \in \mathbb{P}_1$ et $\varphi_S|_{\partial\Omega} = 0$ car sur chaque face du bord, φ_S est nulle aux deux extrémités de la face. Par suite,

$$\varphi_S \in V_h^{(1)}. \tag{4.49}$$

En numérotant les sommets intérieurs du maillage sous la forme $\{S_1, \ldots, S_{N_{s,i}}\}$, on induit une numérotation de la famille de fonctions φ_S sous la forme $\{\varphi_1, \ldots, \varphi_{N_{s,i}}\}$; φ_1 est la fonction qui vaut 1 au sommet S_1 et s'annule sur tous les autres sommets du maillage, etc. En procédant exactement comme dans le cas unidimensionnel, on montre sans peine les résultats suivants.

Proposition 4.20 (Base de $V_h^{(1)}$)**.** *La famille $\{\varphi_1, \ldots, \varphi_{N_{s,i}}\}$ constitue une base de l'espace $V_h^{(1)}$.*

Corollaire 4.21 (Dimension de $V_h^{(1)}$)**.** $\dim(V_h^{(1)}) = N_{s,i}$.

Nous introduisons l'opérateur d'interpolation

$$\mathcal{I}_h^{(1)} : C^0(\overline{\Omega}) \ni v \longmapsto \sum_{i=1}^{N_{s,i}} v(S_i)\varphi_i \in V_h^{(1)}. \tag{4.50}$$

$\mathcal{I}_h^{(1)} v$ est l'unique fonction de $V_h^{(1)}$ prenant la même valeur que v en tous les sommets intérieurs du maillage. Notre objectif est maintenant d'estimer la précision de l'opérateur d'interpolation $\mathcal{I}_h^{(1)}$.

Plus précisément, nous souhaitons obtenir des bornes supérieures pour les quantités $\|v-\mathcal{I}_h^{(1)}v\|_{L^2}$ et $\|v-\mathcal{I}_h^{(1)}v\|_{H^1}$ pour $v \in H^2(\Omega)$. On rappelle (cf. cours d'Analyse) qu'en deux dimensions d'espace, toute fonction de $H^2(\Omega)$ admet un représentant continu si bien que $\mathcal{I}_h^{(1)}v$ est bien défini pour $v \in H^2(\Omega)$.

Contrairement au cas unidimensionnel, l'estimation de *l'erreur d'interpolation* fait intervenir deux paramètres géométriques associés aux mailles, d'une part (comme en une dimension d'espace) le diamètre des mailles K_i, $1 \leqslant i \leqslant N_e$, que nous avons noté h_i, et d'autre part le rayon du cercle inscrit dans la maille K_i que nous noterons ρ_i. Nous posons

$$\sigma_{\{K\}} = \max_{1\leqslant i\leqslant N_e} \frac{h_i}{\rho_i}. \tag{4.51}$$

Cette quantité est, par définition, supérieure à 1. Le rapport h_i/ρ_i est d'autant plus grand que le plus petit angle du triangle K_i est proche de zéro, et ce rapport explose lorsque cet angle tend vers 0. Nous admettons le résultat suivant.

Théorème 4.22 (Interpolation)**.** *Il existe une constante $c_{\mathcal{I}^{(1)}}$, indépendante du maillage, telle que pour toute fonction $v \in H^2(\Omega) \cap H_0^1(\Omega)$,*

$$\|v - \mathcal{I}_h^{(1)}v\|_{H^1} \leqslant c_{\mathcal{I}^{(1)}}\sigma_{\{K\}}h|v|_{H^2} \qquad \text{et} \qquad \|v - \mathcal{I}_h^{(1)}v\|_{L^2} \leqslant c_{\mathcal{I}^{(1)}}h^2|v|_{H^2}. \tag{4.52}$$

Dans l'estimation (4.52) pour la norme H^1 de l'erreur d'interpolation, la borne supérieure peut être atteinte dans certaines situations. Voici un exemple : on considère un unique triangle, de sommets $(0,0)$, $(1,0)$ et $(-1,\epsilon)$ (si bien que $h \sim 1$ et $\rho \sim \epsilon$) et on choisit pour v la fonction $v(x,y) = x^2$. On vérifie facilement que $\mathcal{I}_h^{(1)}v = x + 2y/\epsilon$ si bien que $\|v - \mathcal{I}_h^{(1)}v\|_{H^1}/|v|_{H^2} \sim \epsilon^{-1}$ quand $\epsilon \to 0$. Cette situation peut se produire sur un triangle très aplati avec un angle proche de π et deux angles proches de 0.

Dans la pratique, on ne travaille pas avec un seul maillage fixé $\mathcal{K} = \{K_1, \ldots, K_{N_e}\}$ mais avec une famille de maillages $\{\mathcal{K}^{(1)}, \mathcal{K}^{(2)}, \ldots\}$ obtenus par exemple en effectuant des raffinements successifs afin d'augmenter progressivement la précision du calcul. Sur chaque maillage $\mathcal{K}^{(j)}$, on évalue la constante $\sigma_{\{\mathcal{K}^{(j)}\}}$ selon (4.51).

Définition 4.23 (Régularité)**.** *Soit $\sigma_0 \in \mathbb{R}$ un paramètre fixé. On dit que la famille de maillages $\{\mathcal{K}^{(1)}, \mathcal{K}^{(2)}, \ldots\}$ est* régulière *de paramètre σ_0 si*

$$\forall j = 1, 2, \ldots, \qquad \sigma_{\{\mathcal{K}^{(j)}\}} \leqslant \sigma_0.$$

4.4.4 Application au problème de Dirichlet

Soit $f \in L^2(\Omega)$. Nous souhaitons approcher l'unique fonction $u \in V := H_0^1(\Omega)$ telle que

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv, \qquad \forall v \in V,$$

en utilisant la méthode de Galerkine et l'espace d'approximation $V_h^{(1)}$ construit ci-dessus.

Problème discret et analyse d'erreur

Le problème discret consiste à

$$\begin{cases} \text{Chercher } u_h \in V_h^{(1)} \text{ tel que} \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h = \int_\Omega fv_h, \quad \forall v_h \in V_h^{(1)}. \end{cases} \tag{4.53}$$

Nous avons vu que ce problème revient à la résolution d'un système linéaire de la forme $AU = B$. La matrice de rigidité A est de taille $N := N_{s,i}$, le nombre de sommets intérieurs du maillage, et son terme générique est donné par

$$A_{ij} = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j, \qquad \forall 1 \leqslant i, j \leqslant N. \tag{4.54}$$

Le membre de droite a pour composantes

$$B_i = \int_\Omega f\varphi_i, \qquad \forall 1 \leqslant i \leqslant N. \tag{4.55}$$

On rappelle que de par la proposition 4.5, la matrice A est symétrique définie positive, si bien que le système linéaire $AU = B$ admet une et une seule solution U.

Théorème 4.24 (Estimation d'erreur)**.** *Soit U l'unique solution du système linéaire $AU = B$. On pose $u_h = \sum_{i=1}^N U_i\varphi_i$. On suppose que le maillage $\mathcal{K}$ fait partie d'une famille régulière de maillages de paramètre σ_0 et que le domaine Ω est* convexe*. Alors, on a l'estimation d'erreur suivante : il existe une constante c, pouvant dépendre de Ω, de σ_0 et de f mais pas de h, telle que*

$$\|u - u_h\|_{H^1} \leqslant ch. \tag{4.56}$$

Preuve. En appliquant le Lemme de Céa 4.4, il vient

$$\|u - u_h\|_{H^1} \leqslant (1 + c_\Omega^2) \inf_{v_h \in V_h^{(1)}} \|u - v_h\|_{H^1}.$$

On souhaite majorer le membre de droite en prenant $v_h = \mathcal{I}_h^{(1)}u$ afin d'appliquer le théorème d'interpolation 4.22. Pour cela, il est nécessaire que $u \in H^2(\Omega)$, ce qui est vrai lorsque le polygone Ω est convexe. On montre en effet (et nous l'admettons) que si Ω est convexe et si $f \in L^2(\Omega)$, alors l'unique solution u du problème (4.3) est dans $H^2(\Omega)$ et on a $|u|_{H^2} \leqslant \chi_\Omega \|f\|_{L^2}$ où la constante χ_Ω ne dépend que de Ω. La fin de la preuve est alors immédiate puisqu'il vient

$$\begin{aligned}
\|u - u_h\|_{H^1} &\leqslant (1 + c_\Omega^2)\|u - \mathcal{I}_h^{(1)}u\|_{H^1} \\
&\leqslant (1 + c_\Omega^2)c_{\mathcal{I}^{(1)}}\sigma_0 h|u|_{H^2} \\
&\leqslant \{(1 + c_\Omega^2)c_{\mathcal{I}^{(1)}}\sigma_0\chi_\Omega\|f\|_{L^2}\}h.
\end{aligned}$$

D'où l'estimation (4.56) avec $c = (1 + c_\Omega^2)c_{\mathcal{I}^{(1)}}\sigma_0\chi_\Omega\|f\|_{L^2}$. □

Remarque (Domaine non-convexe)**.** Dans le cas général où le domaine polygonal Ω n'est pas convexe, on montre que la solution exacte u du problème (4.3) est dans $H^{\frac{3}{2}+\epsilon}(\Omega)$ avec $0 < \epsilon \leqslant \frac{1}{2}$, le paramètre ϵ dépendant du plus grand angle obtus dans le polygone Ω. L'approximation par éléments finis de Lagrange $\mathbb{P}_1$ est toujours convergente, mais avec un taux inférieur à un. Plus précisément, on montre que $\|u - u_h\|_{H^1} \leqslant ch^{\frac{1}{2}+\epsilon}$. □

Assemblage de la matrice de rigidité

Une différence importante entre les cas unidimensionnel et bidimensionnel réside dans la structure de la matrice de rigidité, c'est-à-dire dans la disposition des coefficients non-nuls de cette matrice. À cet égard, on distingue deux situations concernant le maillage.

- Les maillages *structurés* s'obtiennent en prenant le produit tensoriel d'un maillage unidimensionnel en x par un maillage unidimensionnel en y. Les sommets sont alors disposés selon une grille tensorielle. Ces maillages sont adaptés aux domaines Ω ayant une forme géométrique relativement simple, par exemple un rectangle ou un carré. Lorsqu'un même pas de maillage constant en x et en y est utilisé, on parle de maillage *uniforme*.
- Dans les maillages *non-structurés*, les sommets forment un nuage de points quelconque sans disposition particulière. De tels maillages sont beaucoup plus intéressants en pratique car ils permettent d'une part de mailler des domaines de géométrie complexe et d'autre part de procéder assez facilement à des raffinements locaux afin d'améliorer localement la précision de l'approximation par éléments finis. Lorsque toutes les mailles ont à peu près la même forme (c'est-à-dire des paramètres h_i et ρ_i sensiblement équivalents), on parle de maillage *quasi-uniforme*.

La figure 4.12 présente un exemple de maillage non-structuré (quasi-uniforme) et deux exemples de maillages structurés uniformes du domaine $\Omega =]0,1[\times]0,1[$. Les sommets intérieurs ont été, dans chaque cas, représentés par des cercles noirs. Les deux maillages structurés ont été construits à partir des maillages unidimensionnels de sommets $\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$ en x et en y, ce qui permet de paver Ω à l'aide de 16 petits carrés, puis chaque carré a été coupé en deux triangles dans un cas et en quatre triangles dans l'autre. Enfin, la figure 4.13 présente un exemple de maillage non-structuré avec un raffinement local autour d'un point (ici le centre du domaine carré). Ce maillage a été construit afin d'approcher avec une bonne précision la solution d'un problème de diffusion hétérogène où la solution exacte présente une singularité au centre du domaine.

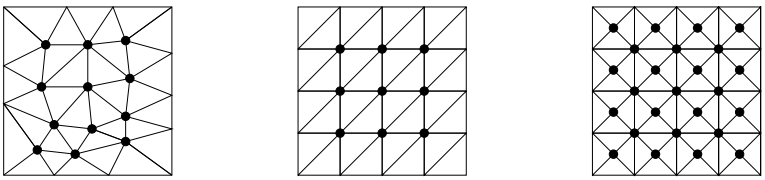


FIGURE 4.12 – Un exemple de maillage non-structuré (quasi-uniforme) et deux exemples de maillages structurés uniformes du carré unité de $\mathbb{R}^2$.

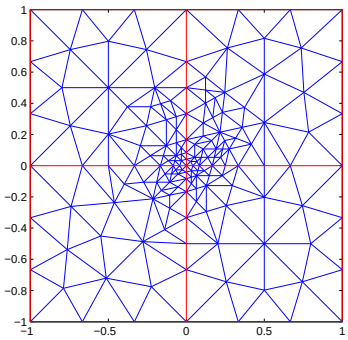


FIGURE 4.13 – Un exemple de maillage non-structuré avec raffinement local.

Examinons maintenant d'un peu plus près la structure de la matrice de rigidité A, dont on rappelle que le terme générique est donné par (4.54). En raisonnant sur les supports des fonctions φ_i et φ_j, il est clair que

$$(A_{ij} \neq 0) \Longrightarrow (\exists K \in \mathcal{K};\ S_i \in K \text{ et } S_j \in K). \tag{4.57}$$

Définition 4.25 (Matrice creuse)**.** *Soit A une matrice d'ordre N. On note N_* le nombre d'éléments non-nuls de la matrice A. On dit que la matrice A est* creuse *si*

$$\frac{N_*}{N^2} \ll 1. \tag{4.58}$$

Pour $i \neq j$, (4.57) montre qu'une condition nécessaire [3] pour que $A_{ij} \neq 0$ est que les sommets S_i et S_j soient situés sur une face du maillage. Chaque face contribue donc au plus à deux coefficients (extra-diagonaux) non-nuls dans la matrice A. De plus, il y a $N_{s,i}$ coefficients diagonaux non-nuls. Par conséquent,

$$N_* \leqslant N_{s,i} + 2N_f,$$

3. À toutes fins utiles, rappelons que si $\mathfrak{P}$ et $\mathfrak{Q}$ sont deux propositions, $\mathfrak{Q}$ est une condition nécessaire pour $\mathfrak{P}$ signifie $\mathfrak{P} \Rightarrow \mathfrak{Q}$ (on dit aussi $\mathfrak{P}$ seulement si $\mathfrak{Q}$) et, par ailleurs, $\mathfrak{Q}$ est une condition suffisante pour $\mathfrak{P}$ signifie $\mathfrak{Q} \Rightarrow \mathfrak{P}$ (on dit aussi $\mathfrak{P}$ si $\mathfrak{Q}$).

et en utilisant (4.46), il vient

$$\frac{N_*}{N^2} \leqslant \frac{N_{s,i} + 2N_f}{N_{s,i}^2} \lesssim \frac{7}{N_{s,i}}.$$

La matrice de rigidité A est donc creuse dès que le maillage est suffisamment fin, ce qui veut dire que dans cette situation, la matrice A contient majoritairement des coefficients nuls. Par ailleurs, on pourra observer qu'en notant $\tau_{0,i}$ le cardinal de l'ensemble $\mathcal{K}(S_i)$, c'est-à-dire le nombre de triangles de $\mathcal{K}$ dont S_i est un sommet, le nombre d'éléments non-nuls dans la ligne i de A est inférieur ou égal à $(\tau_{0,i} + 1)$.

Pour conclure cette section, nous allons évaluer les coefficients non-nuls de la matrice A dans un cas particulier de maillage structuré. Dans les applications, on travaille généralement avec des maillages non-structurés et l'évaluation de la matrice A se fait sur ordinateur. Considérons le maillage suivant (voir figure 4.14) du domaine $\Omega =]0,1[\times]0,1[$ et convenons de numéroter les 9 sommets intérieurs ligne par ligne de la gauche vers la droite en partant de la ligne du bas et en remontant jusqu'à la ligne du haut. Les 32 triangles dans ce maillage sont numérotés de manière analogue. Toutes les mailles sont des triangles rectangles isocèles de côté $h = \frac{1}{4}$ et d'hypoténuse $h' = \frac{\sqrt{2}}{4}$.

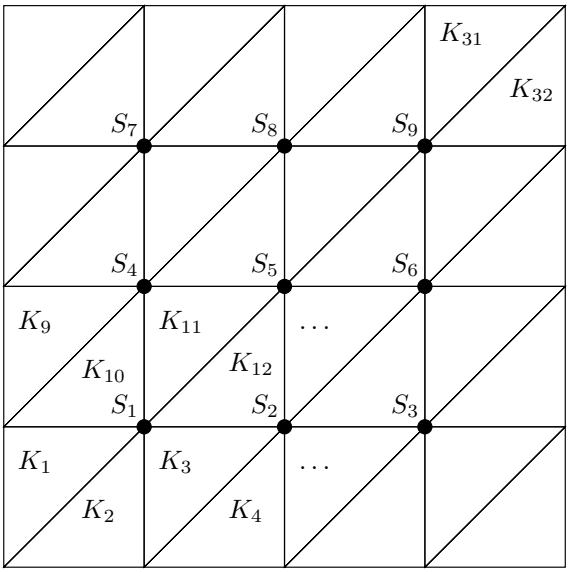


FIGURE 4.14 – Exemple de maillage structuré pour l'évaluation de la matrice A.

La matrice de rigidité est d'ordre $N_{s,i} = 9$ et de par la condition (4.57), nous pouvons d'ores et déjà identifier un certain nombre de coefficients nuls (le symbole $\bullet$ indique comme d'habitude un coefficient *a priori* non-nul) :

$$A = \left[\begin{array}{ccc:ccc:ccc} \bullet & \bullet & 0 & \bullet & \bullet & 0 & 0 & 0 & 0 \\ \bullet & \bullet & \bullet & 0 & \bullet & \bullet & 0 & 0 & 0 \\ 0 & \bullet & \bullet & 0 & 0 & \bullet & 0 & 0 & 0 \\ \hdashline \bullet & 0 & 0 & \bullet & \bullet & 0 & \bullet & \bullet & 0 \\ \bullet & \bullet & 0 & \bullet & \bullet & \bullet & 0 & \bullet & \bullet \\ 0 & \bullet & \bullet & 0 & \bullet & \bullet & 0 & 0 & \bullet \\ \hdashline 0 & 0 & 0 & \bullet & 0 & 0 & \bullet & \bullet & 0 \\ 0 & 0 & 0 & \bullet & \bullet & 0 & \bullet & \bullet & \bullet \\ 0 & 0 & 0 & 0 & \bullet & \bullet & 0 & \bullet & \bullet \end{array}\right].$$

De plus, pour des raisons de symétrie et d'invariance par translation (grâce à l'uniformité du maillage), il vient

$$A = \left[\begin{array}{ccc:ccc:ccc} a & b & 0 & c & d & 0 & 0 & 0 & 0 \\ b & a & b & 0 & c & d & 0 & 0 & 0 \\ 0 & b & a & 0 & 0 & c & 0 & 0 & 0 \\ \hdashline c & 0 & 0 & a & b & 0 & c & d & 0 \\ d & c & 0 & b & a & b & 0 & c & d \\ 0 & d & c & 0 & b & a & 0 & 0 & c \\ \hdashline 0 & 0 & 0 & c & 0 & 0 & a & b & 0 \\ 0 & 0 & 0 & d & c & 0 & b & a & b \\ 0 & 0 & 0 & 0 & d & c & 0 & b & a \end{array}\right].$$

et il nous reste à déterminer les coefficients réels a, b, c et d donnés par les formules suivantes :

$$\begin{aligned} a &= \int_\Omega |\nabla\varphi_1|^2, \\ b &= \int_\Omega \nabla\varphi_1 \cdot \nabla\varphi_2, \\ c &= \int_\Omega \nabla\varphi_1 \cdot \nabla\varphi_4, \\ d &= \int_\Omega \nabla\varphi_1 \cdot \nabla\varphi_5. \end{aligned}$$

La première étape consiste à découper les intégrales sur Ω en une somme d'intégrales sur les mailles et à ne conserver que les mailles intersectant le support des fonctions chapeau à intégrer. Il vient

$$\begin{aligned} a &= \int_{K_1} |\nabla\varphi_1|^2 + \int_{K_2} \ldots + \int_{K_3} \ldots + \int_{K_{10}} \ldots + \int_{K_{11}} \ldots + \int_{K_{12}} \ldots \\ b &= \int_{K_3} \nabla\varphi_1 \cdot \nabla\varphi_2 + \int_{K_{12}} \ldots \\ c &= \int_{K_{10}} \nabla\varphi_1 \cdot \nabla\varphi_4 + \int_{K_{11}} \ldots \\ d &= \int_{K_{11}} \nabla\varphi_1 \cdot \nabla\varphi_5 + \int_{K_{12}} \ldots \end{aligned}$$

Soit φ_i une fonction chapeau et soit $K \in \mathcal{K}(S_i)$ une maille incluse dans le support de φ_i. Sur cette maille, φ_i est affine si bien que son gradient est un vecteur constant. Notons F_{K,S_i} la face de K opposée au sommet S_i et notons n_{K,S_i} la normale à F_{K,S_i} sortante de K. Notons enfin d_{K,S_i} la distance du sommet S_i à la face F_{K,S_i} (cette distance est égale à la longueur de la hauteur du triangle K issue de S_i). Avec ces notations, nous obtenons

$$\nabla\varphi_i|_K = -\frac{1}{d_{K,S_i}} n_{K,S_i}. \tag{4.59}$$

En effet, les vecteurs $\nabla\varphi_i|_K$ et n_{K,S_i} sont colinéaires car ils sont tous deux orthogonaux à la face F_{K,S_i} sur laquelle la fonction φ_i est constante (et égale à 0). De plus, en suivant la hauteur du triangle K issue de S_i jusqu'à la face F_{K,S_i}, la fonction φ passe de la valeur 1 à la valeur 0. La figure 4.15 illustre les considérations ci-dessus.

En appliquant la formule (4.59) au cas de la figure 4.14, nous obtenons par exemple

$$\nabla\varphi_1|_{K_1} = -\frac{1}{h}\begin{pmatrix} -1 \\ 0 \end{pmatrix},$$

et comme K_1 est de mesure égale à $\frac{h^2}{2}$, il vient

$$\int_{K_1} |\nabla\varphi_1|^2 = \frac{h^2}{2}\frac{1}{h^2} = \frac{1}{2}.$$

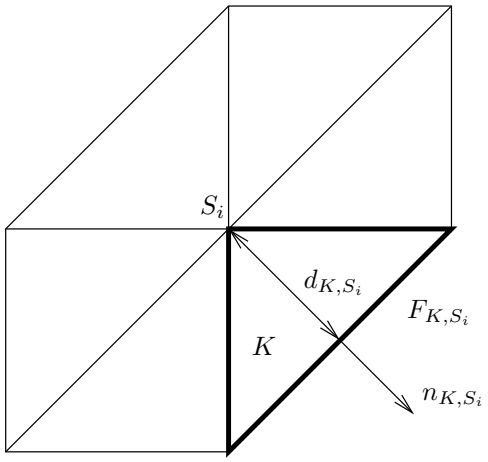


FIGURE 4.15 – Support de la fonction chapeau associée au sommet S_i, triangle K dans $\mathcal{K}(S_i)$ indiqué en gras, face F_{K,S_i}, normale n_{K,S_i} et distance d_{K,S_i}.

De même,

$$\nabla\varphi_1|_{K_3} = -\frac{\sqrt{2}}{h}\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix},$$

si bien que

$$\int_{K_3}|\nabla\varphi_1|^2 = \frac{h^2}{2}\frac{2}{h^2} = 1.$$

Enfin, pour des raisons de symétrie évidentes,

$$\int_{K_1}|\nabla\varphi_1|^2 = \int_{K_2}|\nabla\varphi_1|^2 = \int_{K_{11}}|\nabla\varphi_1|^2 = \int_{K_{12}}|\nabla\varphi_1|^2,$$

et

$$\int_{K_3}|\nabla\varphi_1|^2 = \int_{K_{10}}|\nabla\varphi_1|^2.$$

En rassemblant les contributions ci-dessus, nous obtenons

$$a = 4.$$

En procédant comme ci-dessus pour les trois autres coefficients b, c et d, il vient

$$b = c = -1 \qquad \text{et} \qquad d = 0.$$

La nullité du coefficient d provient du fait que sur les deux triangles K_{11} et K_{12}, les gradients des fonctions chapeau φ_1 et φ_5 sont orthogonaux.

L'exemple ci-dessus s'étend facilement au cas d'un maillage structuré plus fin de pas $h = \frac{1}{n+1}$ ($n = 3$ dans l'exemple ci-dessus). Il y a alors $N_{s,i} = n^2$ sommets intérieurs et la matrice A est d'ordre n^2 (elle contient donc n^4 coefficients, la plupart nuls). En numérotant à nouveau les sommets intérieurs ligne par ligne de la gauche vers la droite en partant de la ligne du bas et en remontant jusqu'à la ligne du haut, nous obtenons une matrice A qui a la structure bloc suivante :

$$A = \begin{pmatrix} B & C & O & \dots & O \\ C & B & C & \ddots & \vdots \\ O & \ddots & \ddots & \ddots & O \\ \vdots & \ddots & C & B & C \\ O & \dots & O & C & B \end{pmatrix}.$$

Les blocs B, C et O sont des matrices d'ordre n et il y a n blocs par ligne dans la structure de la matrice A. De plus, $B = \text{tridiag}(-1, 4, -1)$, $C = -I$ où I est la matrice identité d'ordre n et O est le bloc nul d'ordre n. On dit que la matrice A est *bloc tridiagonale*.

Remarque (Unité des coefficients de la matrice de rigidité)**.** En une dimension d'espace, les coefficients de la matrice de rigidité ont les dimensions de l'inverse d'une longueur ; d'où le coefficient $\frac{1}{h}$ qui apparaît dans cette situation. En deux dimensions d'espace, la situation change puisque les coefficients de la matrice de rigidité résultent d'une intégration sur une surface et non plus sur un segment. Ces coefficients sont alors sans dimension ; d'où l'absence de coefficient $\frac{1}{h}$ dans les expressions ci-dessus. □

Résolution du système linéaire

Lorsque la matrice de rigidité est symétrique définie positive, une méthode très efficace pour résoudre le système linéaire est l'algorithme du gradient conjugué (cf. section 3.3.4). Il s'agit d'une méthode itérative permettant d'approcher la solution du système linéaire. La convergence de l'algorithme étant en général rapide, il est possible d'obtenir une très bonne approximation de la solution en un nombre modéré d'itérations. Rappelons que l'algorithme du gradient conjugué exploite le fait que la solution du système linéaire est le minimiseur de la fonctionnelle quadratique (ou d'énergie) $\frac{1}{2}(Av, v)_{\mathbb{R}^N} - (b, v)_{\mathbb{R}^N}$ où N désigne la taille du système linéaire.

Dans certains cas, notamment pour des problèmes posés en une dimension d'espace et pour des problèmes posés en deux dimensions et de petite taille, il est également possible de considérer une méthode de résolution directe. La méthode la plus générale est celle du *pivot de Gauß*. Elle consiste dans un premier temps à décomposer la matrice A sous la forme $A = LU$ où L est une matrice triangulaire inférieure et U une matrice triangulaire supérieure. On a donc $L_{ij} = 0$ et $U_{ji} = 0$ pour tout $1 \leqslant i \leqslant N$ et pour tout $j > i$. Une fois évaluées ces matrices, la résolution du système linéaire $Au = b$ est relativement simple puisqu'il suffit d'effectuer les résolutions suivantes :

(1) évaluer $v \in \mathbb{R}^N$ tel que $Lv = b$;

(2) évaluer $u \in \mathbb{R}^N$ tel que $Uu = v$ (si bien que $Au = LUu = Lv = b$).

Ces 2 étapes sont très simples à implémenter : le système triangulaire inférieur $Lv = b$ se résout en calculant v_1, puis v_2 jusqu'à v_N, tandis que le système triangulaire supérieur $Uu = v$ se résout en calculant u_N puis u_{N-1} jusqu'à u_1. Lorsque la matrice A est symétrique définie positive, il existe une variante de la méthode du pivot de Gauß, connue sous le nom de méthode de *Choleski*, qui permet d'exploiter la symétrie de la matrice A pour obtenir une décomposition de la forme $A = LL^T$. De plus, dans le cas particulier des matrices tridiagonales, les seuls coefficients non-nuls de la matrice L se trouvent *a priori* sur sa diagonale et sa diagonale inférieure.

Un aspect important dans l'évaluation des performances d'une méthode numérique est son coût. Afin de le quantifier, on peut effectuer un décompte du nombre d'opérations élémentaires (additions, multiplications, etc.) nécessaires à sa mise en œuvre. Dans la limite (pratique) où $N \gg 1$, le coût de calcul des coefficients des matrices L et U est de l'ordre de $N^3/3$ opérations, et il est de l'ordre de $N^3/6$ pour la décomposition de Choleski. La résolution des systèmes triangulaires inférieur et supérieur ne nécessite elle que de l'ordre de N^2 opérations. Par ailleurs, le coût d'une itération de l'algorithme du gradient conjugué est de l'ordre de N^2 opérations. Cet algorithme est donc plus efficace qu'une approche par résolution directe lorsque le nombre d'itérations N_{it} nécessaire afin d'obtenir la convergence est tel que $N_{\text{it}} \ll N$. En pratiquant, cela s'obtient en employant un préconditionneur (cf. à nouveau section 3.3.4). À ce stade, il suffit de savoir que de (très) bons préconditionneurs sont disponibles pour une vaste gamme d'applications et que des bibliothèques informatiques dédiées sont disponibles.

Quadratures

Afin d'évaluer les composantes du membre de droite du système linéaire $AU = B$ (voir la formule (4.55)), il est en général nécessaire d'utiliser une quadrature. Le principe est le même qu'en dimension 1 : une quadrature à l_g points sur une maille K consiste en la donnée de l_g réels $(\omega_1, \ldots, \omega_{l_g})$ appelés *poids de la quadrature* et de l_g points $(\xi_1, \ldots, \xi_{l_g})$ de K appelés *nœuds de la*

quadrature. L'intégrale sur K d'une fonction $\chi : K \to \mathbb{R}$ est alors approchée de la façon suivante :

$$\int_K \chi(x,y)dxdy \simeq \sum_{l=1}^{l_g} \omega_l \chi(\xi_l). \tag{4.60}$$

Introduisons l'espace vectoriel des fonctions polynomiales en x et en y de degré total inférieur ou égal à k :

$$\mathbb{P}_k = \left\{ p : \mathbb{R}^2 \to \mathbb{R};\ p(x,y) = \sum_{\substack{0 \leqslant m,n \leqslant k \\ m+n \leqslant k}} \alpha_{mn} x^m y^n;\ \alpha_{mn} \in \mathbb{R} \right\}.$$

$\mathbb{P}_k$ est un espace vectoriel de dimension $\frac{1}{2}(k+1)(k+2)$. Le degré de la quadrature est alors défini comme le plus grand entier k tel que

$$\forall p \in \mathbb{P}_k, \qquad \int_K p(x,y)dxdy = \sum_{l=1}^{l_g} \omega_l p(\xi_l). \tag{4.61}$$

Comme en dimension 1, une quadrature est d'autant plus précise que son degré est élevé.

En deux dimensions d'espace, il est commode de repérer les nœuds de la quadrature en indiquant leurs coordonnées barycentriques. Le tableau 4.2 fournit quelques exemples de quadratures sur un triangle K de surface $|K|$. Dans ce tableau, nous appelons multiplicité le nombre de permutations qu'il faut réaliser sur les coordonnées barycentriques afin d'obtenir tous les nœuds de la quadrature. Par exemple, la quadrature de degré 2 utilise 3 nœuds dont les coordonnées barycentriques sont $(\frac{1}{2}, \frac{1}{2}, 0)$, $(\frac{1}{2}, 0, \frac{1}{2})$ et $(0, \frac{1}{2}, \frac{1}{2})$, les trois poids correspondants étant tous égaux à $\frac{1}{3}|K|$.

l_g	poids	nœuds	mult.	k_g
1	$\|K\|$	$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$	1	1
3	$\frac{1}{3}\|K\|$	$(1,0,0)$	3	1
3	$\frac{1}{3}\|K\|$	$(\frac{1}{2}, \frac{1}{2}, 0)$	3	2
4	$-\frac{9}{16}\|K\|$	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	1	4
	$\frac{25}{48}\|K\|$	$(\frac{1}{5}, \frac{1}{5}, \frac{3}{5})$	3	

TABLE 4.2 – Quadratures sur un triangle K de surface $|K|$.

En utilisant la première des quadratures à 3 points du tableau 4.2 afin d'évaluer les composantes du vecteur B, il vient

$$B_i = \int_\Omega f\varphi_i = \sum_{K \in \mathcal{K}(S_i)} \int_K f\varphi_i = \sum_{K \in \mathcal{K}(S_i)} \frac{1}{3}|K| f(S_i) = \frac{1}{3}|\mathcal{K}(S_i)| f(S_i),$$

où $|\mathcal{K}(S_i)|$ désigne la surface du polygone $\mathcal{K}(S_i)$. Enfin, comme en dimension 1, on montre qu'il suffit d'utiliser une quadrature de degré $k_g \geqslant 0$ afin d'évaluer le membre de droite pour préserver l'ordre de convergence optimal de l'approximation par éléments finis de Lagrange $\mathbb{P}_1$.

4.5 Exercices

Exercice 1. *(Principe du maximum discret)* Soit A la matrice obtenue par discrétisation du problème (4.21) sur un maillage uniforme de pas $h = \frac{b-a}{N+1}$. On rappelle que A est d'ordre N et

donnée par

$$A = \frac{1}{h}\text{tridiag}(-1, 2, -1).$$

Soit $B \in \mathbb{R}^N$ le membre de droite du système linéaire $AU = B$ de composantes données par (4.24).

1. Pour un vecteur $V \in \mathbb{R}^N$, on note $V \leqslant 0$ si $V_i \leqslant 0$ pour tout $1 \leqslant i \leqslant N$. Soit $V \in \mathbb{R}^N$ tel que $AV \leqslant 0$. Montrer que $V \leqslant 0$.
2. En déduire que si la fonction f du problème (4.21) est telle que $f \leqslant 0$ sur Ω, alors la solution approchée u_h est telle que $u_h \leqslant 0$ sur Ω. Cette propriété porte le nom de principe du maximum discret.
3. On note α_{ij}, $1 \leqslant i, j \leqslant N$, les coefficients de la matrice A^{-1}. Montrer que $\alpha_{ij} \geqslant 0$ pour tout $1 \leqslant i, j \leqslant N$.
4. Montrer que pour tout $1 \leqslant i \leqslant N$,
$$\sum_{j=1}^{N} \alpha_{ij} \leqslant \frac{1}{8h}.$$
(Indication : considérer la fonction $w(x) = \frac{1}{2}x(1-x)$.) En déduire que $\|u_h\|_{L^\infty} \leqslant \frac{1}{8}\|f\|_{L^\infty}$.

Exercice 2. *(Preuve du théorème 4.9)* Soit $\Omega =]a, b[$. On considère le maillage (4.15) et pour tout $0 \leqslant i \leqslant n$, on pose $K_i := [x_i, x_{i+1}]$ et $h_i = x_{i+1} - x_i$. Soit $v \in H^2(\Omega)$.

1. On considère une maille K_i et on définit la fonction $w_i : K_i \to \mathbb{R}$ telle que
$$w_i(s) = v'(s) - \frac{v(x_{i+1}) - v(x_i)}{x_{i+1} - x_i}, \qquad \forall s \in K_i.$$
En partant de l'identité $v(x) - (\mathcal{I}_h^{(1)} v)(x) = \int_{x_i}^{x} w_i(s)ds$ pour tout $x \in K_i$, montrer que
$$\|v - \mathcal{I}_h^{(1)} v\|_{L^2(K_i)} \leqslant h_i \|w_i\|_{L^2(K_i)}.$$
2. Montrer que w_i s'annule sur K_i et en déduire que
$$\|w_i\|_{L^2(K_i)} \leqslant h_i \|w_i'\|_{L^2(K_i)}.$$
3. Conclure.

Exercice 3. *(Convection–diffusion)* On considère le problème suivant

$$\begin{cases} -\nu u'' + \beta u' = f & \text{dans } \Omega :=]0, 1[, \\ u(0) = u(1) = 0, \end{cases}$$

où ν et β sont deux réels positifs et f une fonction dans $L^2(\Omega)$. L'inconnue u représente par exemple la concentration d'une espèce chimique transportée dans un écoulement unidimensionnel de vitesse β, ν est le coefficient de diffusion et f le terme source.

1. Formuler le problème ci-dessus sous forme faible et montrer que ce problème est bien posé. On précisera notamment la forme bilinéaire a et la forme linéaire b.
2. Évaluer la matrice de rigidité A associée au problème approché avec des éléments finis de Lagrange $\mathbb{P}_1$. On mettra le rapport $\frac{\nu}{h}$ en facteur et on fera apparaître le nombre de Péclet
$$\gamma = \frac{h\beta}{\nu}.$$
3. Lorsque $\gamma > 2$, c'est-à-dire lorsque $h > \frac{2\nu}{\beta}$, le coefficient $A_{i,i+1}$ devient positif et on peut montrer que la solution approchée est alors polluée par des oscillations non-physiques. Une première approche pour éviter cette difficulté consiste à prendre le pas du maillage suffisamment fin pour que $h < \frac{2\nu}{\beta}$. Toutefois, lorsque le coefficient de diffusion est très petit, cette

approche comporte l'utilisation de maillages très fins, ce qui s'avère coûteux. Une approche alternative consiste à introduire un terme de diffusion numérique dans le problème approché. On introduit la forme bilinéaire

$$a_h(u_h, v_h) = a(u_h, v_h) + \frac{1}{2}h\beta \int_\Omega u_h' v_h',$$

et on considère le nouveau problème approché qui consiste à

$$\begin{cases} \text{Chercher } u_h \in V_h^{(1)} \text{ tel que} \\ a_h(u_h, v_h) = b(v_h), \quad \forall v_h \in V_h^{(1)}. \end{cases}$$

Evaluer la nouvelle matrice de rigidité et vérifier que ses coefficients extra-diagonaux restent toujours négatifs.

4. Soit $w_h \in V_h^{(1)}$. Montrer que

$$\nu |u_h - w_h|_{H^1} \leqslant \sup_{v_h \in V_h^{(1)}} \frac{a_h(u_h - w_h, v_h)}{|v_h|_{H^1}},$$

où $|\cdot|_{H^1}$ désigne la semi-norme H^1 et vérifier que

$$a_h(u_h - w_h, v_h) = a(u - w_h, v_h) + a(w_h, v_h) - a_h(w_h, v_h).$$

5. Montrer que pour tout couple $(y, z) \in H_0^1(\Omega) \times H_0^1(\Omega)$, $|a(y, z)| \leqslant \max(\nu, \beta)\|y\|_{H^1}|z|_{H^1}$ et en déduire que

$$\nu |u_h - w_h|_{H^1} \leqslant \max(\nu, \beta)\|u - w_h\|_{H^1} + \frac{1}{2}h\beta |w_h|_{H^1}.$$

6. En choisissant $w_h = \mathcal{I}_h^{(1)} u$ et en admettant que $|\mathcal{I}_h^{(1)} u|_{H^1} \leqslant c_0 \|u\|_{H^1}$ où c_0 ne dépend pas de h, montrer l'estimation d'erreur

$$\|u - u_h\|_{H^1} \leqslant ch,$$

avec une constante c indépendante de h.

Exercice 4. *(Élément fini de Hermite)* Soit $\Omega =]a, b[$. On pose $V = \{v \in H^2(\Omega);\ v(a) = v'(a) = v(b) = v'(b) = 0\}$. On admet qu'il existe une constante $\beta > 0$ telle que

$$\forall v \in V, \qquad \beta(\|v\|_{L^2} + \|v'\|_{L^2}) \leqslant \|v''\|_{L^2}.$$

Soit $f \in L^2(\Omega)$. On considère le problème suivant :

$$\begin{cases} \text{Chercher } u \in V \text{ tel que} \\ \displaystyle\int_\Omega u'' v'' = \int_\Omega f v, \quad \forall v \in V. \end{cases} \tag{4.62}$$

Ce problème intervient par exemple dans la modélisation de la flexion d'une poutre encastrée.

1. Montrer que le problème (4.62) est bien posé.
2. Montrer que tout polynôme de $\mathbb{P}_3$ est uniquement déterminé par sa valeur et celle de sa dérivée en deux points distincts.
3. Soit n un entier positif. On considère un maillage uniforme de Ω de pas $h = \frac{b-a}{n+1}$ et de sommets $x_i = ih$ pour tout $0 \leqslant i \leqslant n+1$. On introduit l'espace d'approximation

$$V_h = \{v_h \in C^1(\overline{\Omega});\ \forall 0 \leqslant i \leqslant n,\ v_h|_{[x_i, x_{i+1}]} \in \mathbb{P}_3;\ v_h(a) = v_h'(a) = v_h(b) = v_h'(b) = 0\}.$$

Justifier pourquoi $V_h \subset V$.

4. Construire une famille de fonctions $\{\varphi_1, \dots, \varphi_n\}$ de V_h telle que

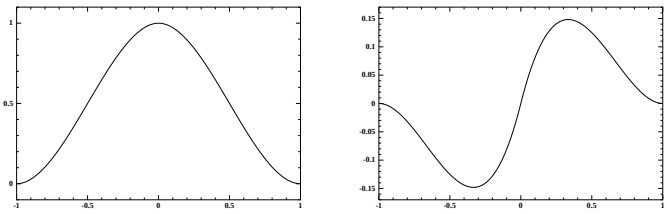
$$\forall 1 \leqslant i, j \leqslant n, \qquad \varphi_i(x_j) = \delta_{ij} \text{ et } \varphi_i'(x_j) = 0.$$

De même, construire une famille de fonctions $\{\psi_1, \dots, \psi_n\}$ de V_h telle que

$$\forall 1 \leqslant i, j \leqslant n, \qquad \psi_i(x_j) = 0 \text{ et } \psi_i'(x_j) = \delta_{ij}.$$

On utilisera pour cela les fonctions de forme $\theta_1(t) = (2|t|+1)(|t|-1)^2$ et $\theta_2(t) = t(|t|-1)^2$ représentées ci-dessous pour $t \in [-1, 1]$ et des transformations affines.

$\theta_1(t) = (2|t|+1)(|t|-1)^2$ $\qquad$ $\theta_2(t) = t(|t|-1)^2$



5. Montrer que la famille $\{\varphi_1, \dots, \varphi_n, \psi_1, \dots, \psi_n\}$ est une base de V_h.
6. On note A la matrice de rigidité associée au problème (4.62) dans la base ci-dessus. Préciser la disposition des coefficients *a priori* non-nuls dans A.
7. On décide maintenant d'ordonner les fonctions de base dans l'ordre suivant

$$\{\varphi_1, \psi_1, \varphi_2, \psi_2, \dots, \varphi_n, \psi_n\}$$

On note $\widetilde{A}$ la matrice de rigidité correspondante. Préciser la disposition des coefficients *a priori* non-nuls dans $\widetilde{A}$.

Exercice 5. *(Erreur d'approximation aux nœuds du maillage)* On considère l'approximation du problème (4.21) par éléments finis de Lagrange $\mathbb{P}_1$ sur le maillage (4.15). Pour tout $1 \leqslant i \leqslant n$, on introduit la fonction $G_i : \Omega \to \mathbb{R}$ telle que

$$G_i(x) = \begin{cases} \dfrac{b - x_i}{b - a}(x - a) & \text{si } a \leqslant x \leqslant x_i, \\ \dfrac{x_i - a}{b - a}(b - x) & \text{si } x_i \leqslant x \leqslant b. \end{cases}$$

1. Montrer que pour tout $v \in H_0^1(\Omega)$,

$$\int_a^b G_i' v' = v(x_i), \qquad \forall 1 \leqslant i \leqslant n.$$

2. En déduire que

$$u_h(x_i) = u(x_i), \qquad \forall 1 \leqslant i \leqslant n.$$

Ce miracle ne se produit qu'en dimension 1.

Exercice 6. *(Éléments finis mixtes)* Soit $\Omega =]0, 1[$. On s'intéresse au problème qui consiste à

$$\begin{cases} \text{Chercher } (u, p) \in H_0^1(\Omega) \times L^2(\Omega) \text{ tel que} \\ -p' = f \qquad \text{dans } \Omega, \\ u' = p \qquad \text{dans } \Omega, \end{cases} \tag{4.63}$$

où f est donnée dans $L^2(\Omega)$. Ce problème est une version simplifiée d'un modèle d'écoulement en milieu poreux (loi de Darcy). On vérifie facilement que si (u, p) est solution de (4.63), u est solution du problème qui consiste à

$$\begin{cases} \text{Chercher } u \in H_0^1(\Omega) \text{ tel que} \\ -u'' = f \qquad \text{dans } \Omega, \end{cases}$$

problème qu'on saurait résoudre numériquement par les méthodes d'éléments finis vues dans ce chapitre. Il peut cependant s'avérer intéressant de conserver explicitement dans la simulation numérique la variable auxiliaire p, qui représente physiquement un débit, surtout si c'est précisément cette quantité qui nous intéresse *in fine*. Le but de cet exercice est de montrer comment résoudre directement le problème (4.63).

1. Reformuler le problème (4.63) sous la forme

$$\begin{cases} \text{Chercher } (u,p) \in H_0^1(\Omega) \times L^2(\Omega) \text{ tel que} \\ \forall q \in L^2(\Omega), \quad a(p,q) + b(q,u) = 0, \\ \forall v \in H_0^1(\Omega), \quad b(p,v) = c(v), \end{cases}$$

   avec des formes bilinéaires a et b et une forme linéaire c que l'on précisera.

2. On réalise une approximation de Galerkine du problème ci-dessus en considérant un maillage uniforme de l'intervalle Ω à l'aide de N segments de longueur $h = 1/N$ et en utilisant l'élément fini $\mathbb{P}_0$ (fonctions constantes par élément) pour approcher p et l'élément fini de Lagrange $\mathbb{P}_1$ pour approcher u. Quelle est la taille du vecteur P discrétisant le champ p ? Même question pour le vecteur U discrétisant le champ u ? Montrer que le problème discret est équivalent à un système linéaire du type

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} P \\ U \end{pmatrix} = \begin{pmatrix} 0 \\ F \end{pmatrix}.$$

   Donner les dimensions et les termes génériques des matrices A et B et du vecteur F en utilisant les fonctions chapeau et les fonctions indicatrices des mailles.

3. Evaluer explicitement les coefficients des matrices A et B.

4. Vérifier que la matrice B est injective. En déduire que la matrice $B^T A^{-1} B$ est définie positive puis que la matrice bloc

$$\begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix}$$

   est inversible.

Exercice 7. *(Estimation d'erreur en norme L^2)* On suppose que pour tout $f \in L^2(\Omega)$, l'unique solution u du problème (4.3) est dans $H^2(\Omega)$ avec $|u|_{H^2} \leqslant \chi_\Omega \|f\|_{L^2}$ où la constante χ_Ω ne dépend que de Ω (une condition suffisante pour que cette hypothèse soit satisfaite est que le polygone Ω soit convexe). Soit u_h la solution approchée du problème (4.3) avec l'élément fini de Lagrange $\mathbb{P}_1$ et un maillage $\mathcal{K}$ faisant partie d'une famille régulière de maillages de paramètre σ_0. Dans ces conditions, on rappelle (cf. théorème 4.24) l'estimation d'erreur suivante : il existe une constante c, pouvant dépendre de Ω, de σ_0 et de f mais pas de h, telle que

$$\|u - u_h\|_{H^1} \leqslant ch.$$

1. Soit ζ l'unique solution du problème (4.3) en prenant pour donnée f l'erreur d'approximation $u - u_h \in L^2(\Omega)$. On a donc

$$\int_\Omega \nabla\zeta \cdot \nabla v = \int_\Omega (u - u_h) v, \qquad \forall v \in H_0^1(\Omega).$$

   Montrer que

$$\|u - u_h\|_{L^2}^2 = \int_\Omega \nabla(u - u_h) \cdot \nabla(\zeta - \mathcal{I}_h^{(1)}\zeta).$$

2. En utilisant le fait que $|\zeta|_{H^2} \leqslant \chi_\Omega \|u - u_h\|_{L^2}$, en déduire que

$$\|u - u_h\|_{L^2} \leqslant \hat{c} h^2,$$

   où la constante $\hat{c}$ peut dépendre de Ω, de σ_0 et de f mais pas de h.

Exercice 8. *(Coordonnées barycentriques)*

1. Préciser les coordonnées barycentriques $\{\lambda_i\}_{1\leqslant i\leqslant 3}$ pour le triangle rectangle isocèle K de sommets $s_1 = (0,0)$, $s_2 = (h,0)$ et $s_3 = (0,h)$.
2. Calculer le gradient de ces trois fonctions et vérifier la formule (4.59). Expliquer pourquoi la somme de ces trois gradients est identiquement nulle.
3. On considère la matrice d'ordre 3 donnée par $\mathcal{E} = (\int_K \nabla\lambda_i \cdot \nabla\lambda_j)_{1\leqslant i,j\leqslant 3}$. Expliquer pourquoi la somme des coefficients de chaque ligne et de chaque colonne de la matrice $\mathcal{E}$ est nulle. Calculer $\mathcal{E}_{11}$ et $\mathcal{E}_{22}$ puis donner la valeur de tous les autres coefficients de la matrice $\mathcal{E}$ sans calculer de nouvelles intégrales.

Exercice 9. *(Projection L^2)* Soient Ω un polygone de $\mathbb{R}^2$, $\mathcal{K}$ un maillage admissible de Ω comprenant N sommets intérieurs et $V_h^{(1)}$ l'espace d'éléments finis de Lagrange $\mathbb{P}_1$ sur ce maillage. Soit $\{\varphi_1, \ldots, \varphi_N\}$ la base usuelle de $V_h^{(1)}$. On considère la projection $\Pi_h : L^2(\Omega) \to V_h^{(1)}$ telle que pour tout $v \in L^2(\Omega)$,

$$\|v - \Pi_h v\|_{L^2} = \inf_{v_h \in V_h^{(1)}} \|v - v_h\|_{L^2}.$$

$\Pi_h v$ est la fonction de $V_h^{(1)}$ qui est la plus proche de v pour la norme L^2. On rappelle (cf. (3.44)) que $\Pi_h v$ est caractérisé par les relations suivantes :

$$\int_\Omega (v - \Pi_h v)\varphi_i = 0, \qquad \forall 1 \leqslant i \leqslant N.$$

1. Soit $M \in \mathbb{R}^{N,N}$ la matrice de terme générique $M_{ij} = \int_\Omega \varphi_i\varphi_j$ pour tout $1 \leqslant i,j \leqslant N$. Montrer que M est définie positive.
2. Montrer que $\Pi_h v$ peut être évalué en résolvant un système linéaire de la forme $MX = V$ où $V \in \mathbb{R}^N$ est un vecteur dont on précisera les composantes.
3. Evaluer la matrice M sur le maillage de la figure 4.16. (Indication : on pourra utiliser la quadrature de degré $k_q = 2$ du tableau 4.2.)

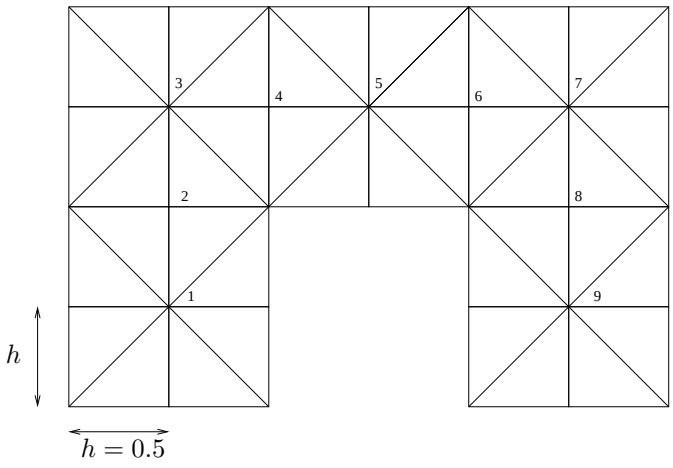


FIGURE 4.16 – Maillage pour l'exercice 9.

.

Exercice 10. *(Élément fini de Crouzeix–Raviart)*

1. Soient K un triangle non-dégénéré et $\{m_i\}_{1\leqslant i\leqslant 3}$ les milieux de ses trois arêtes. Montrer qu'un polynôme de $\mathbb{P}_1$ est uniquement déterminé par les quantités $\{p(m_i)\}_{1\leqslant i\leqslant 3}$. Les fonctions $\{\theta_i\}_{1\leqslant i\leqslant 3}$ de $\mathbb{P}_1$ telles que $\theta_i(m_j) = \delta_{ij}$ sont appelées les fonctions de forme locales sur le triangle K pour l'élément fini de Crouzeix–Raviart.

2. Exhiber les fonctions $\{\theta_i\}_{1\leqslant i\leqslant 3}$ lorsque K est le triangle rectangle isocèle d'arête unité. Calculer dans ce cas les termes de la matrice $\mathcal{E} \in \mathbb{R}^{3,3}$ de terme générique $\mathcal{E}_{ij} = \int_K \nabla\theta_i \cdot \nabla\theta_j$.
3. Soient Ω un polygone de $\mathbb{R}^2$ et $\mathcal{K}$ un maillage admissible de Ω. On note $\mathcal{F}_h$ l'ensemble des faces du maillage situées à l'intérieur de Ω et $\mathcal{F}_h^0$ l'ensemble des faces du maillage situées sur la frontière $\partial\Omega$. Pour une face $f \in \mathcal{F}_h \cup \mathcal{F}_h^0$, on note m_f son point milieu. Pour une face intérieure $f \in \mathcal{F}_h$ séparant deux triangles K_1 et K_2, on note φ_f la fonction affine par morceaux, de support inclus dans $K_1 \cup K_2$ et égale sur chacun de ces deux triangles à la fonction de forme locale θ_i associée au milieu de la face f (voir figure 4.17, le support de φ_f est en trait gras, le graphe en trait fin) Montrer que la famille $\{\varphi_f\}_{f\in\mathcal{F}_h}$ est libre.

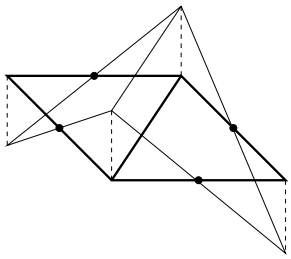


FIGURE 4.17 – Fonction de base pour l'élément fini de Crouzeix–Raviart

4. On introduit l'espace

$$V_h = \{\, v_h \in L^2(\Omega);\ \forall K \in \mathcal{K},\ v_h|_K \in \mathbb{P}_1;\ \forall f \in \mathcal{F}_h,\ [v_h(m_f)] = 0;\ \forall f \in \mathcal{F}_h^0,\ v_h(m_f) = 0 \,\},$$

où $[v_h(m_f)]$ est le saut de v_h en m_f. Montrer que $\{\varphi_f\}_{f\in\mathcal{F}_h}$ est une base de V_h.
5. On cherche une solution approchée du problème (4.3) pour une donnée $f \in L^2(\Omega)$. Pour cela, on considère le problème

$$\begin{cases} \text{Chercher } u_h \in V_h \text{ tel que} \\ \sum_{K\in\mathcal{K}} \int_K \nabla u_h \cdot \nabla v_h = \int_\Omega f v_h, \quad \forall v_h \in V_h. \end{cases} \tag{4.64}$$

Noter qu'on ne peut pas remplacer $\sum_{K\in\mathcal{K}} \int_K$ par $\int_\Omega$ dans le membre de gauche car les fonctions de V_h pouvant être discontinues, leur gradient n'est pas nécessairement de carré sommable sur Ω. On note A la matrice de rigidité associée au problème discret (4.64). Montrer que A est définie positive.
6. Evaluer la matrice A sur le maillage de la figure 4.18 (on numérotera les faces intérieures comme indiqué).

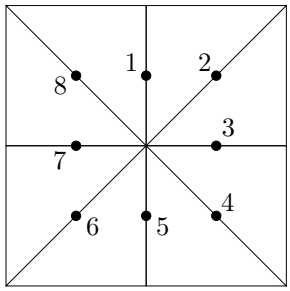


FIGURE 4.18 – Maillage pour l'exercice 10.

7. On considère cette fois un maillage où chaque côté de Ω est subdivisé en N mailles, les mailles carrées ainsi formées étant ensuite découpées en deux triangles selon l'une de leurs diagonales. Quelle est la taille de la matrice de rigidité ? Quel est le nombre maximum d'éléments non-nuls par ligne ?

Exercice 11. *(Élément fini $\mathbb{Q}_1$ en dimension 2)* Soit $\Omega =]\alpha,\beta[\times]\alpha,\beta[$ et considérons deux maillages mono-dimensionnels $\{x_i\}_{0\leqslant i\leqslant N+1}$ et $\{y_j\}_{0\leqslant j\leqslant M+1}$ tels que

$$\alpha = x_0 < x_1 < \ldots x_i \ldots < x_{N+1} = \beta \quad \text{et} \quad \alpha = y_0 < y_1 < \ldots y_j \ldots < y_{M+1} = \beta.$$

On introduit le maillage d'éléments $\{K_{ij}\}_{0\leqslant i\leqslant N, 0\leqslant j\leqslant M}$ définis par $K_{ij} = [x_i, x_{i+1}] \times [y_j, y_{j+1}]$ et on considère l'espace des fonctions polynomiales de degré total $\leqslant 1$,

$$\mathbb{Q}_1 := \{p : \mathbb{R}^2 \to \mathbb{R};\ \exists a,\ b,\ c,\ d \in \mathbb{R},\ p(x,y) = axy + bx + cy + d\}.$$

Une fonction de $\mathbb{Q}_1$ est déterminée de manière unique par la valeur qu'elle prend en les quatre sommets d'un rectangle non dégénéré.

1. On définit l'espace $Q_h^{(1)} := \{v \in C^0(\overline{\Omega});\ v|_{K_{ij}} \in \mathbb{Q}_1,\ 0 \leqslant i \leqslant N,\ 0 \leqslant j \leqslant M,\ v|_{\partial\Omega} = 0\}$. Pour un couple d'indices (i,j), $1 \leqslant i \leqslant N$, $1 \leqslant j \leqslant M$, soit $\varphi_{ij}(x,y) \in Q_h^{(1)}$ la fonction telle que, pour tout $1 \leqslant l \leqslant N$, $1 \leqslant m \leqslant M$,

$$\varphi_{ij}(x_l, y_m) = \begin{cases} 1 & \text{si } (l,m) = (i,j), \\ 0 & \text{sinon.} \end{cases}$$

   Prouver que $\{\varphi_{ij}\}_{1\leqslant i\leqslant N, 1\leqslant j\leqslant M}$ est une base de l'espace $Q_h^{(1)}$. L'espace $Q_h^{(1)}$ est-il un sous-espace vectoriel de $H_0^1(\Omega)$?

2. Pour l'élément de référence $K^* := [0,1]^2$, on définit une base de Lagrange $\{\theta_{lm}\}_{l,m\in\{0,1\}}$ avec $\theta_{lm} : K^* \to \mathbb{R}$ telle que pour $s,t \in \{0,1\}$,

$$\theta_{lm}(s,t) = \begin{cases} 1 & \text{si } (l,m) = (s,t), \\ 0 & \text{sinon.} \end{cases}$$

   Exprimer les fonctions θ_{lm} en fonction de produits de fonctions de forme de l'élément fini de Lagrange $\mathbb{P}_1$ unidimensionnel (que l'on notera θ_0 et θ_1).

3. Soit $f \in L^2(\Omega)$. On considère le problème

$$\begin{cases} \text{Chercher } u_h \in Q_h^{(1)} \text{ tel que} \\ \int_\Omega \nabla u_h \cdot \nabla v_h = \int_\Omega f v_h, \quad \forall v_h \in Q_h^{(1)}. \end{cases} \tag{4.65}$$

   – Exprimer la matrice de rigidité locale $A \in \mathbb{R}^{4,4}$ en utilisant les fonctions θ_{lm} ordonnées de la manière suivante : θ_{00}, θ_{01}, θ_{10}, θ_{11} (on utilisera une double indexation des lignes et des colonnes de A sous la forme $A_{(i_1j_1),(i_2j_2)}$ avec $0 \leqslant i_1,\ j_1,\ i_2,\ j_2 \leqslant 1$).
   – On introduit les matrices de masse $M^{(1)}$ et de rigidité $A^{(1)}$ locales associées à l'élément fini de Lagrange $\mathbb{P}_1$ unidimensionnel. On rappelle que

$$M_{ij}^{(1)} = \int_0^1 \theta_i \theta_j, \quad A_{ij}^{(1)} = \int_0^1 \theta_i' \theta_j', \quad 0 \leqslant i,j \leqslant 1.$$

   Montrer que pour tout couple d'indices (i_1j_1) et (i_2j_2), on a

$$A_{(i_1j_1),(i_2j_2)} = A_{i_1i_2}^{(1)} M_{j_1j_2}^{(1)} + A_{j_1j_2}^{(1)} M_{i_1i_2}^{(1)}, \quad 0 \leqslant i_1, j_1, i_2, j_2 \leqslant 1,$$

   (i_1j_1) et (i_2j_2) étant les indices locaux des fonctions $\theta_{i_1j_1}$ et $\theta_{i_2j_2}$ respectivement (on numérote les lignes et les colonnes des matrices en partant de 0).
   – Calculer les matrices $A^{(1)}$ et $M^{(1)}$ et en déduire la matrice A (Indication : observer que A est symétrique).

4. On introduit l'opérateur d'interpolation

$$I_h^{(1)} : C^0(\overline{\Omega}) \ni v \longmapsto \sum_{1\leqslant i\leqslant N} \sum_{1\leqslant j\leqslant M} v(x_i, y_j)\varphi_{ij}(x,y),$$

et on admet l'estimation d'erreur suivante :

$$\exists c_I, \quad \forall v \in H^2(\Omega) \cap H_0^1(\Omega), \quad \|v - I_h^{(1)} v\|_{H^1(\Omega)} \leqslant c_I h |v|_{H^2(\Omega)},$$

où $h := \max(\max_{0 \leqslant i \leqslant N} |x_{i+1} - x_i|, \max_{0 \leqslant j \leqslant M} |y_{j+1} - y_j|)$. Prouver la convergence de la solution u_h de (4.65) vers la solution du problème continu

$$\begin{cases} \text{Chercher } u \in H_0^1(\Omega) \text{ tel que} \\ \int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v, \quad \forall v \in H_0^1(\Omega). \end{cases} \tag{4.66}$$

Quel est l'ordre de convergence de la méthode en norme H^1 ?

Corrigés

Exercice 1. *(Principe du maximum discret)*

1. On pose $V_0 = V_{N+1} = 0$. Soit $0 \leqslant j \leqslant N+1$ un indice tel que $V_j = \max_{0 \leqslant i \leqslant N+1} V_i$. Si $j = 0$ ou $j = N+1$, la conclusion est immédiate. Si $1 \leqslant j \leqslant N$, on utilise le fait que $AV \leqslant 0$ pour déduire
$$0 \leqslant 2V_j - V_{j+1} - V_{j-1} \leqslant 0,$$
ce qui implique que $V_j = V_{j-1} = V_{j+1}$. De proche en proche, on se ramène à $V_j = V_0$, d'où à nouveau la conclusion.
2. Si $f \leqslant 0$ sur Ω, il vient $B \leqslant 0$ si bien que d'après la question précédente $U \leqslant 0$, d'où $u_h \leqslant 0$ sur Ω.
3. Soit e_i le i-ème vecteur de la base canonique de $\mathbb{R}^N$. Alors, le vecteur $A^{-1}e_i$ a pour composantes $(\alpha_{ij})_{1 \leqslant j \leqslant N}$. Comme $-e_i = A(A^{-1}(-e_i)) \leqslant 0$, il vient $-A^{-1}e_i \leqslant 0$, d'où la conclusion.
4. On introduit le vecteur $W \in \mathbb{R}^N$ de composantes $W_i = w(x_i)$. On constate que $AW = h\sum_{j=1}^N e_j$. D'où, pour tout $1 \leqslant i \leqslant N$,
$$\sum_{j=1}^N \alpha_{ij} \leqslant \frac{1}{h} w(x_i) \leqslant \frac{1}{8h}.$$
Puisque $\alpha_{ij} \geqslant 0$ pour tout $1 \leqslant i, j \leqslant N$, on en déduit
$$u_h(x_i) = \sum_{j=1}^N \alpha_{ij} B_j \leqslant \sum_{j=1}^N \alpha_{ij} |B_j| \leqslant \frac{1}{8h} \max_{1 \leqslant j \leqslant N} |B_j|,$$
et on conclut en observant que
$$|B_j| \leqslant \int_{x_{j-1}}^{x_{j+1}} |f| \varphi_j \leqslant h \|f\|_{L^\infty}.$$

Exercice 2. *(Preuve du théorème 4.9)*

1. On utilise l'inégalité de Cauchy–Schwarz : pour tout $x \in K_i$, on a
$$v(x) - (\mathcal{I}_h^{(1)} v)(x) = \int_{x_i}^x w_i(s) ds,$$
si bien que
$$|v(x) - (\mathcal{I}_h^{(1)} v)(x)| \leqslant \left(\int_{x_i}^x ds \right)^{1/2} \left(\int_{x_i}^x w_i^2(s) ds \right)^{1/2} \leqslant h_i^{1/2} \|w_i\|_{L^2(K_i)}.$$
D'où
$$\|v - \mathcal{I}_h^{(1)} v\|_{L^2(K_i)} \leqslant h_i^{1/2} \|v - \mathcal{I}_h^{(1)} v\|_{L^\infty(K_i)} \leqslant h_i \|w_i\|_{L^2(K_i)}.$$

2. La fonction w_i s'annule sur K_i car les fonctions v et $\mathcal{I}_h^{(1)}v$ coïncident aux deux extrémités de K_i ; le théorème des accroissements finis permet d'affirmer qu'il existe $\xi_i \in K_i$ tel que $v'(\xi_i) = (\mathcal{I}_h^{(1)}v)'(\xi_i)$, c'est-à-dire $w_i(\xi_i) = 0$. On en déduit que pour tout $x \in K_i$,

$$w_i(x) = \int_{\xi_i}^{x} w_i'(s)ds,$$

et en procédant comme ci-dessus, il vient $\|w_i\|_{L^2(K_i)} \leqslant h_i \|w_i'\|_{L^2(K_i)}$.

3. En observant que $w_i = (v - \mathcal{I}_h^{(1)}v)'|_{K_i}$ et que $\|w_i'\|_{L^2(K_i)} = |v|_{H^2(K_i)}$, on obtient

$$\|(v - \mathcal{I}_h^{(1)}v)'\|_{L^2(K_i)} \leqslant h_i |v|_{H^2(K_i)}.$$

En élevant cette inégalité au carré et en sommant sur toutes les mailles, il vient

$$\|(v - \mathcal{I}_h^{(1)}v)'\|_{L^2} \leqslant h|v|_{H^2}.$$

Par ailleurs, en élevant au carré l'estimation obtenue à la question 1 et en sommant sur toutes les mailles, il vient

$$\|v - \mathcal{I}_h^{(1)}v\|_{L^2} \leqslant h\|(v - \mathcal{I}_h^{(1)}v)'\|_{L^2}.$$

D'où la conclusion.

Exercice 3. *(Convection–diffusion)*

1. La formulation faible consiste à

$$\begin{cases} \text{Chercher } u \in H_0^1(\Omega) \text{ tel que} \\ a(u,v) = b(v), \quad \forall v \in H_0^1(\Omega), \end{cases}$$

avec

$$a(u,v) = \int_\Omega \nu u'v' + \int_\Omega \beta u'v \qquad \text{et} \qquad b(v) = \int_\Omega fv.$$

$H_0^1(\Omega)$ équipé de la norme $\|\cdot\|_{H^1}$ est un espace de Hilbert et la forme linéaire b est clairement continue sur $H_0^1(\Omega)$. La forme bilinéaire a est continue avec la constante $\omega = \nu + \beta$ et elle est coercive avec la constante $\alpha = \frac{\nu}{1+c_\Omega^2}$ où c_Ω est la constante intervenant dans l'inégalité de Poincaré (observer que $\int_\Omega u'u = 0$ pour tout $u \in H_0^1(\Omega)$). Le théorème de Lax–Milgram permet d'affirmer que le problème ci-dessus est bien posé.

2. La matrice de rigidité A vaut

$$A = \frac{\nu}{h}\text{tridiag}\left(-1 - \frac{\gamma}{2}, 2, -1 + \frac{\gamma}{2}\right).$$

3. La nouvelle matrice de rigidité A vaut

$$A = \frac{\nu_h}{h}\text{tridiag}\left(-1 - \frac{\beta h}{2\nu_h}, 2, -1 + \frac{\beta h}{2\nu_h}\right),$$

avec $\nu_h = \nu + \frac{1}{2}\beta h$. Comme $\nu_h > \frac{1}{2}\beta h$, les coefficients $A_{i,i+1}$ sont toujours négatifs.

4. Soit $w_h \in V_h^{(1)}$. Comme $\nu \leqslant \nu_h$, il vient

$$\nu|u_h - w_h|_{H^1} \leqslant \nu_h|u_h - w_h|_{H^1} \leqslant \frac{a_h(u_h - w_h, u_h - w_h)}{|u_h - w_h|_{H^1}} \leqslant \sup_{v_h \in V_h^{(1)}} \frac{a_h(u_h - w_h, v_h)}{|v_h|_{H^1}}.$$

Par ailleurs, l'identité

$$a_h(u_h - w_h, v_h) = a(u - w_h, v_h) + a(w_h, v_h) - a_h(w_h, v_h),$$

résulte du fait que $a_h(u_h, v_h) = b(v_h) = a(u, v_h)$.

5. La majoration $|a(y,z)| \leqslant \max(\nu,\beta)\|y\|_{H^1}|z|_{H^1}$ s'obtient en intégrant par parties le terme convectif et en utilisant l'inégalité de Cauchy–Schwarz. Des estimations obtenues à la question précédente, on déduit

$$\nu|u_h - w_h|_{H^1} \leqslant \sup_{v_h \in V_h^{(1)}} \frac{|a(u-w_h,v_h)|}{|v_h|_{H^1}} + \sup_{v_h \in V_h^{(1)}} \frac{|a(w_h,v_h)-a_h(w_h,v_h)|}{|v_h|_{H^1}},$$

d'où la majoration demandée puisque $a(w_h,v_h) - a_h(w_h,v_h) = -\frac{1}{2}\beta h \int_\Omega w_h' v_h'$.

6. En choisissant $w_h = \mathcal{I}_h^{(1)} u$ et en utilisant le fait que $|\mathcal{I}_h^{(1)} u|_{H^1} \leqslant c_0 \|u\|_{H^1}$, il vient de par le théorème d'interpolation 4.9,

$$\nu|u_h - \mathcal{I}_h^{(1)} u|_{H^1} \leqslant c_1 h,$$

avec $c_1 = \max(\nu,\beta) c_{\mathcal{I}^{(1)}} |u|_{H^2} + \frac{1}{2}\beta c_0 |u|_{H^1}$. En utilisant l'inégalité de Poincaré, on obtient

$$\|u_h - \mathcal{I}_h^{(1)} u\|_{H^1} \leqslant c_2 h,$$

avec $c_2 = (1+c_\Omega^2)c_1/\nu$. Enfin, une inégalité triangulaire donne

$$\|u - u_h\|_{H^1} \leqslant \|u_h - \mathcal{I}_h^{(1)} u\|_{H^1} + \|u - \mathcal{I}_h^{(1)} u\|_{H^1},$$

d'où la conclusion en utilisant à nouveau le théorème d'interpolation 4.9.

Exercice 4. *(Élément fini de Hermite)*

1. On applique le théorème de Lax–Milgram.
   – Equipé de la norme $\|v\|_{H^2} = (\|v\|_{L^2}^2 + \|v'\|_{L^2}^2 + \|v''\|_{L^2}^2)^{\frac{1}{2}}$, V est un espace de Hilbert.
   – La forme linéaire $b(v) = \int_\Omega fv$ est clairement continue.
   – La forme bilinéaire $a(u,v) = \int_\Omega u''v''$ est clairement continue.
   – La forme bilinéaire a est coercive. En effet, l'inégalité admise implique que $\forall v \in V$, $\|v''\|_{L^2}^2 \geqslant \beta^2(\|v\|_{L^2}^2 + \|v'\|_{L^2}^2)$; d'où $a(v,v) \geqslant \frac{1}{2}\min(1,\beta^2)\|v\|_{H^2}^2$.
2. L'espace $\mathbb{P}_3$ est de dimension 4. Il suffit donc de montrer que si $p \in \mathbb{P}_3$ est tel que $p(s) = p'(s) = 0$ et $p(t) = p'(t) = 0$ avec $s \neq t$, alors $p \equiv 0$. Or, $p(s) = p'(s) = 0$ implique que $(x-s)^2$ divise p ; de même, $(x-t)^2$ divise p. Comme $s \neq t$ et que p est un polynôme de degré 3 au plus, il vient $p \equiv 0$.
3. Les fonctions de V_h sont continûment différentiables sur $\overline{\Omega}$ et elles sont de classe C^2 par morceaux. La formule des sauts permet d'affirmer que ces fonctions sont dans $H^2(\Omega)$. Par ailleurs, pour tout $v_h \in V_h$, $v_h(a) = v_h'(a) = v_h(b) = v_h'(b) = 0$ si bien que $v_h \in V$.
4. On a

$$\varphi_i = \begin{cases} \theta_1(\frac{x-x_i}{h}) & \text{si } x \in [x_{i-1}, x_{i+1}], \\ 0 & \text{sinon,} \end{cases}$$

et

$$\psi_i = \begin{cases} h\theta_2(\frac{x-x_i}{h}), & \text{si } x \in [x_{i-1}, x_{i+1}], \\ 0 & \text{sinon.} \end{cases}$$

5. On suppose que $\sum_{i=1}^n \alpha_i\varphi_i + \sum_{i=1}^n \beta_i\psi_i \equiv 0$. Alors, en évaluant cette fonction en x_i, il vient $\alpha_i = 0$ et en évaluant la dérivée de cette fonction en x_i, il vient $\beta_i = 0$. La famille $\{\varphi_1,\ldots,\varphi_n,\psi_1,\ldots,\psi_n\}$ est donc libre. Montrons que cette famille est également génératrice de V_h. Soit $v_h \in V_h$. On pose

$$w_h = \sum_{i=1}^n v_h(x_i)\varphi_i + \sum_{i=1}^n v_h'(x_i)\psi_i.$$

Sur chaque maille, les valeurs de v_h et de w_h ainsi que celles de leur dérivée sont les mêmes aux deux extrémités de la maille. D'après la question 2, ces fonctions sont donc identiques. La maille étant quelconque, il en va de même sur tout l'intervalle Ω.

6. La matrice A a une structure bloc
$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$
où les quatre sous-matrices sont d'ordre n et tridiagonales.
7. La matrice $\widetilde{A}$ est hepta-diagonale, ou, plus précisément, bloc-tridiagonale avec des blocs qui sont des matrices d'ordre deux :
$$\widetilde{A} = \begin{pmatrix} \bullet & \bullet & 0 & \dots & 0 \\ \bullet & \bullet & \bullet & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \bullet & \bullet & \bullet \\ 0 & \dots & 0 & \bullet & \bullet \end{pmatrix}$$
où le symbole $\bullet$ représente une matrice d'ordre deux *a priori* non-nulle.

Exercice 5. *(Erreur d'approximation aux nœuds du maillage)*

1. Soit $v \in H_0^1(\Omega)$ et soit $1 \leqslant i \leqslant n$. Puisque $v(a) = v(b) = 0$, on a
$$\int_a^b G_i' v' = \int_a^{x_i} G_i' v' + \int_{x_i}^b G_i' v' = \frac{b - x_i}{b - a}(v(x_i) - v(a)) - \frac{x_i - a}{b - a}(v(b) - v(x_i)) = v(x_i).$$
2. En observant que $G_i \in V_h^{(1)}$ pour tout $1 \leqslant i \leqslant n$ et en utilisant la relation d'orthogonalité de Galerkine, il vient
$$\int_a^b (u - u_h)' G_i' = 0, \qquad \forall 1 \leqslant i \leqslant n.$$
D'où la conclusion.

Exercice 6. *(Éléments finis mixtes)*

1. On obtient
$$a(p,q) = \int_0^1 pq, \qquad \forall (p,q) \in L^2(\Omega) \times L^2(\Omega),$$
$$b(p,v) = -\int_0^1 pv', \qquad \forall (p,v) \in L^2(\Omega) \times H_0^1(\Omega),$$
$$c(v) = -\int_0^1 fv, \qquad \forall v \in H_0^1(\Omega).$$
2. P est de taille N, U de taille $(N-1)$. Le problème discret s'écrit sous la forme matricielle indiquée avec
   - $A \in \mathbb{R}^{N,N}$, $A_{ij} = a(\chi_j, \chi_i) = \int_0^1 \chi_i \chi_j$ où χ_i est la fonction caractéristique de l'intervalle $[(i-1)/N, i/N]$;
   - $B \in \mathbb{R}^{N,N-1}$, $B_{ik} = b(\chi_i, \varphi_k) = -\int_0^1 \chi_i \varphi_k'$ où φ_k est la fonction chapeau qui vaut 1 au sommet $s_k = k/N$ et 0 en les sommets $s_l = l/N$, $l \neq k$;
   - enfin, F est le vecteur de $\mathbb{R}^{N-1}$ de composantes $F_k = \int_0^1 f\varphi_k$.
3. $A = hI_N$ où I_N désigne la matrice identité ($h = 1/N$). La matrice B est bidiagonale : $B_{ii} = -1$ et $B_{i+1,i} = 1$ pour tout $1 \leqslant i \leqslant N$, tous les autres termes étant nuls.
4. Il est clair que B est injective et que A^{-1} est définie positive. Soit $U \in \mathbb{R}^{N-1}$. On constate que $(U, B^T A^{-1} BU)_{\mathbb{R}^{N-1}} = ((BU), A^{-1}(BU))_{\mathbb{R}^{N-1}} \geqslant 0$, ce qui prouve que $B^T A^{-1} B$ est une matrice positive. Comme A^{-1} est définie positive, $(U, B^T A^{-1} BU)_{\mathbb{R}^{N-1}} = 0$ implique $BU = 0$ et donc $U = 0$ puisque B est injective. $B^T A^{-1} B$ est donc définie positive. Soit P et U les composantes d'un vecteur du noyau de la matrice bloc. On a $AP + BU = 0$ et $B^T P = 0$. En éliminant P on obtient $B^T A^{-1} BU = 0$ et donc $U = 0$ puisque $B^T A^{-1} B$ est définie positive. Il en résulte que $AP = 0$ et donc que $P = 0$ puisque A est inversible. Cela démontre l'injectivité et donc l'inversibilité de la matrice bloc.

Exercice 7. *(Estimation d'erreur en norme L^2)*

1. On a
$$\begin{aligned}\|u-u_h\|_{L^2}^2 &= \int_\Omega (u-u_h)(u-u_h)\\ &= \int_\Omega \nabla\zeta\cdot\nabla(u-u_h)\\ &= \int_\Omega \nabla(u-u_h)\cdot\nabla(\zeta-\mathcal{I}_h^{(1)}\zeta),\end{aligned}$$
où on a utilisé la définition de ζ et la relation d'orthogonalité de Galerkine afin de retrancher l'interpolé $\mathcal{I}_h^{(1)}\zeta$ de ζ.
2. En utilisant l'inégalité de Cauchy–Schwarz et le théorème d'interpolation 4.22, il vient
$$\begin{aligned}\|u-u_h\|_{L^2}^2 &\leqslant \|\nabla(u-u_h)\|_{L^2}\|\nabla(\zeta-\mathcal{I}_h^{(1)}\zeta)\|_{L^2}\\ &\leqslant \|u-u_h\|_{H^1} c_{\mathcal{I}^{(1)}}\sigma_0 h|\zeta|_{H^2}\\ &\leqslant \|u-u_h\|_{H^1} c_{\mathcal{I}^{(1)}}\sigma_0 h\chi_\Omega\|u-u_h\|_{L^2}.\end{aligned}$$
D'où
$$\begin{aligned}\|u-u_h\|_{L^2} &\leqslant (c_{\mathcal{I}^{(1)}}\sigma_0\chi_\Omega)h\|u-u_h\|_{H^1}\\ &\leqslant (cc_{\mathcal{I}^{(1)}}\sigma_0\chi_\Omega)h^2.\end{aligned}$$

Exercice 8. *(Coordonnées barycentriques)*

1. L'expression d'une coordonnée barycentrique est de la forme $\alpha q(x,y)$ où $\alpha\in\mathbb{R}$ et $q(x,y)=0$ est l'équation de la droite sur laquelle s'annule la coordonnée barycentrique. Par exemple, la coordonnée barycentrique λ_1 s'annule sur l'arête du triangle reliant les sommets s_2 et s_3, si bien que $q(x,y)=x+y-h$. Le coefficient α se détermine en imposant $\lambda_1(s_1)=1$, ce qui donne $\alpha=-1/h$. D'où
$$\lambda_1(x,y)=\frac{1}{h}(h-x-y).$$
De même,
$$\lambda_2(x,y)=\frac{x}{h},\qquad \lambda_3(x,y)=\frac{y}{h}.$$
2. Les gradients des coordonnées barycentriques sont donnés par
$$\nabla\lambda_1=\frac{1}{h}\begin{pmatrix}-1\\-1\end{pmatrix},\qquad \nabla\lambda_2=\frac{1}{h}\begin{pmatrix}1\\0\end{pmatrix},\qquad \nabla\lambda_3=\frac{1}{h}\begin{pmatrix}0\\1\end{pmatrix}.$$
Les 3 normales sortantes de K sont données par
$$n_{K,s_1}=\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\end{pmatrix},\qquad n_{K,s_2}=\begin{pmatrix}-1\\0\end{pmatrix},\qquad n_{K,s_3}=\begin{pmatrix}0\\-1\end{pmatrix},$$
et les longueurs des trois hauteurs par
$$d_{K,s_1}=\frac{h}{\sqrt{2}},\qquad d_{K,s_2}=h,\qquad d_{K,s_3}=h,$$
ce qui permet de vérifier la formule (4.59). La somme des gradients des trois coordonnées barycentriques est nulle car par construction, $\lambda_1+\lambda_2+\lambda_3\equiv 1$, si bien qu'en prenant le gradient $\nabla\lambda_1+\nabla\lambda_2+\nabla\lambda_3\equiv 0$.
3. La nullité de la somme des coefficients de chaque ligne et de chaque colonne de la matrice $\mathcal{E}$ est due au fait que $\nabla\lambda_1+\nabla\lambda_2+\nabla\lambda_3\equiv 0$. Par ailleurs, un calcul direct donne
$$\mathcal{E}_{11}=\int_K|\nabla\lambda_1|^2=\frac{h^2}{2}\frac{2}{h^2}=1,$$

et

$$\mathcal{E}_{22} = \int_K |\nabla \lambda_2|^2 = \frac{h^2}{2}\frac{1}{h^2} = \frac{1}{2}.$$

Pour des raisons de symétrie, on a

$$\mathcal{E} = \begin{pmatrix} a & b & b \\ b & c & d \\ b & d & c \end{pmatrix},$$

avec $a = 1$ et $c = \frac{1}{2}$ d'après les calculs ci-dessus. Par ailleurs, l'argument de nullité de la somme des lignes et des colonnes de $\mathcal{E}$ implique que $a + 2b = 0$ et $b + c + d = 0$. D'où

$$\mathcal{E} = \frac{1}{2}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

Exercice 9. *(Projection L^2)*

1. Soit $X \in \mathbb{R}^N$ de composantes $(X_i)_{1\leqslant i\leqslant N}$. Posons $\xi = \sum_{i=1}^N X_i\varphi_i$. Il vient
$$(MX, X)_{\mathbb{R}^N} = \int_\Omega |\xi|^2,$$
d'où l'on déduit immédiatement que M est définie positive.
2. Le vecteur V a pour composantes $V_i = \int_\Omega v\varphi_i$ pour tout $1 \leqslant i \leqslant N$. Une fois calculé le vecteur X, on a $\Pi_h v = \sum_{i=1}^N X_i\varphi_i$.
3. Soit K une maille et $\{\lambda_i\}_{1\leqslant i\leqslant 3}$ les coordonnées barycentriques de K. En utilisant la quadrature de degré $k_g = 2$ (qui est exacte puisque $\lambda_i\lambda_j \in \mathbb{P}_2$), on obtient
$$\int_K \lambda_i^2 = \frac{|K|}{6}, \qquad i \in \{1, 2, 3\},$$
et
$$\int_K \lambda_i\lambda_j = \frac{|K|}{12}, \qquad i, j \in \{1, 2, 3\},\ i \neq j,$$
où $|K| = \frac{1}{8}$ est la surface de K. La matrice M est tridiagonale et d'ordre 9. On pose $\gamma = |K|/12 = 1/96$. Sur la sous-diagonale et la sur-diagonale de M les coefficients sont tous égaux à 2γ. Sur la diagonale, on a successivement 16γ et 8γ.

Exercice 10. *(Élément fini de Crouzeix–Raviart)*

1. Les trois $\{m_i\}_{1\leqslant i\leqslant 3}$ étant non-alignés, un polynôme de $\mathbb{P}_1$ est uniquement déterminé par la valeur qu'il prend en ces trois points.
2. On obtient $\theta_1(x, y) = 2(x + y) - 1$, $\theta_2(x, y) = 1 - 2x$, $\theta_3(x, y) = 1 - 2y$. La matrice $\mathcal{E}$ vaut
$$\mathcal{E} = \begin{pmatrix} 4 & -2 & -2 \\ -2 & 2 & 0 \\ -2 & 0 & 2 \end{pmatrix}.$$
3. Supposons que la fonction $w = \sum_{f\in\mathcal{F}_h} \alpha_f\varphi_f$ soit identiquement nulle. Alors, pour tout $f \in \mathcal{F}_h$, $w(m_f) = \alpha_f = 0$. La famille $\{\varphi_f\}_{f\in\mathcal{F}_h}$ est donc libre.
4. Il est clair que pour tout $f \in \mathcal{F}_h$, $\varphi_f \in V_h$. Soit $v_h \in V_h$ et posons
$$w_h = \sum_{f\in\mathcal{F}_h} v_h(m_f)\varphi_f.$$
Sur chaque triangle $K \in \mathcal{K}$, v_h et w_h appartiennent à $\mathbb{P}_1$ et leur valeur aux trois milieux des arêtes coïncident. D'après la question 1, ces deux fonctions sont donc égales sur K. Le triangle K étant quelconque, elles sont égales sur Ω.

5. Numérotons les faces intérieures du maillage sous la forme $\{f_1, \ldots, f_N\}$. Soit $X \in \mathbb{R}^N$. Posons $\xi_h = \sum_{i=1}^N X_i \varphi_{f_i}$. Alors, il est clair que

$$(AX, X)_{\mathbb{R}^N} = \sum_{K \in \mathcal{K}} \int_K |\nabla \xi_h|^2 \geqslant 0.$$

De plus, $(X, AX)_{\mathbb{R}^N} = 0$ implique que ξ_h est constant sur chaque triangle ; ξ_h est alors nul sur les triangles ayant une arête sur la frontière par construction de V_h. On propage le résultat de triangle en triangle en utilisant la continuité en m_f ; d'où $\xi_h = 0$ sur Ω et donc $X = 0$.

6. La matrice de rigidité est d'ordre 8 et donnée par

$$A = \begin{pmatrix} 4 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & -2 \\ -2 & 8 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & 4 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & 8 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 & 4 & -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 & 8 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -2 & 4 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & 8 & -2 \\ -2 & 0 & 0 & 0 & 0 & 0 & 0 & -2 & 4 \end{pmatrix}.$$

7. La matrice de rigidité est d'ordre $N^2 + 2N(N-1)$, le nombre de faces intérieures du maillage. Il y a au plus 5 éléments non-nuls par ligne.

Exercice 11. *(Élément fini $\mathbb{Q}_1$ en dimension 2)*

1. Il est clair que la famille $\{\varphi_{ij}\}_{1 \leqslant i \leqslant N, 1 \leqslant j \leqslant M}$ est libre, car

$$\sum_{1 \leqslant i \leqslant N} \sum_{1 \leqslant j \leqslant M} \xi_{ij} \varphi_{ij}(x) \equiv 0$$

implique $\xi_{ij} = 0$ pour tout $1 \leqslant i \leqslant N$, $1 \leqslant j \leqslant M$ (évaluer l'expression ci-dessus en un sommet intérieur (i, j) générique). Par ailleurs, soit $v_h \in Q_h^{(1)}$ et notons que la restriction de v_h à un élément générique K_{ij} du maillage est définie uniquement par les valeurs de v_h aux quatre sommets de la maille (x_i, y_j), (x_{i+1}, y_j), (x_i, y_{j+1}) et (x_{i+1}, y_{j+1}). Considérons la fonction w_h définie par

$$w_h(x, y) = \sum_{1 \leqslant i \leqslant N} \sum_{1 \leqslant j \leqslant M} v_h(x_i, y_j) \varphi_{ij}(x).$$

Puisque, sur chaque maille K_{ij}, v_h et w_h coïncident aux quatre sommets (x_i, y_j), (x_{i+1}, y_j), (x_i, y_{j+1}) et (x_{i+1}, y_{j+1}), elles coïncident sur toute la maille. Enfin, il est clair que $\mathbb{Q}_1 \subset H_0^1(\Omega)$.

2. On obtient $\theta_{lm}(s, t) = \theta_l(s)\theta_m(t)$ pour $l, m \in \{0, 1\}$ (où $\{\theta_l\}_{l \in \{0,1\}}$ sont les fonctions de forme locales pour l'élément fini de Lagrange $\mathbb{P}_1$ en dimension 1).

3. – La matrice de rigidité locale A est donnée par

$$A = \begin{pmatrix} \int_{K^*} \nabla\theta_{00} \cdot \nabla\theta_{00} & \int_{K^*} \nabla\theta_{00} \cdot \nabla\theta_{01} & \int_{K^*} \nabla\theta_{00} \cdot \nabla\theta_{10} & \int_{K^*} \nabla\theta_{00} \cdot \nabla\theta_{11} \\ \int_{K^*} \nabla\theta_{01} \cdot \nabla\theta_{00} & \int_{K^*} \nabla\theta_{01} \cdot \nabla\theta_{01} & \int_{K^*} \nabla\theta_{01} \cdot \nabla\theta_{10} & \int_{K^*} \nabla\theta_{01} \cdot \nabla\theta_{11} \\ \int_{K^*} \nabla\theta_{10} \cdot \nabla\theta_{00} & \int_{K^*} \nabla\theta_{10} \cdot \nabla\theta_{01} & \int_{K^*} \nabla\theta_{10} \cdot \nabla\theta_{10} & \int_{K^*} \nabla\theta_{10} \cdot \nabla\theta_{11} \\ \int_{K^*} \nabla\theta_{11} \cdot \nabla\theta_{00} & \int_{K^*} \nabla\theta_{11} \cdot \nabla\theta_{01} & \int_{K^*} \nabla\theta_{11} \cdot \nabla\theta_{10} & \int_{K^*} \nabla\theta_{11} \cdot \nabla\theta_{11} \end{pmatrix}.$$

– Notons que, pour deux couples d'indices (i_1, j_1) et (i_2, j_2),

$$\int_{K^*} \nabla\theta_{i_1 j_1} \cdot \nabla\theta_{i_2 j_2} = \int_0^1 \theta'_{i_1}(s)\theta'_{i_2}(s)\mathrm{d}s \times \int_0^1 \theta_{j_1}(t)\theta_{j_2}(t)\mathrm{d}t + \int_0^1 \theta_{i_1}(s)\theta_{i_2}(s)\mathrm{d}s \times \int_0^1 \theta'_{j_1}(t)\theta'_{j_2}(t)\mathrm{d}t.$$

La matrice A peut donc être obtenue à partir des matrices de masse et de rigidité locales pour l'élément finis Lagrange sur le segment $[0,1]$ en dimension 1 selon la formule proposée.

– On trouve

$$A = \begin{pmatrix} \frac{2}{3} & -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{3} \\ -\frac{1}{6} & \frac{2}{3} & -\frac{1}{3} & -\frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{3} & \frac{2}{3} & -\frac{1}{6} \\ -\frac{1}{3} & -\frac{1}{6} & -\frac{1}{6} & \frac{2}{3} \end{pmatrix}.$$

4. Les hypothèses du Lemme de Céa sont satisfaites. On a donc

$$\|u - u_h\|_{H^1(\Omega)} \leqslant (1 + c_\Omega^2) \inf_{v_h \in Q_h^{(1)}} \|u - v_h\|_{H^1(\Omega)} \leqslant (1 + c_\Omega^2) \|u - I_h^{(1)} u\|_{H^1(\Omega)},$$

c_Ω étant la constante de Poincaré (cf. (3.14)). En outre, Ω étant convexe, la régularité de f garantit que la solution de (4.66) appartient à $H^2(\Omega) \cap H_0^1(\Omega)$. On conclut en utilisant l'estimation d'erreur fournie pour l'opérateur d'interpolation $I_h^{(1)}$. L'ordre de convergence en norme H^1 est égal à 1.

Table des matières

Index