

Première partie

**Introduction à la méthodes
des différences finies**

Introduction

Nous allons présenter dans cette partie les idées de base de la méthode des différences finies qui est sans doute la méthode la plus intuitive, la plus simple et la plus utilisée pour résoudre numériquement des équations aux dérivées partielles (EDP ou edp dans le jargon des spécialistes). En fait, résoudre numériquement une EDP signifie calculer de bonnes approximations de la solution (encore faut-il que le problème ait une solution unique) en un nombre (généralement grand) de points bien répartis sur l'ensemble où elle est définie.

Pour simplifier la présentation, nous ne considérerons presque exclusivement que des équations en dimension 1 d'espace, une variable temporelle pouvant éventuellement s'y rajouter. Des exercices seront néanmoins proposés en dimensions supérieures. Nous nous intéresserons à la résolution des trois types classiques d'EDP : elliptique, parabolique et hyperbolique. Chacune de ces classes d'équations possèdent des propriétés particulières que nous rappellerons brièvement. Commençons par indiquer quelques problèmes modèles pour ces trois classes : *Ne faudrait-il pas mettre néanmoins quelques pistes pour la dimension supérieure ? On pourrait rajouter une phrase du style :*

Équations elliptiques :

Le problème modèle est ici l'équation de Poisson :

$$(1) \quad -u''(x) = f(x) \quad \forall x \in (0, 1)$$

où f est une fonction continue sur l'intervalle $[0, 1]$ à valeurs dans \mathbb{R} et on cherche une fonction u aussi régulière que possible qui satisfait (1).

Pour résoudre (1), il faut lui associer des *conditions aux limites*. Nous étudierons en détails le cas où (1) est associé à des conditions de Dirichlet *homogènes* :

$$(2) \quad u(0) = u(1) = 0.$$

Plus généralement on peut prescrire les valeurs de $u(0)$ et $u(1)$, ce sont les conditions de Dirichlet non-homogènes :

$$(3) \quad u(0) = \alpha \quad u(1) = \beta ,$$

où α et β sont deux nombres réels donnés.

Il existe aussi des *conditions de Neumann* où ce sont les dérivées qui sont prescrites sur le bord de l'intervalle :

$$(4) \quad u'(0) = \alpha \quad u'(1) = \beta ,$$

ou encore des *conditions mixtes*, par exemple :

$$(5) \quad u'(0) = \alpha \quad u(1) = \beta ,$$

Dans le premier chapitre, nous étudierons les propriétés du problème de Dirichlet (1)-(2) : formule (quasi-)explicite pour la solution, principe du maximum, caractérisation variationnelle, . . . etc. Puis nous montrerons comment résoudre numériquement (1)-(2) par la méthode des différences finies. Nous introduirons à cette occasion les notions de grille (ou maillage), de pas de discrétisation, de discrétisation de l'équation, de schémas numériques, de consistance et de monotonie. Nous discuterons également des propriétés de convergence de certains schémas. L'ouvrage essayant d'être aussi "self-contained" que possible, nous indiquerons un certain nombre de méthodes de résolutions des systèmes linéaires auxquels conduisent inévitablement les différents schémas. Bien entendu, des références seront proposées pour que le lecteur puisse en découvrir d'autres. Les exercices proposés seront l'occasion de généraliser ou d'approfondir quelques notions (en particulier en analyse fonctionnelle) ou de découvrir les propriétés de l'équation (1) associée à d'autres conditions aux limites.

Équations hyperboliques : L'exemple modèle est ici celui de l'équation de transport :

$$(6) \quad \frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0 \quad \text{dans } \mathbb{R} \times]0, T[$$

où la vitesse de transport $c \in \mathbb{R}$ et le temps $T > 0$ sont donnés. On cherche une solution u qui est une fonction de $x \in \mathbb{R}$ et $t \in [0, T[$, à valeurs dans \mathbb{R} .

Il s'agit d'une équation d'évolution et, pour pouvoir la résoudre, il faut lui associer une *donnée initiale* qui décrit la fonction à l'instant initial supposé ici être $t = 0$. Cette donnée initiale s'écrit :

$$(7) \quad u(x, 0) = u_0(x) \quad \text{dans } \mathbb{R},$$

la fonction u_0 sera supposée au moins continue (pour des généralisations, voir les exercices du chapitre).

Dans le deuxième chapitre, nous étudierons les propriétés de l'équation (6) : méthode des caractéristiques, vitesse finie de propagation, . . . etc. Puis nous montrerons comment résoudre (6)-(7) par la méthode des différences finies. Nous introduirons les notions de schémas stable, monotone, implicite, explicite et celles de dissipation et de dispersion. Les propriétés de convergence de certains schémas seront étudiées, d'autres vues en exercice. Les exercices proposeront aussi l'étude de l'équation des ondes, de la résolution d'autres équations par la méthode des caractéristiques, . . . etc.

Équations paraboliques : L'exemple modèle est ici celui de l'équation de la chaleur :

$$(8) \quad \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{dans } \mathbb{R} \times (0, T)$$

à laquelle on associe la donnée initiale :

$$(9) \quad u(x, 0) = u_0(x) \text{ dans } \mathbb{R},$$

où u_0 est ici aussi une fonction continue donnée, satisfaisant certaines conditions de croissance à l'infini. Dans le troisième chapitre seront étudiées là encore d'une part les propriétés de l'équation puis celles de quelques schémas numériques envisagés pour résoudre numériquement (8)-(9). Compte tenu des analogies avec les équations hyperboliques, nous insisterons plus particulièrement sur les différences -vitesse infinie de propagation en particulier-.

L'intérêt de ces trois équations est de posséder des solutions explicites (et simples) au moins si les données ont une assez grande régularité. Cela permet de tester la validité des méthodes numériques proposées, en particulier leurs précisions. Testées sur es exemples simples, les méthodes peuvent être ensuite adaptées pour traiter des cas plus complexes. [Le choix de se limiter à la dimension un pour l'espace pour cette partie est motivé par le désir de ne pas obliger le lecteur à ce plonger dans un premier temps dans la théorie des distributions et l'analyse fonctionnelle "pure et dure"]

Chapitre 1

Equations Elliptiques

1.1 Le problème de Dirichlet : étude théorique

On s'intéresse ici à l'équation (1) avec la condition de Dirichlet homogène (2) et, comme nous l'avons annoncé dans l'introduction, nous allons calculer une solution explicite de ce problème (en fait, LA solution de ce problème).

Si u est une solution de (1)-(2), on commence par intégrer (1) de 0 à $y \in (0, 1)$:

$$u'(0) - u'(y) = \int_0^y f(t) dt ,$$

puis on ré-intègre de 0 à $x \in (0, 1)$:

$$(1.1) \quad u'(0)x - u(x) = \int_0^x \int_0^y f(t) dt dy ,$$

où, au passage, on a utilisé le fait que $u(0) = 0$. Reste à calculer $u'(0)$, ce qui est immédiat en faisant $x = 1$ dans (1.1) et en tenant compte du fait que $u(1) = 0$:

$$u'(0) = \int_0^1 \int_0^y f(t) dt dy .$$

On obtient finalement :

$$(1.2) \quad u(x) = \left(\int_0^1 \int_0^y f(t) dt dy \right) x - \int_0^x \int_0^y f(t) dt dy .$$

En intégrant par parties les deux termes du membre de droite de (1.2), on peut réécrire cette formule sous la forme :

$$(1.3) \quad u(x) = \int_0^1 (1-y)xf(y) dy - \int_0^x (x-y)f(y) dy .$$

On retrouve ainsi la forme générale des solutions d'équations du second ordre, dites de *Sturm-Liouville*, qui conduirait à écrire de manière

plus générale :

$$u(x) = \int_0^1 K(x, y) f(y) \, dy$$

où K est un “noyau” positif et symétrique (voir ([6])). Dans notre cas, il est défini par :

$$K(x, y) = \begin{cases} (1-x)y & \text{si } y \leq x \\ (1-y)x & \text{sinon} \end{cases}$$

1.1.1 Quelques propriétés des solutions du problème de Dirichlet

Commençons par énoncer les propriétés élémentaires qui découlent des formules (1.2) ou (1.3).

Proposition 1.1.

1. Pour toute fonction $f \in \mathcal{C}([0, 1])$, il existe une unique solution $u \in \mathcal{C}^2([0, 1])$ du problème de Dirichlet homogène (1)-(2).

2. Si $f \in \mathcal{C}^k([0, 1])$ pour $k \geq 1$, la solution u de (1)-(2) est de classe \mathcal{C}^{k+2} sur $[0, 1]$ et :

$$\|u^{(l)}\|_\infty \leq \|f^{(l-2)}\|_\infty \quad \text{pour } 2 \leq l \leq k+2.$$

3. Si $g \in \mathcal{C}([0, 1])$ et si v est l'unique solution de (1)-(2) associée au second membre g , on a :

- Si $f \leq g$ sur $[0, 1]$ alors $u \leq v$ sur $[0, 1]$,
- $\|u - v\|_\infty \leq \frac{1}{8} \|f - g\|_\infty$,
- $\|u' - v'\|_\infty \leq \frac{3}{2} \|f - g\|_\infty$.

Preuve :

1. L'existence, l'unicité et la régularité de u découlent de l'expression trouvée précédemment.

2. C'est encore une conséquence de (1.2), l'inégalité s'obtenant soit par (1.2), soit par l'équation (dérivée un certain nombre de fois) pour les dérivées d'ordre supérieures à 2.

3. La linéarité de l'équation entraîne que la fonction $w = u - v$ est la solution de l'équation associée à la fonction $h = f - g$. On utilise alors l'égalité (1.3) pour w ,

$$w(x) = \int_0^x [(1-y)x - (x-y)]h(y) \, dy + \int_x^1 (1-y)xh(y) \, dy$$

ou encore

$$w(x) = \int_0^x [(1-x)y]h(y) \, dy + \int_x^1 (1-y)xh(y) \, dy$$

Comme $(1-x)y \geq 0$ et $(1-y)x \geq 0$ si $x, y \in [0, 1]$ (le noyau K est positif), on déduit que la fonction w a le même signe sur $[0, 1]$ que h . De plus, en utilisant encore la positivité de K :

$$\max_{x \in [0,1]} w(x) \leq \max_{x \in [0,1]} \left(\int_0^x [(1-x)y] dy + \int_x^1 (1-y)x dy \right) \|h\|_\infty$$

Pour obtenir le résultat pour w' , on dérive l'équivalent de l'égalité (1.2) obtenue pour w et on obtient :

$$w'(x) = \int_0^1 \int_0^y h(t) dt dy - \int_0^x h(t) dt$$

D'où :

$$|w'(x)| \leq \int_0^1 \int_0^y \|h\|_\infty dt dy + \int_0^x \|h\|_\infty dt \leq \frac{3}{2} \|h\|_\infty$$

La première propriété du point 3) est un cas particulier d'un résultat plus général : le *principe du maximum* que l'on énonce maintenant.

Théorème 1.1. Soient u et v deux fonctions de $\mathcal{C}^2(]0, 1[) \cap \mathcal{C}([0, 1])$ satisfaisant :

1. $-u'' \leq f$ sur $(0, 1)$ (on dit que u est sous-solution de (1))
2. $-v'' \geq g$ sur $(0, 1)$ (on dit que v est sur-solution de (1))
3. $u(0) \leq v(0)$ et $u(1) \leq v(1)$.

Alors, si les seconds membres f et g satisfont $f \leq g$ sur $[0, 1]$, on a $u \leq v$ sur $[0, 1]$.

Preuve : On pourrait⁽¹⁾ faire une preuve basée sur l'expression explicite de la solution comme précédemment mais on va faire une preuve plus générale qui fonctionnerait même dans le cas où cette stratégie est inenvisageable (équation avec des coefficients non constants, par exemple).

Pour montrer que $u \leq v$ sur $[0, 1]$, on considère $M = \max_{x \in [0,1]} (u - v)$ et on va prouver que $M \leq 0$.

On commence par supposer que l'on a $f < g$ sur $[0, 1]$, donc une hypothèse un peu plus forte que celle du Théorème 1.1. Comme u, v sont continues sur le compact $[0, 1]$, il existe un réel $x_0 \in [0, 1]$ tel que $M = u(x_0) - v(x_0)$. Deux cas se présentent :

- ou bien $x_0 \in \{0, 1\}$ et dans ce cas le résultat est acquis par la 3ème propriété car on sait que $u(0) \leq v(0)$ et $u(1) \leq v(1)$,
- ou bien $0 < x_0 < 1$. Comme x_0 est un point de maximum local de $w = u - v$, on a, par des résultats d'Analyse classique :

$$w'(x_0) = 0 \text{ et } w''(x_0) \leq 0.$$

(1). C'est bien vrai ?

Or, en combinant les propriétés 1 et 2 satisfaites par u et v , on a aussi :

$$-w''(x_0) \leq f(x_0) - g(x_0) < 0.$$

Les deux inégalités précédentes sur $w''(x_0)$ sont incompatibles et donc ce cas-là ne peut pas se produire ; le maximum ne peut être atteint que sur le "bord" de $[0, 1]$ et on a bien $M \leq 0$.

Si on a seulement $f \leq g$ sur $[0, 1]$, il faut se ramener au cas d'une inégalité stricte et on va même le faire en remplaçant f par $f - \alpha$ pour $\alpha > 0$; on aura bien $f(x) - \alpha < g(x)$, pour tout $x \in [0, 1]$. Un calcul élémentaire (ou la formule (1.2)) montre que la fonction $\phi_\alpha(x) = \frac{\alpha}{2}x(1-x)$ est solution du problème avec condition de Dirichlet homogène (1)-(2) avec second membre la fonction constante égale α . Considérons la fonction $\tilde{u}_\alpha = u - \phi_\alpha$: elle satisfait une équation de (1) avec second membre égal à $f - \alpha$ et par ailleurs $\tilde{u}_\alpha(0) = u(0)$ et $\tilde{u}_\alpha(1) = u(1)$. Les arguments précédents montrent donc que

$$\tilde{u}_\alpha(x) \leq v(x) \quad \text{sur } [0, 1]$$

ce qui donne pour tout $\alpha > 0$

$$u(x) - \frac{\alpha}{2}x(1-x) \leq v(x) \quad \text{sur } [0, 1]$$

et on conclut en faisant tendre α vers zéro.

Remarque : Le lecteur pourra constater dans la liste d'exercice que les arguments de la preuve du Théorème 1.1 peuvent être généralisés pour traiter des cas beaucoup plus complexes, y compris en dimension supérieure ; en analysant cette preuve, on voit, en effet, qu'elle ne repose que sur la compacité de $[0, 1]$ (pour que le maximum soit atteint) et les propriétés des dérivées premières et secondes en un point de maximum local. Il s'agit donc de deux résultats basiques qui s'étendent sans difficulté en toutes dimensions. Il est à noter enfin que, par des arguments de perturbation du type de celui que nous avons utilisé dans la preuve (cf. $u \rightarrow \tilde{u}_\alpha$), on peut parfois se passer de la compacité du domaine...

Exercice 1. Soit a est une fonction de classe C^1 sur $[0, 1]$ telle que $a(x) \geq \alpha > 0$ sur $[0, 1]$ et f une fonction continue sur $[0, 1]$.

1. Résoudre explicitement le problème de Dirichlet :

$$\begin{cases} -(a(x)u')' = f & \text{dans }]0, 1[\\ u(0) = u(1) = 0. \end{cases}$$

2. En déduire les propriétés de la solution en fonction de celles de a et f .

3. Montrer que cette équation satisfait le principe du maximum.

(Si v est une sous-solution et w une sursolution (notions à définir), on pourra considérer $\max_{[0,1]} (v(x) - w(x) + \eta \exp(Kx))$ pour $0 < \eta \ll 1$ et pour une constante $K > 0$ bien choisie.)

Exercice 2.

1. Résoudre explicitement le problème de Dirichlet :

$$\begin{cases} -u'' + u = f & \text{dans }]0, 1[\\ u(0) = u(1) = 0. \end{cases}$$

(On pourra utiliser l'équation différentielle ordinaire satisfaite par le couple $(u(t), u'(t))$.)

2. Formuler et prouver le principe du maximum pour cette équation.

Exercice 3. On considère l'équation :

$$-u'' + u = f \quad \text{dans } \mathbb{R},$$

où f est une fonction périodique, au moins continue. Résoudre cette équation en utilisant des développements en séries de Fourier dont on étudiera soigneusement la convergence. (Le lecteur pusillanime pourra commencer par le cas où f est de classe C^1 ou même C^∞ puis refouiller ses anciens cours à la recherche des meilleures hypothèses sur f ...)

1.1.2 L'approche variationnelle du problème de Dirichlet

On va montrer maintenant que la solution u de (1)-(2) possède une caractérisation variationnelle c'est-à-dire que u est l'unique solution d'un problème d'optimisation.

On introduit d'abord l'espace fonctionnel :

$$\mathcal{C}_0^1([0, 1]) = \{v \in \mathcal{C}^1(]0, 1[) \cap \mathcal{C}^0([0, 1]); v(0) = 0, v(1) = 0\}.$$

Il est à noter que cet espace est défini à la fois par une contrainte de régularité (\mathcal{C}^1 dans $]0, 1[$ et continu sur $[0, 1]$) et par la valeur imposée des fonctions en 0 et 1. De plus, $u \in \mathcal{C}_0^1([0, 1])$. Pour toute fonction $v \in \mathcal{C}_0^1([0, 1])$, on pose :

$$J(v) = \frac{1}{2} \int_0^1 [v'(t)]^2 dt - \int_0^1 f(t)v(t) dt$$

La première caractérisation variationnelle de u est donnée par la :

Proposition 1.2. La solution u de (1)-(2) satisfait :

$$J(u) = \min_{w \in \mathcal{C}_0^1([0, 1])} J(w)$$

Ce type de résultat sera important à la fois du point de vue théorique que numérique car il signifie que pour calculer u , on peut envisager de résoudre un problème d'optimisation.

Preuve : Si $w \in \mathcal{C}_0^1([0, 1])$, il existe une fonction h de $\mathcal{C}_0^1([0, 1])$ telle que $w = u + h$. On a alors :

$$J(w) = J(u) + \int_0^1 u'(t)h'(t) dt - \int_0^1 f(t)v(t) dt + \frac{1}{2} \int_0^1 (h'(t))^2 dt.$$

Grâce à la régularité des fonctions et la nullité au bord de h , on peut intégrer par parties pour obtenir :

$$J(w) = J(u) + \int_0^1 (-u''(t) - f(t))h(t) dt + \frac{1}{2} \int_0^1 (h'(t))^2 dt.$$

Comme u est solution de (1) et que $\int_0^1 (h'(t))^2 dt \geq 0$, on obtient bien que $J(w) \geq J(u)$. \square

Remarque : On montre facilement qu'une réciproque partielle est vraie, à savoir que si une fonction $u \in \mathcal{C}^2([0, 1]) \cap \mathcal{C}_0^1([0, 1])$ satisfait :

$$J(u) \leq J(w) \quad \text{pour tout } w \in \mathcal{C}_0^1([0, 1])$$

alors elle est solution ⁽²⁾ du problème (1).

En effet, le calcul précédent montre qu'alors :

$$\int_0^1 (-u''(t) - f(t))h(t) dt + \frac{1}{2} \int_0^1 (h'(t))^2 dt \geq 0 \quad \text{pour tout } h \in \mathcal{C}_0^1([0, 1]).$$

En remplaçant h par αh où $\alpha > 0$, on obtient :

$$\alpha \int_0^1 (-u''(t) - f(t))h(t) dt + \frac{1}{2} \alpha^2 \int_0^1 (h'(t))^2 dt \geq 0 \quad \text{pour tout } h \in \mathcal{C}_0^1([0, 1]),$$

et en divisant par α puis en faisant tendre α vers zéro, on a :

$$\int_0^1 (-u''(t) - f(t))h(t) dt \geq 0 \quad \text{pour tout } h \in \mathcal{C}_0^1([0, 1]).$$

Enfin, on change h en $-h$ ce qui conduit à :

$$\int_0^1 (-u''(t) - f(t))h(t) dt = 0 \quad \text{pour tout } h \in \mathcal{C}_0^1([0, 1]).$$

Cette propriété s'interprète de la manière suivante : la fonction $-u''(t) - f(t)$ en tant que fonction de $L^2([0, 1])$ est orthogonale à $\mathcal{C}_0^1([0, 1])$ qui est dense dans $L^2([0, 1])$ ⁽³⁾ donc $-u''(t) - f(t) = 0$ presque partout et en fait partout puisque la fonction $-u'' - f$ est continue.

(2). ce n'est pas une réciproque complète car on doit supposer un peu plus de régularité sur u , plus précisément u de classe \mathcal{C}^2 .

(3). Se souvenir que l'espace des fonctions \mathcal{C}^∞ , à supports compacts est dense dans $L^2([0, 1])$!

Remarque : On peut préférer démontrer le résultat à la main, au lieu d'invoquer la densité (cela revient à redémontrer cette densité) : on procède par troncature et régularisation pour construire une suite $(h_\varepsilon)_\varepsilon$ de fonctions continues qui converge dans $L^2([0, 1])$ vers la fonction $-u'' - f$. Plus précisément, pour $0 \leq \varepsilon \ll 1$, on pose d'abord $\tilde{h}_\varepsilon = (-u'' - f)\mathbb{1}_{[\varepsilon, 1-\varepsilon]}$ (phase de troncature) et on montre facilement que $\tilde{h}_\varepsilon \rightarrow -u'' - f$ dans $L^2([0, 1])$. Puis on régularise \tilde{h}_ε par convolution, ce qui revient à poser $h_\varepsilon = \tilde{h}_\varepsilon * \rho_\varepsilon$ où $(\rho_\varepsilon)_\varepsilon$ est votre suite d'approximation de l'unité favorite, en ayant soin de prendre ρ_ε à support compact dans $] - \varepsilon, \varepsilon[$ et, pour faire bonne mesure, les ρ_ε de classe \mathcal{C}^∞ .

En fait, $\mathcal{C}_0^1([0, 1])$, bien que commode pour les intégrations par parties, n'est pas le bon espace fonctionnel pour attaquer le problème d'optimisation. Cela provient du fait que rien n'assure qu'une suite minimisante va converger vers un élément de $\mathcal{C}_0^1([0, 1])$. On va donc introduire le bon espace fonctionnel : c'est l'espace $H^1(]0, 1[)$ ou plus exactement pour notre problème $H_0^1(]0, 1[)$.

L'espace $H^1(]0, 1[)$ est constitué des fonctions v de $L^2(]0, 1[)$ pour lesquelles il existe un nombre réel μ et une fonction w de $L^2(]0, 1[)$ tels que :

$$v(x) = \mu + \int_0^x w(t) dt \quad \text{p.p. dans }]0, 1[$$

Proposition 1.3. *Si $v \in H^1(]0, 1[)$ alors le couple (μ, w) est unique.*

Preuve : Etant donné $v \in H^1(]0, 1[)$, supposons l'existence de deux couples (μ, w) et $(\tilde{\mu}, \tilde{w})$ tels que :

$$v(x) = \mu + \int_0^x w(t) dt = \tilde{\mu} + \int_0^x \tilde{w}(t) dt \quad \text{p.p. dans }]0, 1[.$$

En faisant tendre x vers zéro⁽⁴⁾, on voit immédiatement que $\mu = \tilde{\mu}$. Il reste à montrer que la fonction $h(t) = w(t) - \tilde{w}(t)$ est nulle presque partout dans l'intervalle $[0, 1]$. On sait, du fait que $\mu = \tilde{\mu}$, que $\int_0^x h(t) dt = 0$ pour presque tout x . Donc, si φ désigne une fonction continue sur $[0, 1]$, on a :

$$\int_0^1 \varphi(x) \int_0^x h(t) dt dx = 0$$

Mais le théorème de Fubini (dont le lecteur ou la lectrice vérifiera aisément qu'il s'applique) donne :

$$(1.4) \quad \int_0^1 h(t) \int_t^1 \varphi(x) dx dt = 0.$$

(4). Attention tout de même car l'égalité n'est vraie que presque partout...

En invoquant à nouveau la densité de $\mathcal{C}_0^1([0, 1])$ dans $L^2(]0, 1[)$, on peut construire une suite de fonctions ψ_ε de $\mathcal{C}_0^1([0, 1])$ qui convergent vers h dans $L^2(]0, 1[)$. On utilise alors (1.4) avec le choix $\varphi = -\psi'_\varepsilon$ pour tout ε et on obtient que $\int_0^1 \psi_\varepsilon(t)h(t) dt = 0$ ce qui conduit à la limite à $\int_0^1 [h(t)]^2 dt = 0$. On a donc bien $h = 0$. \square

Si v est une fonction de classe $\mathcal{C}^1([0, 1]) \cap \mathcal{C}([0, 1])$, elle est dans l'espace $H^1(]0, 1[)$ avec $\mu = v(0)$ et $w(t) = v'(t)$. Pour une fonction v de $H^1(]0, 1[)$ quelconque, la fonction w s'interprète donc comme une dérivée généralisée que l'on notera v' sans ambiguïté maintenant que l'on sait que cette dérivée est unique.

Un lecteur plus averti saura immédiatement que le théorème de Lebesgue (voir le chapitre (?) de [3]) montre que si $w \in L^1(]0, 1[)$ alors la fonction qui à x associe $\int_0^x w(t) dt$ est dérivable presque partout et de dérivée presque partout égale à w ce qui donne une deuxième justification de la désignation de dérivée généralisée puisque en dimension 1, on sait, grâce à l'inégalité de Cauchy-Schwarz, que l'espace $L^2(]0, 1[)$ est contenu dans $L^1(]0, 1[)$.

On peut munir $H^1(]0, 1[)$ de la norme $\|v\|_1$ définie par :

$$\|v\|_1^2 = \|v\|_{L^2(]0,1])}^2 + \|v'\|_{L^2(]0,1])}^2$$

Proposition 1.4. *Muni de la norme $\|\cdot\|_1$, l'espace $H^1(]0, 1[)$ est un espace de Hilbert dans lequel l'espace $\mathcal{C}^1([0, 1])$ est dense.*

Admettons pour l'instant ce résultat dont la preuve utilisera les techniques ou les résultats ci-dessous et qui concernent la régularité classique des fonctions de $H^1(]0, 1[)$.

Proposition 1.5. *L'injection de l'espace $H^1(]0, 1[)$, muni de la norme $\|\cdot\|_1$ dans l'espace des fonctions höldériennes $\mathcal{C}^{0, \frac{1}{2}}(]0, 1[)$ est continue.*

“Rappel :” Si $0 < \alpha \leq 1$, l'espace $\mathcal{C}^{0, \alpha}([0, 1])$ est l'espace constitué des fonctions continues $v : [0, 1] \rightarrow \mathbb{R}$ qui satisfont la propriété suivante : il existe une constante $C \in \mathbb{R}$ telle que, pour tous $x, y \in [0, 1]$:

$$(1.5) \quad |v(x) - v(y)| \leq C|x - y|^\alpha .$$

En d'autres termes, les fonction de $\mathcal{C}^{0, \alpha}([0, 1])$ sont celles qui ont un module de continuité $\omega_v(t)$ de la forme Ct^α .

On munit $\mathcal{C}^{0, \alpha}([0, 1])$ de la norme $\|\cdot\|_{0, \alpha}$ définie pour $v \in \mathcal{C}^{0, \alpha}([0, 1])$ par :

$$\|v\|_{0, \alpha} = \|v\|_\infty + |v|_{0, \alpha} ,$$

où la semi-norme $|v|_{0, \alpha}$ est donnée par :

$$|v|_{0, \alpha} = \sup_{x \neq y} \frac{|v(x) - v(y)|}{|x - y|^\alpha} .$$

Une autre façon de voir cette semi-norme, c'est comme la plus petite des constantes C telle que la propriété (1.5) soit satisfaite.

Muni de cette norme, $\mathcal{C}^{0,\alpha}([0, 1])$ est un espace de Banach.

Les exos : prouver que la norme est une norme + Banach + autour d'Ascoli (si les $\|v_n\|_{0,\alpha}$ sont bornés on peut extraire une sous-suite convergente)

Preuve : On veut estimer $v(x) - v(y)$. Par un simple calcul :

$$v(x) - v(y) = \int_x^y v'(t) dt = \int_0^1 \mathbb{1}_{\{t \in]x, y[\}} v'(t) dt .$$

On applique alors l'inégalité de Cauchy-Schwarz dans $L^2(]0, 1[)$ qui donne :

$$|v(x) - v(y)| \leq \left(\int_0^1 \mathbb{1}_{\{t \in]x, y[\}}^2 dt \right)^{\frac{1}{2}} \|v'\|_{L^2(]0, 1[)}$$

ou encore

$$(1.6) \quad |v(x) - v(y)| \leq |x - y|^{\frac{1}{2}} \|v'\|_{L^2(]0, 1[)}$$

Ceci montre qu'une fonction de $H^1(]0, 1[)$ peut être vue comme une fonction de $\mathcal{C}^{0,\frac{1}{2}}(]0, 1[)$ et donc que l'injection est bien définie.

Montrons maintenant la continuité de l'injection (qui est évidemment une application linéaire) : la norme qui fait de $\mathcal{C}^{0,\frac{1}{2}}(]0, 1[)$ un espace de Banach (voir le rappel ci-dessus et ([6])) est :

$$\|v\|_{0,\frac{1}{2}} = \|v\|_{\infty} + |v|_{0,\frac{1}{2}}$$

où :

$$|v|_{0,\frac{1}{2}} = \sup_{x \in [0, 1]} \frac{|v(x) - v(y)|}{|x - y|^{\frac{1}{2}}} .$$

Nous devons montrer qu'il existe une constante $K \in \mathbb{R}$ telle que :

$$\|v\|_{0,\frac{1}{2}} \leq K \|v\|_{H^1(]0, 1[)}$$

pour tout $v \in H^1(]0, 1[)$.

Nous remarquons d'abord que l'inégalité (1.6) peut être interprétée comme :

$$|v|_{0,\frac{1}{2}} \leq \|v'\|_{L^2(]0, 1[)} \leq \|v\|_{H^1(]0, 1[)}$$

et donc il ne nous reste plus qu'à estimer $\|v\|_{\infty}$ en fonction de $\|v\|_{H^1(]0, 1[)}$.

Comme $v \in H^1(]0, 1[)$, v s'écrit :

$$v(x) = \mu + \int_0^x v'(t) dt \quad \text{p.p. dans }]0, 1[$$

et en combinant l'inégalité triangulaire avec les arguments développés plus haut, on voit que :

$$|v(x)| \leq |\mu| + \left| \int_0^x v'(t) dt \right| \leq |\mu| + x^{1/2} \|v'\|_{L^2(]0,1])} ,$$

d'où :

$$\|v\|_\infty \leq |\mu| + \|v'\|_{L^2(]0,1])} \leq |\mu| + \|v\|_{H^1(]0,1])} ,$$

Il nous reste à estimer $|\mu|$ pour conclure. On écrit :

$$\mu = v(x) - \int_0^x v'(t) dt ,$$

et on considère cette égalité comme une égalité de fonctions de $L^2(]0,1])$, μ étant vu comme une fonction constante. Par l'inégalité triangulaire pour la norme L^2 :

$$|\mu| = \|\mu\|_{L^2(]0,1])} \leq \|v\|_{L^2(]0,1])} + \left\| \int_0^x v'(t) dt \right\|_{L^2(]0,1])} .$$

Or, on a vu que $\left| \int_0^x v'(t) dt \right| \leq x^{1/2} \|v'\|_{L^2(]0,1])}$ et donc :

$$(1.7) \quad \left\| \int_0^x v'(t) dt \right\|_{L^2(]0,1])} \leq \frac{\|v'\|_{L^2(]0,1])}}{\sqrt{2}} .$$

Finalement :

$$(1.8) \quad |\mu| = \|\mu\|_{L^2(]0,1])} \leq \|v\|_{L^2(]0,1])} + \frac{\|v'\|_{L^2(]0,1])}}{\sqrt{2}} \leq \left(1 + \frac{1}{\sqrt{2}}\right) \|v\|_{H^1(]0,1])}$$

et :

$$\|v\|_\infty \leq \left(2 + \frac{1}{\sqrt{2}}\right) \|v\|_{H^1(]0,1])} .$$

En rassemblant toutes les informations obtenues sur $\|v\|_\infty$ et $|v|_{0, \frac{1}{2}}$, on conclut :

$$(1.9) \quad \|v\|_{0, \frac{1}{2}} \leq \left(3 + \frac{1}{\sqrt{2}}\right) \|v\|_{H^1(]0,1])} .$$

□

Revenons maintenant à la preuve de la Proposition 1.4 :

Preuve : Tout d'abord, la norme définie sur $H^1(]0,1])$ dérive clairement d'un produit scalaire lui-même construit sur le produit scalaire défini sur $L^2(]0,1])$ que nous noterons $(\cdot, \cdot)_{L^2(]0,1])}$:

$$\|v\|_1^2 = (v, v)_{H^1(]0,1])} = (v, v)_{L^2(]0,1])} + (v', v')_{L^2(]0,1])} .$$

Ensuite, il nous faut montrer que $H^1(]0,1])$ est complet pour cette norme (ce qui légitimera l'envie de prouver les injections). Considérons donc une

suite de Cauchy $(v_n)_{n \in \mathbb{N}}$ d'éléments de $H^1(]0, 1[)$. A chaque fonction v_n est associée le réel μ_n et la fonction $w_n = v'_n$. En regardant la preuve ci-dessus et l'inégalité (1.8), on voit que d'une part la suite $(\mu_n)_{n \in \mathbb{N}}$ est de Cauchy dans \mathbb{R} donc converge vers un réel μ . Par ailleurs, l'espace $H^1(]0, 1[)$ s'injecte naturellement dans l'espace $L^2(]0, 1[)$ de façon continue puisque :

$$\|v\|_{L^2(]0,1[)} \leq \|v\|_1$$

et on a aussi :

$$\|v'\|_{L^2(]0,1[)} \leq \|v\|_1$$

On a donc aussi que les suites $(v_n)_{n \in \mathbb{N}}$ d'une part et $(v'_n)_{n \in \mathbb{N}}$ sont de Cauchy dans $L^2(]0, 1[)$ donc elles convergent vers respectivement une fonction $v \in L^2(]0, 1[)$ et $w \in L^2(]0, 1[)$.

En utilisant (1.7), on voit que $\int_0^x v'_n(t) dt$ converge vers $\int_0^x w(t) dt$ dans $L^2(]0, 1[)$ et, quitte à extraire une sous-suite qui converge presque partout [à la fois pour v_n et $\int_0^x v'_n(t) dt$], on peut passer à la limite dans l'égalité :

$$v_n(x) = \mu_n + \int_0^x v'_n(t) dt \quad \text{p.p.}$$

et obtenir :

$$v(x) = \mu + \int_0^x w(t) dt \quad \text{p.p.}$$

ce qui montre bien que $v \in H^1(]0, 1[)$ et donc que $H^1(]0, 1[)$ est complet.

Reste à prouver la densité de $\mathcal{C}^1([0, 1])$ dans $H^1(]0, 1[)$, ce que nous ne ferons pas en détails ici ⁽⁵⁾. \square

Définition 1.1. On désigne par $H_0^1(]0, 1[)$ le sous-espace fermé de $H^1(]0, 1[)$ des fonctions qui s'annulent en $x = 0$ et en $x = 1$.

Remarque : Il est d'abord important de remarquer que $H_0^1(]0, 1[)$ est bien défini car les fonctions de $H^1(]0, 1[)$ peuvent être vue comme des fonctions continues à cause de l'injection de $H^1(]0, 1[)$ dans $\mathcal{C}^{0, \frac{1}{2}}(]0, 1[)$, ce qui donne un sens à leurs valeurs en 0 et 1 (alors que la définition initiale de $H^1(]0, 1[)$ ne définissait les fonctions que presque partout...). Le fait que $H_0^1(]0, 1[)$ soit fermé est un corollaire de la continuité de cette même injection.

Nous admettrons dans la suite le résultat suivant :

Théorème 1.2. L'espace $\mathcal{C}_0^1([0, 1])$ est dense dans $H_0^1(]0, 1[)$ au sens de la norme $H^1(]0, 1[)$.

La preuve de ce résultat n'est pas si délicate : comme pour la densité de $\mathcal{C}^1([0, 1])$ dans $H^1(]0, 1[)$ (preuve que nous n'avons pas faite non plus!), on

(5). car c'est un excellent exercice pour le lecteur! Indication : utiliser la méthode de troncature et régularisation sur v' et ajoutez une pincée de (1.7).

doit tronquer et régulariser v' mais, ici, on a une contrainte supplémentaire du type $\int_0^1 v'(t) dt = 0$ qu'il faut préserver et même étendre pour les régularisées à un intervalle du type $[\varepsilon, 1 - \varepsilon]$ pour $0 < \varepsilon \ll 1$. Sans être inatteignable, cette preuve nécessite un petit bagage technique...

Le résultat principal de cette section est le :

Théorème 1.3. *Si u est la solution du problème (1) avec la condition de Dirichlet homogène (2), on a :*

$$J(u) = \min_{w \in H_0^1(]0,1[)} J(w).$$

De plus, toute suite minimisante d'éléments de $H_0^1(]0,1[)$ converge vers u qui est l'unique solution du problème de minimisation de la fonctionnelle J dans $H_0^1(]0,1[)$.

Preuve : La première propriété provient simplement de la densité de l'espace $C_0^1(]0,1[)$ dans $H_0^1(]0,1[)$ et de la continuité de J .

La deuxième partie nécessite le résultat suivant :

Lemme 1.1. *(Inégalité de Poincaré) Pour tout élément $w \in H_0^1(]0,1[)$, on a :*

$$\|w\|_{H^1(]0,1[)} \leq \left(1 + \frac{1}{\sqrt{2}}\right) \|w'\|_{L^2(]0,1[)}$$

Une autre manière d'interpréter (ou même de formuler) ce résultat, c'est de dire que, sur $H_0^1(]0,1[)$, l'application $w \mapsto \|w'\|_{L^2(]0,1[)}$ est une norme qui est équivalente à la norme $H^1(]0,1[)$, puisque on a clairement :

$$\|w'\|_{L^2(]0,1[)} \leq \|w\|_{H^1(]0,1[)} \leq \left(1 + \frac{1}{\sqrt{2}}\right) \|w'\|_{L^2(]0,1[)}.$$

On notera $\|w\|_{H_0^1(]0,1[)} = \|w'\|_{L^2(]0,1[)}$ et ce sera désormais la norme que nous utiliserons sur $H_0^1(]0,1[)$. Une dernière conséquence du lemme (ou plus exactement de sa preuve) est que l'on a aussi :

$$(1.10) \quad \|w\|_{L^2(]0,1[)} \leq \frac{1}{\sqrt{2}} \|w\|_{H_0^1(]0,1[)}.$$

Preuve du Lemme : Si $w \in H_0^1(]0,1[)$, puisque $w(0) = 0$, on a (pour tout $x \in [0, 1]$ car w est continu) :

$$w(x) = \int_0^x w'(t) dt$$

et donc, par Cauchy-Schwarz :

$$|w(x)| \leq x^{\frac{1}{2}} \|w'\|_{L^2(]0,1[)}.$$

On en déduit :

$$\|w\|_{L^2(]0,1])} \leq \frac{1}{\sqrt{2}} \|w\|_{H_0^1(]0,1])}.$$

Et le résultat du lemme est ainsi établi. \square

Retour sur la preuve du Théorème : commençons par montrer que J est minorée sur $H_0^1(]0,1])$:

$$J(w) = \frac{1}{2} \|w\|_{H_0^1(]0,1])}^2 - \int_0^1 f(t)w(t) dt$$

D'où, par Cauchy-Schwarz :

$$J(w) \geq \frac{1}{2} \|w\|_{H_0^1(]0,1])}^2 - \|f\|_{L^2(]0,1])} \|w\|_{L^2(]0,1])}$$

et en utilisant (1.10) :

$$J(w) \geq \frac{1}{2} \|w\|_{H_0^1(]0,1])}^2 - \|f\|_{L^2(]0,1])} \frac{\|w\|_{H_0^1(]0,1])}}{\sqrt{2}}$$

Or $ab \leq \frac{\lambda a^2}{2} + \frac{1}{2\lambda} b^2$ pour tout $\lambda > 0$ donc :

$$J(w) \geq \frac{1}{2} (1 - \lambda) \|w\|_{H_0^1(]0,1])}^2 - \frac{1}{4\lambda} \|f\|_{L^2(]0,1])}.$$

En choisissant $\lambda = 1$, on a la minoration souhaitée.

Soit maintenant $(w_\varepsilon)_\varepsilon$ une suite minimisante pour J . En utilisant l'inégalité précédente avec $\lambda = \frac{1}{2}$, on voit que $(w_\varepsilon)_\varepsilon$ est bornée dans $H_0^1(]0,1])$ qui est un espace de Hilbert. Donc il existe une sous-suite $(w'_{\varepsilon'})_{\varepsilon'}$ qui converge faiblement dans $H_0^1(]0,1])$ vers w_∞ . Comme J est convexe et continue, elle est sci pour la topologie faible de $H_0^1(]0,1])$ et donc :

$$J(w_\infty) \leq \underline{\lim} J(w_{\varepsilon'}) = \min_{v \in H_0^1(]0,1])} J(v).$$

Il en résulte que w_∞ est une solution du problème d'optimisation.

Lemme 1.2. *La fonction J est strictement convexe. Plus précisément, pour tout couple $(w_1, w_2) \in H_0^1(]0,1]) \times H_0^1(]0,1])$ et tout réel $\alpha \in [0, 1]$, on a :*

$$J(\alpha w_1 + (1 - \alpha)w_2) - \alpha J(w_1) - (1 - \alpha)J(w_2) = -\frac{\alpha(1 - \alpha)}{2} \|w_1 - w_2\|_{H_0^1(]0,1])}^2$$

Le lemme résulte d'un simple calcul et du caractère quadratique de J .

Comme conséquence si w_1 et w_2 sont deux solutions distinctes du problème d'optimisation et si on pose $m = \min_{v \in H_0^1(]0,1[)} J(v)$ et en prenant $\alpha = \frac{1}{2}$:

$$J\left(\frac{w_1 + w_2}{2}\right) - \frac{m}{2} - \frac{m}{2} < 0$$

ou encore

$$J\left(\frac{w_1 + w_2}{2}\right) < m = \min_{v \in H_0^1(]0,1[)} J(v)$$

D'où la contradiction.

Donc le problème d'optimisation a une solution unique, u , et comme toutes les sous-suites convergentes de la suite minimisante converge vers u , un argument de compacité classique montre que toute le suite $(w_\varepsilon)_\varepsilon$ converge vers u . \square

Exercice 4. On considère le problème de Dirichlet :

$$(P) \quad \begin{cases} -a(x)u'' + b(x)u' + c(x)u = f & \text{dans }]0, 1[\\ u(0) = u(1) = 0, \end{cases}$$

où $a, b, c, f \in C([0, 1])$ et le problème variationnel :

$$(P') \quad \begin{cases} \text{Trouver } \tilde{u} \in H_0^1(]0, 1[) \text{ tel que :} \\ J(\tilde{u}) = \min_{v \in H_0^1(]0, 1[)} J(v), \end{cases}$$

avec :

$$J(v) = \frac{1}{2} \int_0^1 [\alpha(t)[v'(t)]^2 + 2\beta(t)v'(t)v(t) + \gamma(t)[v(t)]^2] dt - \int_0^1 f(t)v(t)dt,$$

où α, β et γ sont des fonctions continues (ou plus régulières si nécessaire) avec $\alpha(t) \geq \delta > 0$ pour tout $t \in [0, 1]$.

1. À quelle condition liant α, β, γ et a, b, c , peut-on penser FORMELLEMENT que les deux problèmes sont équivalents, c'est-à-dire que les solutions de (P) et de (P') sont les mêmes.
2. Trouver une condition "naturelle" sur α, β, γ pour que J soit coercive.
3. Prouver que, sous cette même condition, J est aussi strictement convexe et C^1 .
4. En déduire que (P') admet une unique solution $\tilde{u} \in H_0^1(]0, 1[)$.
5. Trouver, de même, une condition pour que l'équation satisfasse le principe du maximum (on pourra songer à un changement de variable du type $v(x) = u(x) \exp(Kx)$ ou bien $v(x) = u(x) \sin(k_1x + k_2)$ pour des constantes K, k_1, k_2 bien choisie).

Exercice 5.

1. Prouver le Théorème 1.2.

2. Démontrer l'inégalité de Hardy : si $u \in H_0^1(]0, 1[)$,

$$\left\| \frac{u(x)}{x} \right\|_{L^2(]0,1[)} \leq c \|u'\|_{L^2(]0,1[)} ,$$

pour une certaine constante c à déterminer.

Exercice 6. (Exercice récapitulatif pour ceux qui se sentent bien musclés !)

Soit $h : \mathbb{R} \rightarrow \mathbb{R}$ une fonction de classe C^∞ qui satisfait :

$$0 < \alpha \leq h''(t) \leq \beta \quad \text{pour tout } t \in \mathbb{R} ,$$

pour certaines constantes α et β . On admettra qu'on a alors :

$$C_1 t^2 - D_1 \leq h(t) \leq C_2 t^2 + D_2 \quad \text{pour tout } t \in \mathbb{R} ,$$

pour certaines constantes strictement positives C_1, C_2, D_1, D_2 .

On considère le problème variationnel :

$$(P) \quad \begin{cases} \text{Trouver } u \in H_0^1(]0, 1[) \text{ tel que :} \\ J(u) = \min_{v \in H_0^1(]0,1[)} J(v), \end{cases}$$

avec :

$$J(v) = \int_0^1 h(v'(t)) dt + \frac{1}{2} \int_0^1 [v(t)]^2 dt - \int_0^1 f(t)v(t) dt ,$$

avec $f \in C([0, 1])$.

1. Prouver que le problème (P) admet au moins une solution $u \in H_0^1(]0, 1[)$ qui satisfait :

$$\forall v \in H_0^1(]0, 1[), \quad \int_0^1 h'(u'(t))v'(t) dt + \int_0^1 u(t)v(t) dt = \int_0^1 f(t)v(t) dt .$$

2. Pour $x \in [0, 1]$, on pose :

$$w(x) = h'(u'(x)) - \int_0^x u(t) dt + \int_0^x f(t) dt .$$

Rappeler pourquoi u est une fonction continue et montrer que :

$$(1.11) \quad \text{pour tout } v \in C_0^1([0, 1]), \quad \int_0^1 w(t)v'(t) dt = 0 .$$

3. En déduire que la fonction w est constante sur $[0, 1]$. (On pourra poser

$\bar{w} = \int_0^1 w(t) dt$ et prouver que $w - \bar{w}$ satisfait également (1.11).

(NB : Courtesy of Du Bois-Raymond !)

4. Démontrer, en utilisant la question 3, que $u \in C^1([0, 1])$ puis que $u \in C^2([0, 1])$. (On pourra remarquer que h' est une fonction strictement croissante sur \mathbb{R} .)

5. Prouver que u est solution du problème de Dirichlet :

$$(P') \quad \begin{cases} -h''(u')u'' + u = f & \text{dans }]0, 1[\\ u(0) = u(1) = 0. \end{cases}$$

6. Démontrer que le problème (P') admet une unique solution dans $C^2([0, 1])$ et en déduire que le problème (P) admet aussi une unique solution.

7. Prouver que :

$$\|u\|_{L^\infty]0,1[} \leq \|f\|_{L^\infty]0,1[}.$$

et que $u \in C^4([0, 1])$ si $f \in C^2([0, 1])$.

1.2 Résolution numérique du problème par la méthode des différences finies

1.2.1 Discrétisation de l'équation : le schéma numérique et ses propriétés

Un mot d'abord sur la terminologie "différences finies" : l'idée de cette méthode est de remplacer les dérivées usuelles par des "dérivées discrètes". On obtient d'abord une approximation des dérivées premières en les remplaçant par des quotients différentiels : typiquement, si $h > 0$ est petit,

$$u'(x) \simeq \frac{u(x+h) - u(x)}{h},$$

et le second membre est une différence "finie" car on n'a pas encore pris h infiniment petit. En itérant le processus, on peut définir des dérivées discrètes d'ordres supérieurs, ce qui permet d'avoir des approximations de toutes les dérivées.

Deux remarques : d'abord en pratique, on utilisera plutôt la formule de Taylor pour obtenir des approximations des dérivées d'ordres supérieurs car c'est plus simple. Ensuite, on voit dans l'exemple ci-dessus de l'approximation de la dérivée première que l'on a aussi :

$$u'(x) \simeq \frac{u(x) - u(x-h)}{h} \quad \text{ou} \quad u'(x) \simeq \frac{u(x+h) - u(x-h)}{2h},$$

et, en fait, on se rend vite compte qu'il y a souvent beaucoup de possibilités différentes pour approcher une dérivée. C'est une richesse qui amènera parfois des difficultés (voir le chapitre sur les équations de transport) car le choix de telle ou telle approximation ne sera pas toujours anodin et on pourra avoir des comportements très différents suivant les cas. Il faut aussi

prendre en compte la complexité des calculs et ne pas choisir de formules trop compliquées qui pourront générer des calculs longs.

Après cette introduction, nous revenons à la résolution du problème de Dirichlet (1)-(2). Comme l'ordinateur ne peut manipuler qu'un nombre fini de valeurs, on choisit N points de l'intervalle $[0, 1]$:

$$x_0 = 0 < x_1 < x_2 \cdots < x_{N-1} < x_N < x_{N+1} = 1 ,$$

et on va chercher à calculer une “bonne” approximation u_j de $u(x_j)$ pour tout $1 \leq i \leq N$ via un schéma numérique, c'est-à-dire un certain nombre d'équations permettant de déterminer les u_j . La notion de “bonne” approximation dépend des applications, et donc, en général, de l'ordre de grandeur des $u(x_j)$ et de la précision souhaitée.

L'ensemble des points x_j s'appelle une grille ou un maillage et chaque intervalle $]x_j, x_{j+1}[$ est souvent appelé une “maille”. L'exemple le plus simple est celui de la grille uniforme où l'on a :

$$x_j = \frac{j}{N+1} \quad \text{pour } 0 \leq j \leq N+1 .$$

Dans ce cas, la taille de la maille ($x_{j+1} - x_j$) est constante et on la notera suivant les cas Δx ou h . La quantité Δx est souvent appelée “pas de discrétisation”. Dans le cas général, la finesse du maillage est mesurée par la quantité :

$$D_N := \max_{0 \leq j \leq N} (x_{j+1} - x_j) .$$

Dans toute la suite, pour simplifier la présentation, nous utiliserons une grille uniforme.

Nous passons maintenant au schéma numérique : il s'agit de remplacer l'équation “continue” (1) et la condition aux limites (2) par un système d'équations sur les x_j . Pour cela, on considère le point x_j et on écrit la formule de Taylor pour u au point x_j , en utilisant les points voisins x_{j+1} et x_{j-1} :

$$u(x_{j+1}) = u(x_j + \Delta x) = u(x_j) + u'(x_j)\Delta x + \frac{1}{2}u''(x_j)(\Delta x)^2 + o((\Delta x)^2) ,$$

$$u(x_{j-1}) = u(x_j - \Delta x) = u(x_j) - u'(x_j)\Delta x + \frac{1}{2}u''(x_j)(\Delta x)^2 + o((\Delta x)^2) .$$

En sommant ces deux égalités, en retranchant $2u(x_j)$ et en divisant par $(\Delta x)^2$, on voit que :

$$(1.12) \quad \frac{u(x_{j+1}) + u(x_{j-1}) - 2u(x_j)}{(\Delta x)^2} = u''(x_j) + o(1) .$$

Cette formule donne une approximation de $u''(x_j)$ par une quantité ne dépendant que des valeurs de u ($u(x_{j+1})$, $u(x_j)$ et $u(x_{j-1})$) et il est naturel

de remplacer l'équation (1) par :

$$(1.13) \quad -\frac{u_{j+1} + u_{j-1} - 2u_j}{(\Delta x)^2} = f_j \quad \text{pour } 1 \leq j \leq N,$$

où $f_j = f(x_j)$, puisque l'on cherche des u_k qui sont des approximations des $u(x_k)$. Nous avons pu écrire cette équation pour tout j , et en particulier pour $j = 1$ et $j = N$, en utilisant la condition aux limites et la convention $u_0 = u(x_0) = u(0) = 0$ et $u_{N+1} = u(x_{N+1}) = u(1) = 0$.

Le système d'équations sur les u_j est donc un système linéaire de N équations à N inconnues. Avant de l'étudier, théoriquement et de manière plus pratique, un peu de vocabulaire : dans (1.12), la différence finies du membre de gauche est une approximation *consistante* de $u''(x_j)$ car elle ne diffère de cette dérivée seconde que d'un $o(1)$. Comme déjà mentionné plus haut (1.13) est appelé *schéma d'approximation numérique* de (1)-(2). Ce schéma est *consistant* car, en remplaçant dans (1.13) les u_k par des $u(x_k)$ ET en supposant u très régulier, on retrouve (1) à un $o(1)$ près.

Donnons une définition plus formelle de la consistance. On considère l'équation générale :

$$(E) \quad Lu = f \quad \text{dans }]a, b[,$$

et une grille $x_0 = a < x_1 < x_2 \cdots < x_{N-1} < x_N < x_{N+1} = b$; on note toujours u_k une approximation de $u(x_k)$. Un schéma numérique approchant (E) peut être écrit sous la forme :

$$(SN) \quad G_j^N(u_1, u_2, \dots, u_{N-1}, u_N) = 0 \quad \text{pour } 1 \leq j \leq N,$$

où, disons, les G_j^N sont des fonctions continues. La $j^{\text{ème}}$ équation $G_j^N = 0$ représente l'approximation de l'équation $Lu(x_j) = f(x_j)$, donc de l'équation (E) au point x_j , le N faisant référence au nombre de points de la grille.

Définition 1.2. *Le schéma numérique (SN) est dit consistant avec l'équation (E) si, pour tout $1 \leq j \leq N$, il existe une suite $(\rho_j^N)_N$ de réels telle que, pour toute fonction ϕ de classe C^∞ sur \mathbb{R} , on ait :*

$$\rho_j^N G_j^N(\phi(x_1), \phi(x_2), \dots, \phi(x_{N-1}), \phi(x_N)) = Lu(x_j) - f(x_j) + o(1)$$

quand N tend vers l'infini.

Cette définition est très imparfaite et elle est plus compliquée que la réalité sous-jacente : si, dans le schéma numérique, on remplace les u_k par des $\phi(x_k)$ où ϕ est une fonction très régulière, on doit retrouver l'équation à un $o(1)$ près, quitte à renormaliser le schéma par des ρ_j^N . Cette renormalisation est due au fait que si $G_j^N = 0$ est un schéma consistant, $NG_j^N = 0$ qui a les mêmes solutions doit l'être aussi, de même que $N^{-2}G_j^N = 0$. Les ρ_j^N prennent en compte ces modifications éventuelles du schéma.

La vérification de la consistance s'effectue toujours à l'aide de la formule de Taylor, ce qui ne pose aucun problème théorique puisque ϕ est de classe C^∞ sur \mathbb{R} et donc elle satisfait des formules de Taylor à tous les ordres. La consistance est donc une propriété quasi-formelle puisque tous les calculs sont justifiés a priori vu la régularité de ϕ . Il est à noter que la solution n'intervient pas du tout.

Il est naturel de penser que plus le schéma reproduit fidèlement l'équation, plus il sera précis et donc il est intéressant de mesurer cette proximité équation-schéma ; ceci donne lieu à la notion d'ordre qui précise le "o(1)" de la consistance.

Définition 1.3. *Le schéma numérique (SN) est d'ordre $k \geq 1$ si, pour tout $1 \leq j \leq N$, il existe une suite $(\rho_j^N)_N$ de réels telle que, pour toute fonction ϕ de classe C^∞ sur \mathbb{R} , on ait :*

$$\rho_j^N G_j^N(\phi(x_1), \phi(x_2), \dots, \phi(x_{N-1}), \phi(x_N)) = Lu(x_j) - f(x_j) + o((\Delta x)^k)$$

quand N tend vers l'infini.

Cette définition est naturelle dans le cas d'un maillage uniforme ; on peut l'étendre au cas d'un maillage quelconque en utilisant les D_N introduit plus haut.

La vérification de la consistance et de l'ordre s'effectue simultanément : la seule différence, c'est qu'il faut pousser un peu plus loin le développement de Taylor pour avoir l'ordre. Dans le cas de (1.13), on a :

$$\begin{aligned} G_j^N(\phi(x_1), \phi(x_2), \dots, \phi(x_{N-1}), \phi(x_N)) &= \frac{\phi(x_{j+1}) + \phi(x_{j-1}) - 2\phi(x_j)}{(\Delta x)^2} \\ &\quad - f(x_j) \\ &= -\phi''(x_j) - f(x_j) - \frac{1}{12}\phi^{(4)}(x_j)(\Delta x)^2 \\ &\quad + o((\Delta x)^4) . \end{aligned}$$

Le schéma est donc d'ordre 2.

Remarque : On peut penser que plus un schéma numérique est d'ordre élevé, plus il est précis ; c'est en général vrai avec trois réserves :

1. Un BON schéma d'ordre élevé est plus précis : il y a des cas où des schémas d'ordres élevés ne convergent même pas...
2. L'ordre de convergence va dépendre de la régularité de la solution : si la solution u n'est pas C^4 , on ne voit pas pourquoi le calcul ci-dessus impliquerait une précision d'ordre $(\Delta x)^2$. Nous verrons cette influence plus en détails dans la partie consacrée aux estimations de convergence.
3. Un schéma d'ordre élevé utilise beaucoup de points (exo : construire pour (1) un schéma d'ordre 3 et d'ordre 4 pour le vérifier) et ceci rend les calculs

plus complexes et plus coûteux en temps. Il faut penser au rapport qualité/prix...

Exercice 7.

1. Donner une approximation d'ordre 3 de u' et d'ordre 4 de u'' .
2. Proposer plusieurs schémas numériques consistants pour les équations des exercices 1, 4, 6.

1.2.2 Étude du système linéaire donné par le schéma

On pose $U = (u_j)_{1 \leq j \leq N}$ et $F = (f_j)_{1 \leq j \leq N}$. U et F sont deux vecteurs de \mathbb{R}^N . On introduit la matrice :

$$A = \begin{pmatrix} -2 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & & 0 & 1 & -2 \end{pmatrix}$$

Le schéma numérique se réécrit sous la forme :

$$-\frac{1}{(\Delta x)^2}AU = F.$$

Il est à noter que la matrice A est symétrique et (plus important encore numériquement) elle est *creuse* : seuls $3N - 2$ coefficients sont non nuls sur les N^2 possibles. On n'a donc à stocker en mémoire que $3N - 2$ nombres au lieu de N^2 . Bien sûr, ici, la simplicité de la matrice fait que l'on n'en stocke pas vraiment $3N - 2$...

Avant de rappeler comment résoudre en pratique ce système linéaire, on va en étudier certaines propriétés.

1. Les valeurs propres de A .

Comme A est symétrique, toutes ses valeurs propres sont réelles et A est diagonalisable dans une base orthonormée. On va calculer toutes les valeurs propres de A et les vecteurs propres associés.

Si $U = (u_j)_{1 \leq j \leq N}$ est un vecteur propre de A associé à la valeur propre λ , on a, en posant $u_0 = u_{N+1} = 0$:

$$u_{j+1} + u_{j-1} - 2u_j = \lambda u_j \quad \text{pour } 1 \leq j \leq N,$$

i.e.

$$u_{j+1} + u_{j-1} - (2 + \lambda)u_j = 0 \quad \text{pour } 1 \leq j \leq N.$$

Pour calculer les u_j , l'idée est de considérer cette égalité comme une suite récurrente d'ordre 2. On introduit l'équation caractéristique :

$$r^2 - (2 + \lambda)r + 1 = 0 .$$

Si l'équation caractéristique a une racine double, comme :

$$\Delta = (2 + \lambda)^2 - 4 ,$$

alors on a $\lambda = 0$ ou $\lambda = -4$. La racine double est 1 si $\lambda = 0$ et -1 si $\lambda = -4$. Montrons que ces deux cas sont impossibles.

Si $\lambda = 0$ alors la théorie des suite récurrentes nous dit qu'il existent $\alpha, \beta \in \mathbb{R}$ tels que, pour tout j :

$$u_j = \alpha(1)^j + \beta j(1)^j = \alpha + \beta j .$$

Comme $u_0 = u_{N+1} = 0$, on en déduit d'abord $\alpha = 0$ en utilisant $u_0 = 0$ puis $\beta = 0$ en utilisant $u_{N+1} = 0$. Donc $u_j = 0$ pour tout j et donc on n'a pas affaire à un vecteur propre. La preuve est analogue dans le cas $\lambda = -4$.

Si l'équation caractéristique a deux racines distinctes, r_1, r_2 , alors on a, pour tout j :

$$u_j = \alpha r_1^j + \beta r_2^j .$$

pour certaines constantes $\alpha, \beta \in \mathbb{R}$. Comme $u_0 = 0$, on a $\beta = -\alpha$ et, puisque nous cherchons U non nul (donc $\alpha \neq 0$), $u_{N+1} = 0$ conduit à $r_1^{N+1} = r_2^{N+1}$. Mais, par le lien entre coefficients et racines de l'équation du deuxième ordre, on a $r_1 r_2 = 1$ et en multipliant l'égalité précédente par r_1^{N+1} , on a forcément :

$$r_1^{2(N+1)} = 1 .$$

Il existe donc un entier k tel que $r_1 = \exp(2ik\pi/2(N+1))$, ce qui donne en utilisant encore le lien entre coefficients et racines de l'équation du deuxième ordre :

$$2 + \lambda = r_1 + r_2 = r_1 + r_1^{-1} = 2 \cos \left(\frac{k\pi}{N+1} \right) ,$$

et :

$$u_j = \alpha(r_1^j - r_2^j) = 2\alpha \sin \left(\frac{jk\pi}{N+1} \right) .$$

On vérifie facilement qu'en faisant varier k de 1 à N , on obtient ainsi N valeurs propres réelles λ_k et N vecteurs propres $U_k = (u_j^k)_j$ associés :

$$\lambda_k = 2 \left(\cos \left(\frac{k\pi}{N+1} \right) - 1 \right) ,$$

et :

$$u_j^k = \sin \left(\frac{jk\pi}{N+1} \right) .$$

On remarque que toutes les valeurs propres sont strictement négatives et la plus grande (donc la plus proche de 0) est :

$$\lambda_1 = 2 \left(\cos \left(\frac{\pi}{N+1} \right) - 1 \right) \sim -\frac{\pi^2}{(N+1)^2} .$$

Si on fait le produit scalaire de l'équation :

$$-\frac{1}{(\Delta x)^2} AU = F ,$$

par U , on obtient :

$$-\frac{\lambda_1}{(\Delta x)^2} \|U\|^2 \leq -\frac{1}{(\Delta x)^2} (AU, U) = (F, U) \leq \|F\| \cdot \|U\| ,$$

en utilisant le produit scalaire standard sur \mathbb{R}^N et la norme euclidienne associée. Il en résulte :

$$\|U\| \leq -\frac{(\Delta x)^2}{\lambda_1} \|F\| \sim \frac{1}{\pi^2} \|F\| .$$

Cette inégalité donne une estimation de $\|U\|$ et nous amène à nous poser la question de la *stabilité* du schéma : les calculs concrets sur ordinateurs vont fonctionner si les u_j calculés restent bornés et même relativement “petits” ; sinon l'ordinateurs arrêtera le calcul car il ne pourra plus gérer les nombres trop grands qui apparaîtront.

La propriété ci-dessus implique-t-elle la stabilité du schéma ? La réponse est : oui et non (mais plutôt non). Si on examine $\|F\|^2 = \sum_{i=1}^N f_j^2$, on voit que les f_j étant seulement bornés [Rappel : $f_j = f(x_j)$], a priori $\|F\|^2$ ne reste pas borné quand N tend vers l'infini et donc $\|U\|^2$ non plus. Par contre, la quantité :

$$\Delta x \|F\|^2 = \Delta x \sum_{i=1}^N [f(x_j)]^2 \sim \int_0^1 [f(t)]^2 dt ,$$

(pensez à la méthode des rectangles) reste borné et donc $\Delta x \|U\|^2$ aussi. Mais cette propriété n'implique pas que les u_j soient bornés. En fait, on a une stabilité L^2 , c'est-à-dire que la fonction constante par morceaux dont les valeurs sur chaque maille sont données par les u_j est bornée dans l'espace L^2 . Mais pas forcément dans L^∞ ...

Or nous souhaitons avoir cette stabilité L^∞ au sens de la définition suivante :

Définition 1.4. *Le schéma numérique (SN) est dit stable s'il admet une solution et si cette solution satisfait :*

$$\max_{1 \leq j \leq N} |u_j| \leq C \quad (\text{constante indépendante de } N).$$

2. Monotonie et stabilité.

On va d'abord montrer que le schéma numérique satisfait le *principe du maximum discret*. Pour cela, on introduit sur \mathbb{R}^N la relation d'ordre suivante : si $F = (f_j)_j$ et $G = (g_j)_j$, on dira que $F \geq G$ si $f_j \geq g_j$ pour tout j et, en particulier, on dira que $F \geq 0$ si $f_j \geq 0$ pour tout j .

Proposition 1.6. *Si U est la solution de (1.13) associée à F et si $F \geq 0$ alors $U \geq 0$.*

Corollaire 1.1. *Si U est la solution de (1.13) associée à F , si V est la solution de (1.13) associée à $G = (g_j)_j$ et si $F \leq G$ alors $U \leq V$.*

La preuve du corollaire est immédiate en appliquant la Proposition 1.6 à $V - U$ et $G - F$.

Preuve : On raisonne comme pour le principe du maximum pour l'équation (1). Avec la convention habituelle $u_0 = u_{N+1} = 0$, on considère :

$$\min_{0 \leq j \leq N+1} u_j,$$

qui est atteint en un certain u_{j_0} . Si $j_0 = 0$ ou $N + 1$, c'est terminé puisque le min serait nul donc tous les $u_j \geq 0$. Sinon on peut écrire le schéma en j_0

$$-\frac{u_{j_0+1} + u_{j_0-1} - 2u_{j_0}}{(\Delta x)^2} = f_{j_0} \geq 0,$$

et donc :

$$u_{j_0+1} + u_{j_0-1} - 2u_{j_0} \leq 0.$$

Mais, par la propriété de minimum, $u_{j_0} \leq u_{j_0+1}$ et $u_{j_0} \leq u_{j_0-1}$ et donc le premier membre est positif; il est même strictement positif si l'une des deux inégalités $u_{j_0} \leq u_{j_0+1}$ et $u_{j_0} \leq u_{j_0-1}$ est stricte. Il en résulte que nécessairement $u_{j_0} = u_{j_0+1}$ et $u_{j_0} = u_{j_0-1}$.

Mais alors, si le minimum des u_j est atteint pour un certain j_0 , il l'est aussi pour $j_0 + 1$ et $j_0 - 1$ et en considérant le plus petit (ou le plus grand) de ces j_0 on obtient une contradiction car forcément le min est atteint pour $j_0 = 0$ et $j_0 = N + 1$.

Déduisons maintenant la stabilité du schéma du principe du maximum discret.

Théorème 1.4. *Si U est la solution de (1.13) associée à F , on a :*

$$\max_{1 \leq j \leq N} |u_j| \leq \frac{1}{8} \|f\|_\infty.$$

Preuve : Montrons que, pour tout j , $u_j \leq \frac{1}{8} \|f\|_\infty$; le résultat complet s'obtient à partir de cette inégalité en changeant U en $-U$.

Notons W la solution associée à $G = (\|f\|_\infty)_j$ (toutes les coordonnées de G sont les mêmes et elles sont égales à $\|f\|_\infty$). Ce schéma est associé au problème de Dirichlet :

$$-w'' = \|f\|_\infty \quad \text{avec } w(0) = w(1) = 0,$$

dont la solution se calcule facilement :

$$w(x) = \frac{1}{2}\|f\|_\infty x(1-x).$$

Des calculs assez simples montrent qu'en fait $W = (w(x_j))_j$; on a donc une résolution exacte dans ce cas particulier ⁽⁶⁾.

Comme $F \leq G$, on a $U \leq W$ et une estimation facile de la norme L^∞ de w permet de conclure.

En examinant de plus près la preuve de la Proposition 1.6, on constate que le point clé est la monotonie du schéma que nous définissons maintenant.

Définition 1.5. *Le schéma numérique (SN) est dit monotone si, pour tout j , la fonction G_j^N est décroissante par rapport à u_k pour tout $k \neq j$.*

Exercice 8.

1. On considère l'équation (1) associée à une condition de Neumann ou une condition mixte (Dirichlet en 0 et Neumann en 1). Comment se modifie le schéma numérique et le système linéaire associé dans ces deux cas ? Est-il toujours inversible ? (Si la réponse est non, on pourra expliquer pourquoi et trouver un moyen de le rendre inversible.)

2. Reprendre les schémas numériques consistants obtenus dans l'exercice 7 pour les équations des exercices 1, 4, 6 ; étudier leur propriétés et, en particulier, leur stabilité L^∞ . (Si ce n'est déjà fait, on trouvera pour ces équations des schémas numériques monotones pour lesquels on montrera que le principe du maximum discret est satisfait.)

3. Résoudre numériquement le problème de l'exercice 3 (on n'oubliera pas d'utiliser la périodicité et d'indiquer ce qu'elle apporte).

1.3 Rappels sur quelques méthodes numériques de résolution de systèmes linéaires

L'erreur naïve que pourrait faire un lecteur peu averti serait de croire que le fait que l'on sache qu'une matrice $A \in \mathbb{M}_n(\mathbb{R})$ est inversible (par exemple en calculant son déterminant) résout le problème puisque l'on dispose des *Formules de Cramer*. S'il y réfléchit quelques instants, il conviendra que le

(6). Le lecteur pourra relier cette propriété avec l'étude de la consistance et de l'ordre du schémas en se souvenant que, pour un polynôme, la formule de Taylor est *exacte*.

coût de calcul de *chaque* coefficient de A^{-1} par la méthode des cofacteurs, est égal, outre la division par le déterminant de A qu'il faut aussi calculer, à celui C_{n-1} du calcul d'un déterminant $(n-1) \times (n-1)$. Il est clair que si on utilise le développement par rapport à une colonne on obtient ainsi $C_{n-1} = (n-1)C_{n-2} + n-2$ ce qui prouve que cette formule conduit à un coût d'inversion supérieur à $n!$ opérations!!!! Même avec les ordinateurs les plus puissants, on ne pourrait pas traiter de grandes matrices.

Donc la résolution d'un système de la forme $Ax = b$ ne se fera JAMAIS par le calcul de A^{-1} puis de $A^{-1}b$.

1.3.1 Les méthodes directes

Les méthodes directes nous donnent l'occasion d'illustrer le propos ci-dessus. On peut généralement les décrire de la manière suivante : si on veut résoudre dans \mathbb{R}^N le système $Ax = b$, on procède comme suit :

- **1ère Etape : élimination.** On détermine une matrice M inversible (que l'on ne calcule jamais dans la pratique) de telle sorte que la matrice MA soit facile à inverser (typiquement triangulaire supérieure).

- **2ème Etape : remontée.** On résout le système linéaire :

$$MAx = Mb$$

par une méthode dite "de remontée" que nous décrivons maintenant : comme MA est triangulaire supérieure, le système linéaire est de la forme :

$$(S) : \begin{cases} t_{11}x_1 + t_{12}x_2 + \dots + t_{1N}x_n & = c_1 \\ & \vdots \\ t_{ii}x_i + \dots + t_{iN}x_n & = c_i \\ & \vdots \\ t_{NN}x_n & = c_N \end{cases}$$

Et chaque coefficient $t_{i,i}$ est non nul puisque la matrice MA est inversible comme produit de matrice inversible.

Pour résoudre ce système, on commence par la dernière équation qui permet de calculer d'abord x_N , puis on calcule x_{N-1} par l'avant-dernière et on "remonte" ainsi, en calculant successivement tous les x_i .

Rappels sur la méthode de Gauss :

Ce que l'on décrit ici, c'est évidemment la première étape dite d'élimination, la deuxième étant complètement automatique.

On pose $A = (a_{i,j})_{i,j}$. Comme A est inversible, la première colonne de A est non nulle ; il existe donc un indice i_0 pour lequel $a_{i_0,1} \neq 0$. Pour

des raisons d'erreur d'arrondis sur lesquelles nous reviendrons plus tard, on choisit en fait un indice i_0 tel que :

$$|a_{i_0,1}| = \max_{1 \leq i \leq N} |a_{i,1}|$$

puis on permute la première ligne et la ligne i_0 de A ce qui revient à multiplier la matrice A par la matrice de permutation :

$$P_{ij} = \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & \dots & \vdots \\ 0 & 0 & \mathbf{0} & 0 & \mathbf{1} & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \mathbf{1} & 0 & \mathbf{0} & 0 & \vdots \\ 0 & \dots & \dots & \dots & 0 & 1 & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix} \begin{array}{l} \leftarrow \text{ligne } i \\ \\ \\ \\ \leftarrow \text{ligne } j \\ \\ \end{array}$$

$$\begin{array}{cc} \uparrow & \uparrow \\ \text{col. } i & \text{col. } j \end{array}$$

qui échange la ligne i et la ligne j avec, dans notre cas, $i = 1$ et $j = i_0$. On a évidemment $\det P_{ij} = -1$.

On se retrouve avec une matrice $P_{1,i_0}A = (\alpha_{i,j})_{i,j}$ dont le coefficient $\alpha_{1,1}$ est non nul et donc (c'est une autre façon de le voir) avec le système linéaire :

$$(S) : \begin{cases} \alpha_{1,1}x_1 + \alpha_{1,2}x_2 + \dots + \alpha_{1,N}x_n = c_1 \\ \dots \dots \\ \alpha_{j,1}x_1 + \alpha_{j,2}x_2 + \dots + \alpha_{j,N}x_n = c_j \\ \dots \dots \\ \alpha_{N,1}x_1 + \alpha_{N,2}x_2 + \dots + \alpha_{N,N}x_n = c_N \end{cases}$$

On élimine alors u_1 des $(N-1)$ dernières équations en changeant la ligne j , L_j du système en $L_j - \frac{\alpha_{j,1}}{\alpha_{1,1}}L_1$, ce qui revient à multiplier la matrice $P_{1,i_0}A$ à gauche par la matrice :

$$E_1 = \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ -\frac{\alpha_{2,1}}{\alpha_{1,1}} & 1 & 0 & 0 & \dots & \dots & \vdots \\ -\frac{\alpha_{3,1}}{\alpha_{1,1}} & 0 & 1 & 0 & 0 & \dots & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 & \ddots & 0 & \vdots \\ -\frac{\alpha_{N-1,1}}{\alpha_{1,1}} & 0 & \dots & \dots & 0 & 1 & 0 \\ -\frac{\alpha_{N,1}}{\alpha_{1,1}} & 0 & \dots & \dots & \dots & 0 & 1 \end{pmatrix}$$

On note $P_1 = P_{1,i_0}$. Le bilan matriciel de cette première étape s'écrit :

$$E_1 P_1 A u = E_1 P_1 b$$

où, maintenant, la matrice $A_1 = E_1 P_1 A$ a pour forme :

$$\begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,N} \\ 0 & \boxed{\tilde{A}_1} \\ \vdots & & & \\ 0 & & & \end{pmatrix}.$$

On applique alors le même argument à la matrice \tilde{A}_1 qui est inversible puisque $\alpha_{1,1} \det(\tilde{A}_1) = \det E_1 \det P_1 \det A = -\det A$ et on se retrouve avec :

$$A_2 = E_2 P_2 E_1 P_1 A = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \cdots & \alpha_{1,N} \\ 0 & \beta_{2,2} & \cdots & \cdots & \beta_{2,N} \\ 0 & 0 & \boxed{\tilde{A}_2} \\ \vdots & \vdots & & & \\ 0 & 0 & & & \end{pmatrix}$$

où P_2 est une matrice de permutation de la forme $P_{j,2}$ avec $j \geq 2$ et E_2 est de la forme :

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & * & 1 & 0 & 0 \\ 0 & \vdots & 0 & \ddots & 0 \\ 0 & * & 0 & 0 & 1 \end{pmatrix}$$

Et on continue : pour la $k^{\text{ième}}$ étape, on a une matrice du type :

$$\begin{pmatrix} a_{1,1}^{(k)} & a_{1,2}^{(k)} & \cdots & \cdots & \cdots & a_{1,N}^{(k)} \\ 0 & a_{2,2}^{(k)} & \cdots & \cdots & \cdots & a_{2,N}^{(k)} \\ \vdots & 0 & \ddots & \cdots & \cdots & \cdots \\ \vdots & \vdots & 0 & a_{k,k}^{(k)} & \cdots & a_{k,N}^{(k)} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & a_{N,k}^{(k)} & \cdots & a_{N,N}^{(k)} \end{pmatrix}$$

On permute éventuellement les lignes pour que $a_{k,k}^{(k)} \neq 0$, ce qui revient à multiplier à gauche par une matrice de permutation P_k puis on multiplie à gauche par une matrice E_k de la forme :

$$\begin{pmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & \vdots & * & \ddots & \cdots & 0 \\ 0 & 0 & \vdots & 0 & \ddots & 0 \\ 0 & 0 & * & 0 & 0 & 1 \end{pmatrix}$$

ce qui revient à faire des combinaisons de lignes pour annuler les $a_{j,k}^{(k)}$ pour $k + 1 \leq j \leq N$.

La matrice :

$$A_N := E_{N-1}P_{N-1} \cdots E_1P_1A$$

que l'on obtient finalement est triangulaire supérieure et on a réalisé l'étape d'élimination avec :

$$M := E_{N-1}P_{N-1} \cdots E_1P_1.$$

Remarque : 1. On ne peut être que frappé par la simplicité de toutes les étapes de cette procédure : combinaisons linéaires et permutations de lignes, il n'y a que des étapes élémentaires. Bien sûr, le calcul de Mb se fait de la même manière, en appliquant les mêmes opérations à b .

2. À cause des erreurs d'arrondies, il vaut mieux choisir le "meilleur pivot", celui dont la taille est la plus grande comme le montre l'exemple caricatural suivant :

$$\begin{cases} 10^{-4}x_1 + x_2 = 1 \\ x_1 + x_2 = 2 \end{cases}$$

La solution est $x_1 = (1 - 10^{-4})^{-1} \simeq 1$ et $x_2 = (1 - 210^{-4})(1 - 10^{-4})^{-1} \simeq 1$. Le pivot de Gauss conduit au système :

$$\begin{cases} 10^{-4}x_1 + x_2 = 1 \\ -9999x_1 + x_2 = -9998 \end{cases}$$

Si les nombres 9998 et 9999 sont arrondis de la même manière tous les deux alors $x_2 = 1$ et la première équation conduit à $x_1 = 0$, ce qui ne redonne pas vraiment la solution du système ! Éviter des pivots trop petits paraît donc très utile...

3. La méthode de Gauss nécessite de l'ordre de :

$$N^3/3 \text{ additions, } N^3/3 \text{ multiplications, } N^2/2 \text{ divisions}$$

alors que les formules de Cramer requièrent :

$$(N + 1)! \text{ additions, } (N + 2)! \text{ multiplications, } N \text{ divisions}$$

Faire $N = 10$ et comparer ! (700 vs 400 000 000)

Malgré ses avantages, la méthode de Gauss n'est vraiment utilisée en Analyse Numérique que pour résoudre des problèmes où la matrice A n'a aucune propriété particulière. Sinon des algorithmes plus performants lui seront préférés (cf. la factorisation de Cholesky plus bas).

Factorisation LU d'une matrice

Si l'on doit résoudre non pas un seul système linéaire mais un grand nombre de systèmes linéaires avec la même matrice (voir les sections consacrées aux problèmes d'évolution), il peut être intéressant de conserver en mémoire

les opérations faites sur la matrice au cours du pivot de Gauss et donc ne faire l'étape d'élimination qu'une seule fois.

En fait, si on n'a pas de problème de pivot nul (c'est-à-dire si $a_{k,k}^{(k)} \neq 0$ pour tout k) et si l'on peut donc éviter de permuter les lignes, on peut choisir $P_1 = \dots = P_{N-1} = Id$. La matrice $M = E_{N-1} \dots E_1$ est triangulaire inférieure et MA est triangulaire supérieure : en posant $U = MA$ et $L = M^{-1}$, on voit que : $A = LU$, L est triangulaire inférieure et U est triangulaire supérieure. On a bien factorisé A comme produit d'une matrice triangulaire inférieure et d'une matrice triangulaire supérieure.

Dans ces conditions, la résolution de $Ax = b$ se fait simplement par une "descente-remontée" : en effet, si on pose $w = Ux$, le système $Ax = b$ se décompose en :

- $Lw = b$ (qui se résout par une descente)
- $Ux = w$ (qui se résout par une remontée)

Une fois connus L et U , on résout donc $Ax = b$ de manière extrêmement efficace, d'où l'intérêt de la factorisation LU , surtout si on a 100 000 systèmes linéaires à résoudre...

La méthode de Gauss nous fournit déjà U et pour L , on remarque que, d'une part :

$$L = E_1^{-1} \dots E_{N-1}^{-1},$$

et d'autre part que le calcul des E_k^{-1} est simple puisque si :

$$E_k = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & 0 & \dots & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & \vdots & -l_{k+1,k} & \ddots & \dots & 0 \\ 0 & 0 & \vdots & 0 & \ddots & 0 \\ 0 & 0 & -l_{N,k} & 0 & 0 & 1 \end{pmatrix}$$

alors :

$$E_k^{-1} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & 0 & \dots & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & \vdots & l_{k+1,k} & \ddots & \dots & 0 \\ 0 & 0 & \vdots & 0 & \ddots & 0 \\ 0 & 0 & l_{N,k} & 0 & 0 & 1 \end{pmatrix}$$

Le calcul de L est donc immédiat à partir de la méthode de Gauss.

Il reste à se demander quand on peut vraiment prendre $P_1 = \dots = P_{N-1} = Id$. Le résultat est le suivant :

Théorème 1.5. *Si A est une matrice carrée $N \times N$ telle que les N sous-matrices :*

$$\Delta_k = \begin{pmatrix} a_{1,1} & \cdots & a_{1,k} \\ \vdots & & \vdots \\ a_{1,1} & \cdots & a_{k,k} \end{pmatrix}$$

soient inversibles, alors il existe une unique factorisation $A = LU$ où U est matrice $N \times N$ triangulaire supérieure et L est une matrice $N \times N$ triangulaire inférieure avec $L_{i,i} = 1$ pour $1 \leq i \leq N$.

Remarque : La condition imposant des 1 sur la diagonale de L sert à démontrer l'unicité car, dans ce type de décomposition, U et L sont définies à une multiplication par une matrice diagonale près. Il faut donc introduire une "normalisation" pour avoir l'unicité.

Preuve : On procède par récurrence : comme $a_{1,1} = \det(\Delta_1) \neq 0$, le premier pivot est bien non nul et P_1 peut être pris égal à l'identité. Supposons maintenant que l'on a pu prendre $P_1 = \cdots = P_{k-1} = Id$ et vérifions que le $k^{\text{ième}}$ pivot est non nul. L'égalité $A_k = E_{k-1} \cdots E_1 A$ s'écrit :

$$\left(\begin{array}{ccc|cc} a_{1,1}^{(k)} & \cdots & a_{1,k}^{(k)} & * & * \\ 0 & \ddots & \vdots & * & * \\ 0 & 0 & a_{k,k}^{(k)} & * & * \\ \hline 0 & 0 & * & * & * \\ 0 & 0 & * & * & * \end{array} \right) = \left(\begin{array}{ccc|cc} 1 & 0 & \cdots & 0 & 0 \\ * & \ddots & \vdots & \vdots & \vdots \\ * & * & 1 & 0 & 0 \\ \hline * & * & 0 & 1 & 0 \\ * & * & 0 & 0 & 1 \end{array} \right) \left(\begin{array}{ccc|cc} & & & * & * \\ & & \Delta_k & * & * \\ \hline * & * & * & * & * \\ * & * & * & * & * \end{array} \right)$$

En utilisant les règles de multiplications des matrices par blocs et en calculant le déterminant, on a :

$$a_{1,1}^{(k)} \cdots a_{k,k}^{(k)} = 1 \cdot \det(\Delta_k)$$

d'où $a_{k,k}^{(k)} \neq 0$ puisque $\det(\Delta_k) \neq 0$, Δ_k étant inversible. On peut donc prendre $P_k = Id$ et l'existence est démontrée.

Pour l'unicité, si $A = L_1 U_1 = L_2 U_2$ avec L_1, U_1 et L_2, U_2 satisfaisant les conditions du théorème, on en déduit :

$$L_2^{-1} L_1 = U_2 U_1^{-1}.$$

Or le membre de gauche est triangulaire inférieur alors que celui de droite est triangulaire supérieur ; il en résulte que les deux membres sont des matrices diagonales et il s'agit en fait de l'identité car $L_2^{-1} L_1$ a des 1 sur sa diagonale puisque c'est le cas de L_1 et L_2 .

Donc $L_1 = L_2$ et $U_1 = U_2$, l'unicité est démontrée.

Factorisation de Cholesky

Théorème 1.6. *Si A est une matrice symétrique définie positive, il existe une unique matrice réelle triangulaire inférieure telle que :*

$$A = BB^T ,$$

et telle que $B_{i,i} > 0$ pour tout $1 \leq i \leq N$.

La factorisation de Cholesky est donc une factorisation de type “ LU ” particulière puisque, dans ce cas, $U = L^T$. Un premier avantage clair est l’économie de stockage des coefficients des matrices de la factorisation car on la divise ici par 2. On en verra un autre avantage plus loin lié au calcul de B qui s’effectuera de manière plus économique que pour une factorisation LU classique.

Preuve : On remarque tout d’abord que les N sous-matrices Δ_k introduite dans le résultat de la factorisation LU sont toutes symétriques, définies positives (donc inversibles). En fait, Δ_k est la matrice de la forme quadratique $q(x) = (Ax, x)$ mais restreinte à l’espace vectoriel $\text{Vect}(e_1, \dots, e_k)$ où $(e_i)_i$ désigne la base canonique de \mathbb{R}^N .

Une autre façon de le voir est de se rappeler qu’il existe $\alpha > 0$ tel que :

$$(Ax, x) \geq \alpha \|x\|^2 \quad \text{pour tout } x = (x_1, \dots, x_N) \in \mathbb{R}^N ,$$

et qu’en faisant $x_{k+1} = \dots = x_N = 0$, on voit que Δ_k satisfait la même propriété.

Donc A admet une factorisation LU . De plus, si $U = (u_{i,j})_{i,j}$ alors les $u_{i,i}$ sont strictement positifs car la preuve de la factorisation LU montre que, pour tout k :

$$\det(\Delta_k) = \prod_{i=1}^k u_{i,i} .$$

De plus, comme A est symétrique, on a :

$$A = A^T = (LU)^T = U^T L^T .$$

U^T est une matrice triangulaire inférieure alors que L^T est triangulaire supérieure mais on ne peut pas appliquer directement le résultat d’unicité car U^T n’a pas forcément des 1 sur la diagonale. Pour se ramener à ce cas, on introduit la matrice $\Lambda = \text{diag}(\sqrt{u_{i,i}})$ dont on “intercale” le carré dans cette égalité :

$$A = (U^T \Lambda^{-2})(\Lambda^2 L^T) .$$

Maintenant $U^T \Lambda^{-2}$ est triangulaire inférieure avec des 1 sur la diagonale, alors que $\Lambda^2 L^T$ est triangulaire supérieure et, par unicité, on déduit $U^T \Lambda^{-2} = L$, $\Lambda^2 L^T = U$.

Enfin, on écrit :

$$A = (L\Lambda)(\Lambda^{-1}U)$$

et on pose $B = L\Lambda$, $C = \Lambda^{-1}U$. Le calcul de B^T donne :

$$B^T = (L\Lambda)^T = \Lambda L^T = \Lambda^{-1}U = C$$

ce qui donne le résultat voulu.

Calcul pratique de B :

On pose :

$$B = \begin{pmatrix} b_{1,1} & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & \\ \vdots & b_{i,j} & \ddots & 0 \\ b_{N,1} & \cdots & \cdots & b_{N,N} \end{pmatrix}$$

et on déduit de $A = BB^T$ les relations suivantes pour la première ligne de $A = (a_{i,j})_{i,j}$:

$$\begin{aligned} a_{1,1} &= b_{1,1}^2 & \text{d'où } b_{1,1} &= \sqrt{a_{1,1}} \\ a_{1,2} &= b_{1,1}b_{2,1} & \text{d'où } b_{2,1} &= a_{1,2}/\sqrt{a_{1,1}} \\ & \vdots & & \\ a_{1,N} &= b_{1,1}b_{N,1} & \text{d'où } b_{N,1} &= a_{1,N}/\sqrt{a_{1,1}} \end{aligned}$$

On détermine donc la première colonne de B en utilisant la première ligne de A .

De manière générale :

$$a_{i,j} = \sum_{k=1}^i b_{i,k}b_{j,k} ,$$

et on voit immédiatement qu'en utilisant la deuxième ligne de A ($i = 2$), on peut calculer successivement tous les coefficients de la deuxième colonne de B ; puis, de proche en proche, on va pouvoir obtenir les coefficients de la $k^{\text{ième}}$ colonne de B en examinant la $k^{\text{ième}}$ ligne de A .

Cette méthode ne requiert que :

$$\begin{aligned} &N^3/6 \text{ additions, } N^3/6 \text{ multiplications, } N^2/2 \text{ divisions,} \\ &N \text{ extractions de racines carrées} \end{aligned}$$

elle est donc presque deux fois plus efficace que la méthode de Gauss dans ce cas.

Bien sûr, la résolution de $Ax = b$ s'effectue ici aussi par une "descente-remontée", en résolvant successivement les systèmes linéaires :

- $Bw = b$ (descente)
- $B^T x = w$ (remontée)

Et nous ne quitterons pas cette section sans faire remarquer au lecteur que la matrice de notre schéma numérique est symétrique définie positive et donc la factorisation de Cholesky est une bonne méthode pour calculer les u_j .

1.3.2 Les méthodes itératives

Le principe des méthodes itératives est de considérer une suite $(x_k)_k$ d'éléments de \mathbb{R}^N vérifiant une relation de récurrence du type :

$$(1.14) \quad x_{k+1} = Bx_k + c, \quad x_0 \text{ arbitraire}$$

où B est une matrice $N \times N$ et $c \in \mathbb{R}^N$ qui sont construits à partir des données A et b du système linéaire $Ax = b$. Bien évidemment, on souhaite que la suite $(x_k)_k$ converge vers l'unique solution du système linéaire (et si possible assez rapidement...) et, pour cela, on va appliquer le théorème du point fixe de Picard dans sa version matricielle (théorème du point fixe pour les applications contractantes).

De nombreuses méthodes (Jacobi, Gauss-Seidel, gradient..) repose sur une décomposition de la matrice sous la forme $A = M - N$ où M est non seulement une matrice inversible, mais surtout une matrice facile à inverser *numériquement*. Dans la pratique, M sera *diagonale ou triangulaire*.

Ainsi x est solution de $Ax = b$ si et seulement si $Mx = Nx + b$ i.e. $x = M^{-1}Nx + M^{-1}b$ soit $x = Bx + c$. Ce qui donne une manière "naturelle" de construire B et c .

Le résultat de base pour la convergence de la suite $(x_k)_k$ est le suivant :

Théorème 1.7. *Pour tout choix de vecteur x_0 , la méthode numérique définie par le schéma (1.14) :*

$$x_{n+1} = Bx_n + c, \quad x_0 \text{ donné}$$

converge vers l'unique solution u du problème

$$(1.15) \quad x = Bx + c$$

si et seulement si l'une des deux propriétés équivalentes ci-dessous est satisfaite :

- a) *Le rayon spectral de B vérifie : $\rho(B) < 1$*
- b) *Il existe une norme matricielle assujettie $|||\cdot|||$ pour laquelle $|||B||| < 1$*

Preuve : La preuve contient deux parties distinctes : il faut d'abord comprendre pourquoi les points a) et b) sont équivalents. Ensuite, si on utilise la

propriété b), il s'agit d'une application du théorème des applications contractantes (via la méthode des itérations de Picard). En effet, le point b) dit que l'application linéaire associée est contractante si on munit \mathbb{R}^N de la bonne norme, celle qui donne $|||\cdot|||$ comme norme assujettie.

“Rappel” : on dit que la norme $|||\cdot|||$ est assujettie s'il existe une norme $\|\cdot\|$ sur \mathbb{R}^N telle que :

$$|||B||| = \sup_{x \in \mathbb{R}^N} \frac{\|Bx\|}{\|x\|}$$

autrement dit lorsque B est vu comme la matrice d'une application linéaire de \mathbb{R}^N dans \mathbb{R}^N . Dans ce cas-là, on a alors :

$$\forall y \in \mathbb{R}^N \quad \|By\| \leq |||B||| \|y\|, \quad \text{ou} \quad \forall y, z \in \mathbb{R}^N \quad \|By - Bz\| \leq |||B||| \|y - z\|.$$

On voit donc que le point b) assure précisément que B est contractante donc possède un unique point fixe x , ou si l'on préfère que la matrice $I - B$ est inversible⁽⁷⁾.

D'autre part, la condition $\rho(B) < 1$ est certainement nécessaire pour qu'il y ait convergence quelque soit la valeur x_0 choisie, le vecteur c étant donné. En effet sinon il existerait une valeur propre λ de module supérieur ou égal à 1 associée au vecteur propre e et le choix de $x_0 = c + e$ conduirait à une suite $(x_k)_k$ non convergente. \square

Avant de voir pour quoi ces deux points sont équivalents, remarquons qu'en itérant $k - 1$ fois l'inégalité ci-dessus, on a, bien sûr :

$$\|x_k - x\| = \|B^k(x_0 - x)\| \leq |||B|||^k \|x_0 - x\|$$

ce qui dit que la convergence est exponentiellement rapide et qu'elle est d'autant plus rapide que la norme $|||B|||$ est petite. D'où l'idée de choisir au mieux la matrice M de sorte que $\rho(M^{-1}N)$ soit le plus petit possible (puisque a) et b) sont équivalents).

Précisons l'équivalence des points a) et b) en démontrant le lemme suivant :

Lemme 1.3. 1) Soit $|||\cdot|||$ une norme assujettie. Alors on a toujours : $\rho(B) \leq |||B|||$.
2) Pour tout $\varepsilon > 0$, il existe une norme assujettie $|||\cdot|||$ de sorte que : $|||B||| \leq \rho(B) + \varepsilon$

Preuve : Soit λ une valeur propre de B de module maximal et e un vecteur propre associé. Alors $\|Be\| = |\lambda| \|e\| = \rho(B) \|e\|$ ce qui établit le point 1) compte tenu du fait que la norme est assujettie.

Le deuxième point est un peu plus technique : il repose sur deux ingrédients,

(7). suggestion de révision pour le lecteur : théorèmes de points fixe dans \mathbb{R}^n et convergence des séries dans les espaces de Banach ; application : (re-)démontrer l'assertion du texte.

le calcul explicite de la norme infinie d'une matrice et la triangularisation des matrices.

On sait en effet qu'il existe une matrice P inversible et une matrice triangulaire $T = (t_{i,j})_{i,j}$ ayant pour spectre le même que celui de B , telles que $PBP^{-1} = T$.

Soit t un petit nombre réel. Introduisons la matrice diagonale $\Delta_t = (t^{i-1}\delta_{i,j})_{i,j}$ $\delta_{i,j}$ étant le symbole de Kronecker. Il est facile de voir que pour t non nul Δ_t est inversible et $(\Delta_t)^{-1} = \Delta_{t^{-1}}$. De plus multiplier à gauche par Δ_t une matrice M , c'est multiplier la ligne i de la matrice M par t^{i-1} tandis que multiplier à droite c'est multiplier la colonne j de M par t^{j-1} . Ainsi la matrice $T_\varepsilon = \Delta_\varepsilon T (\Delta_\varepsilon)^{-1}$ est triangulaire supérieure comme T mais :

$$T_\varepsilon = \begin{pmatrix} t_{1,1} & \varepsilon t_{1,2} & \varepsilon^2 t_{1,3} & \dots & \dots & \varepsilon^{n-2} t_{1,n-1} & \varepsilon^{n-1} t_{1,n} \\ 0 & t_{2,2} & \varepsilon t_{2,3} & \varepsilon^2 t_{2,4} & \dots & \dots & \varepsilon^{n-2} t_{2,n} \\ 0 & 0 & t_{3,3} & \varepsilon t_{3,4} & \varepsilon^2 t_{3,5} & \dots & \varepsilon^{n-3} t_{3,n} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \dots & 0 & 0 & t_{n-2,n-2} & \varepsilon t_{n-2,n-1} & \varepsilon^2 t_{n-2,n} \\ 0 & \dots & \dots & \dots & 0 & t_{n-1,n-1} & \varepsilon t_{n-1,n} \\ 0 & \dots & \dots & \dots & \dots & 0 & t_{n,n} \end{pmatrix}$$

où $\{t_{1,1}, \dots, t_{n,n}\} = \{\lambda_1, \dots, \lambda_n\}$, les λ_i étant les valeurs propres de B . Maintenant, on rappelle que :

$$\|A\|_\infty = \max_i \sum_j |a_{i,j}|$$

où $\|\cdot\|_\infty$ est la norme matricielle assujettie à la norme infinie sur \mathbb{R}^n , et on voit alors que la norme $\|T_\varepsilon\|_\infty = \max_i \sum_j \varepsilon^{j-i} |t_{i,j}|$ converge vers $\max |t_{i,i}| = \rho(B)$. \square

Quelques exemples de méthodes itératives

On considère une matrice inversible A telle que tous les $a_{i,i}$ soient non nuls.

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & \dots & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} & \dots & \dots & a_{2,n} \\ a_{3,1} & a_{3,2} & a_{3,3} & a_{3,4} & a_{3,5} & \dots & a_{3,n} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \dots & \ddots & a_{n-2,n-3} & a_{n-2,n-2} & a_{n-2,n-1} & \vdots \\ a_{n-1,1} & \dots & \dots & \dots & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ a_{n,1} & \dots & \dots & \dots & \dots & a_{n,n-1} & a_{n,n} \end{pmatrix}$$

On écrit $A = D - E - F$ où D est la diagonale de A , $-E$ la matrice triangulaire sous la diagonale de A et $-F$ la matrice triangulaire située au

dessus de la diagonale i.e. :

$$D_{ij} = a_{ij}\delta_{ij} ; \quad -E_{ij} = a_{ij}\delta_{i>j} ; \quad -F_{ij} = a_{ij}\delta_{i<j}$$

a) **Méthode de Jacobi**

Dans cette méthode $M = D$ et $N = E + F$ et la suite itérative définie est :

$$x_{n+1} = Jx_n + c \quad J = D^{-1}(E + F) = I - D^{-1}A, c = D^{-1}b.$$

On la définit en fait par

$$Dx_{n+1} = (E + F)x_n + b$$

en utilisant une procédure (écrivez-là dans votre langage préféré) qui inverse la matrice diagonale D .

b) **Méthode de Gauss-Seidel**

En écrivant explicitement la méthode de Jacobi ci-dessus , on voit que l'on pourrait mieux utiliser certaines quantités déjà calculées : par exemple si on calcule d'abord x_1^{k+1} que l'on peut ensuite utiliser à la place de x_1^k On est conduit ainsi à la méthode de Gauss-Seidel

$$(1.16) \quad (D - E)y_{n+1} = Fy_n + b$$

qui correspond à $M = D - E$ et $N = F$ et la suite itérative est définie aussi par :

$$y_{n+1} = Gy_n + c \quad G = (D - E)^{-1}F, \quad c = (D - E)^{-1}b.$$

Mais bien sûr , on utilisera la forme (1.16) et une procédure qui inverse la matrice triangulaire $M = D - E$ (écrivez en une). Intuitivement, la méthode de Gauss-Seidel est plus implicite que celle de Jacobi donc doit être plus performante. On verra en exercice qu'en général la méthode de Gauss-Seidel, quand elle converge, converge mieux que la méthode de Jacobi car le rayon spectral de G est plus petit que celui de J .

c) **Méthode de relaxation**

On peut essayer d'améliorer la méthode de Gauss-Seidel en introduisant un paramètre réel $\omega \neq 0$ et en considérant la méthode itérative :

$$\left(\frac{D}{\omega} - E\right) x_{k+1} = \left(\frac{1-\omega}{\omega}D + F\right) x_k + b.$$

En effet, si la méthode de Gauss-Seidel (qui correspond à $\omega = 1$) converge, on peut peut-être améliorer la convergence en choisissant un "meilleur" paramètre de relaxation $\omega \neq 1$.

Si on pose :

$$\mathcal{L}_\omega := \left(\frac{D}{\omega} - E \right)^{-1} \left(\frac{1-\omega}{\omega} D + F \right) ,$$

l'idée serait de déterminer un ω_0 optimal, c'est-à-dire tel que :

$$\rho(\mathcal{L}_{\omega_0}) = \inf_{\omega \in I} \rho(\mathcal{L}_\omega) ,$$

où I est un intervalle de \mathbb{R} ne contenant pas 0.

En pratique, il n'est pas toujours évident de calculer un tel ω_0 mais on peut étudier les propriétés de convergence de la méthode de relaxation, ce qui donnera une idée de la localisation d'un tel ω_0 .

Un pas essentiel dans cette direction est la :

Proposition 1.7. *Soit A une matrice symétrique définie positive décomposée sous la forme :*

$$A = M - N ,$$

où M est une matrice inversible. Si la matrice symétrique $M^T + N$ est définie positive alors :

$$\rho(M^{-1}N) < 1 .$$

Preuve : La matrice $M^T + N$ est effectivement symétrique puisque :

$$M^T + N = (A + N)^T + N = A + N^T + N ,$$

et A et $N^T + N$ sont symétriques.

On va prouver que $\|M^{-1}N\| < 1$, en utilisant une norme matricielle $\|\cdot\|$ subordonnée à la norme $|\cdot|$ sur \mathbb{R}^n , définie par :

$$|v|^2 = (Av, v) \quad \text{pour tous } v \in \mathbb{R}^n .$$

On remarque d'abord que $M^{-1}N = M^{-1}(M - A) = I - M^{-1}A$. Si $w = M^{-1}Av$, on calcule $|M^{-1}Nv|^2 = |v - w|^2$ pour $|v| = 1$:

$$\begin{aligned} |v - w|^2 &= |v|^2 - 2(Av, w) + (Aw, w) && \text{(identité remarquable) ,} \\ &= 1 - 2(Mw, w) + (Aw, w) && \text{(définition de } w \text{) ,} \\ &= 1 - (Mw, w) + ((A - M)w, w) && \text{(réarrangement des termes) ,} \\ &= 1 - (M^T w, w) - (Nw, w) && \text{((} Mw, w \text{) = (} M^T w, w \text{)) ,} \\ &= 1 - ((M^T + N)w, w) && \text{(réarrangement des termes)} \end{aligned}$$

Comme $M^T + N$ est une matrice symétrique définie positive et que A, M sont inversibles, $((M^T + N)w, w) > 0$ pour tout v tels que $|v| = 1$ et donc :

$$|M^{-1}Nv|^2 < 1 \quad \text{pour tous } v \in \mathbb{R}^n \text{ tels que } |v| = 1 .$$

Il en résulte que $\|M^{-1}N\| < 1$ puisque $\{v : |v| = 1\}$ est compact. \square

Application pratique : Nous revenons d'abord à la matrice du schéma numérique (renormalisée en laissant tomber le $\frac{1}{(\Delta x)^2}$) :

$$A = \begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & & 0 & -1 & 2 \end{pmatrix}.$$

La méthode de Gauss-Seidel conduit à la décomposition de A sous la forme $M = D - E$ et $N = F$, i.e. :

$$M = \begin{pmatrix} 2 & 0 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & & 0 & -1 & 2 \end{pmatrix} \text{ et } N = \begin{pmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & & 0 & 0 & 0 \end{pmatrix}.$$

On remarque alors que M est inversible et que :

$$M^T + N = \begin{pmatrix} 2 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 2 & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & & \ddots & \ddots & 0 \\ 0 & \cdots & & 0 & 0 & 2 \end{pmatrix}.$$

est bien définie positive. Donc $\rho(M^{-1}N) < 1$ et la méthode converge.

De même, dans le cas de la méthode de relaxation :

$$M = \begin{pmatrix} \frac{2}{\varepsilon} & 0 & 0 & \cdots & \cdots & 0 \\ -1 & \frac{2}{\varepsilon} & 0 & 0 & \cdots & 0 \\ 0 & -1 & \frac{2}{\varepsilon} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & & 0 & -1 & \frac{2}{\varepsilon} \end{pmatrix}$$

et :

$$N = \begin{pmatrix} 2\frac{(1-\omega)}{\omega} & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 2\frac{(1-\omega)}{\omega} & 1 & 0 & \cdots & 0 \\ 0 & 0 & 2\frac{(1-\omega)}{\omega} & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & \ddots & 1 \\ 0 & \cdots & & 0 & 0 & 2\frac{(1-\omega)}{\omega} \end{pmatrix}.$$

De même, M est inversible. De plus :

$$M^T + N = \begin{pmatrix} 2\frac{(2-\omega)}{\omega} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 2\frac{(2-\omega)}{\omega} & 0 & 0 & \cdots & 0 \\ 0 & 0 & 2\frac{(2-\omega)}{\omega} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & & 0 & 0 & 2\frac{(2-\omega)}{\omega} \end{pmatrix}.$$

est définie positive si $2\frac{(2-\omega)}{\omega} > 0$, donc si $0 < \omega < 2$. Finalement on a $\rho(M^{-1}N) < 1$ et convergence si et seulement si $0 < \omega < 2$.

On va voir maintenant que ce cas particulier reflète le cas général, au moins pour des matrices symétriques définies positives.

Théorème 1.8. (Condition suffisante de convergence pour la méthode de relaxation).

Si la matrice A est symétrique définie positive alors la méthode de relaxation converge pour tous $0 < \omega < 2$.

Preuve : La décomposition de $A = M - N$ associée à la méthode de relaxation s'écrit :

$$A = \left(\frac{D}{\omega} - E \right) - \left(\frac{1-\omega}{\omega} D + F \right).$$

De telle sorte que :

$$\begin{aligned} M^T + N &= \left(\frac{D}{\omega} - E \right)^T + \left(\frac{1-\omega}{\omega} D + F \right) \\ &= \frac{D}{\omega} - E^T + \frac{1-\omega}{\omega} D + F \\ &= \frac{2-\omega}{\omega} D, \end{aligned}$$

car A étant symétrique, $E^T = F$.

Comme A est définie positive, ses coefficients diagonaux sont positifs car ils sont de la forme (Ae_i, e_i) où $(e_i)_i$ est la base canonique de \mathbb{R}^n . Donc $M^T + N$ est définie positive si $\frac{2-\omega}{\omega} > 0$, i.e. si $0 < \omega < 2$. \square

Cette propriété ($0 < \omega < 2$) est aussi nécessaire comme le montre le :

Théorème 1.9. (Condition nécessaire de convergence pour la méthode de relaxation).

Pour toute matrice A , on a :

$$\rho(\mathcal{L}_\omega) \geq |\omega - 1| ,$$

et donc la méthode de relaxation ne peut converger que si $0 < \omega < 2$.

Preuve : Si les $\lambda_i(\mathcal{L}_\omega)$ sont les valeurs propres de \mathcal{L}_ω , on a :

$$\det(\mathcal{L}_\omega) = \prod_{i=1}^n \lambda_i(\mathcal{L}_\omega) = \frac{\det\left(\frac{1-\omega}{\omega}D + F\right)}{\det\left(\frac{D}{\omega} - E\right)} = (1-\omega)^n ,$$

en tenant compte de la structure particulière des matrices D, E, F . Mais $\rho(\mathcal{L}_\omega) = \max_i |\lambda_i(\mathcal{L}_\omega)|$ et donc :

$$|\rho(\mathcal{L}_\omega)|^n \geq \prod_{i=1}^n |\lambda_i(\mathcal{L}_\omega)| = |1-\omega|^n .$$

Et le résultat en découle facilement. \square

Remarque : On peut aussi montrer que les méthodes de Jacobi et Gauss-Seidel convergent ou sont divergentes simultanément et que, dans le cas où elles convergent, la méthode de Gauss-Seidel converge plus rapidement.

1.4 Une autre méthode pour calculer la solution du problème de Dirichlet : l'approche variationnelle

1.4.1 Discrétisation du problème variationnel

Nous avons vu que la solution du problème de Dirichlet (1)-(2) est aussi solution du problème d'optimisation :

$$\min_{v \in H_0^1(]0,1[)} J(v),$$

où :

$$J(v) = \frac{1}{2} \int_0^1 [v'(t)]^2 dt - \int_0^1 f(t)v(t) dt.$$

Cette propriété nous donne une deuxième approche pour calculer u numériquement, en discrétisant ce problème d'optimisation. On utilise les notations de la section 1.2 et on combine la méthode des différences finies avec les techniques standard de calcul approché d'intégrales. Plus précisément, on va écrire :

$$\begin{aligned} \int_0^1 [v'(t)]^2 dt &= \sum_{i=0}^N \int_{x_i}^{x_{i+1}} [v'(t)]^2 dt \\ &\simeq \sum_{i=0}^N (x_{i+1} - x_i) \left| \frac{v(x_{i+1}) - v(x_i)}{x_{i+1} - x_i} \right|^2, \end{aligned}$$

en utilisant la méthode des rectangles et de même :

$$\begin{aligned} \int_0^1 f(t)v(t) dt &= \sum_{i=0}^N \int_{x_i}^{x_{i+1}} f(t)v(t) dt \\ &\simeq \sum_{i=0}^N (x_{i+1} - x_i) f(x_i)v(x_i). \end{aligned}$$

On voit qu'ici aussi on a de nombreuses variantes : notre double choix n'est d'ailleurs pas très cohérent car la méthode des rectangles pour la première intégrale revient à supposer que v' est constant sur chaque sous-intervalle $]x_i, x_{i+1}[$, donc a priori que v est affine sur chacun de ces sous-intervalles ce qui devrait nous conduire à une méthode de type plutôt trapèze pour la seconde intégrale... Mais, bien qu'incohérent, ce choix a l'avantage d'être simple.

Nous n'allons utiliser ici encore que grilles uniformes. En divisant tous les termes par Δx , ce qui ne change pas le problème d'optimisation, nous sommes donc conduits à minimiser la fonction :

$$J^N(V) = \frac{1}{2} \sum_{i=0}^N \frac{(v_{i+1} - v_i)^2}{(\Delta x)^2} - \sum_{i=0}^N f_i v_i,$$

où $V = (v_i)_{1 \leq i \leq N}$ avec la convention $v_0 = v_{N+1} = 0$ et les f_i ne sont rien d'autres que les $f(x_i)$.

En fait, ce problème d'optimisation est très lié au système linéaire via la

Proposition 1.8. *Pour tout $V \in \mathbb{R}^N$, on a :*

$$\sum_{i=0}^N \frac{(v_{i+1} - v_i)^2}{(\Delta x)^2} = -(AV, V)$$

où :

$$A = \begin{pmatrix} -2 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & & & & \ddots & 0 \\ 0 & \cdots & & 0 & 1 & -2 \end{pmatrix}.$$

Cette proposition permet d'affirmer que, pour tout $V \in \mathbb{R}^N$:

$$J^N(V) = -\frac{1}{2(\Delta x)^2}(AV, V) - (F, V),$$

et la section 1.4.2 ci-dessous montre que minimiser J^N est équivalent à résoudre $AU = F$ puisque $-A$ est une matrice symétrique, définie positive.

Preuve : La preuve est simple et elle ne nécessite que des manipulations élémentaires. Mais il est bon d'avoir un peu de recul sur le résultat lui-même qui est l'analogie discret de l'égalité :

$$\int_0^1 [v'(t)]^2 dt = - \int_0^1 v(t)v''(t) dt.$$

Cette dernière égalité s'obtenant (formellement) par intégration par parties, on va faire de même en discret en appliquant une *transformation d'Abel*.

On écrit :

$$\sum_{i=0}^N (v_{i+1} - v_i)^2 = \sum_{i=0}^N (v_{i+1} - v_i)v_{i+1} - \sum_{i=0}^N (v_{i+1} - v_i)v_i,$$

puis on fait un changement d'indice $i + 1 \rightarrow i$ dans la première somme du second membre qui devient :

$$\sum_{i=1}^{N+1} (v_i - v_{i-1})v_i - \sum_{i=0}^N (v_{i+1} - v_i)v_i.$$

Enfin, on réécrit ce résultat sous la forme :

$$(v_{N+1} - v_N)v_{N+1} - (v_1 - v_0)v_0 + \sum_{i=1}^N (v_{i+1} + 2v_i - v_{i-1})v_i .$$

On reconnaît la forme typique que l'on obtient après une intégration par parties : "terme tout intégré" + "nouvelle intégrale" ; Comme $v_{N+1} = v_0 = 0$, le terme tout intégré est nul et le résultat est prouvé. \square

1.4.2 Systèmes linéaires et problèmes d'optimisation : un résultat fondamental

Cette section étant d'un intérêt général et pouvant être lue indépendamment de ce qui précède, nous allons utiliser des notations différentes : A désignera une matrice $n \times n$ symétrique dont toutes les valeurs propres sont strictement positives, $b \in \mathbb{R}^n$ et x ou y seront les variables dans \mathbb{R}^n .

On a le résultat suivant :

Théorème 1.10. $x \in \mathbb{R}^n$ est solution du système linéaire $Ax = b$ si et seulement si on a :

$$f(x) = \min_{y \in \mathbb{R}^n} f(y) ,$$

où la fonction f est donnée par :

$$f(y) = \frac{1}{2}(Ay, y) - (b, y) .$$

Preuve : Cette preuve est tout à fait standard mais elle va nous permettre de revoir certains résultats fondamentaux d'optimisation dans \mathbb{R}^n .

On va montrer que le problème d'optimisation a une unique solution x et que cette solution satisfait l'égalité $Ax = b$. Le système linéaire étant inversible, on aura bien le résultat escompté.

Existence : Un résultat classique nous dit que le problème d'optimisation :

$$\text{Trouver } x \in \mathbb{R}^n \text{ tel que } f(x) = \min_{y \in \mathbb{R}^n} f(y) ,$$

a au moins une solution si f est continue⁽⁸⁾ et *coercive*, i.e.

$$f(y) \rightarrow +\infty \quad \text{quand } |y| \rightarrow +\infty .$$

On veut appliquer ce résultat à $f(y) = \frac{1}{2}(Ay, y) - (b, y)$; nous devons donc en vérifier les hypothèses. La continuité est triviale car f est une fonction polynomiale en les x_i . La coercivité résulte du :

Lemme 1.4. Si $\lambda_1 > 0$ est la plus petite valeur propre de A , on a, pour tout $y \in \mathbb{R}^n$:

$$(Ay, y) \geq \lambda_1 \|y\|^2 .$$

(8). semi-continue inférieurement serait suffisant.

Preuve : La matrice A étant symétrique, elle est diagonalisable dans une base orthonormée $(e_i)_i$ et si on décompose $y \in \mathbb{R}^n$ dans cette base sous la forme $y = y_1 e_1 + y_2 e_2 + \dots + y_n e_n$, on a :

$$\begin{aligned} (Ay, y) &= (y_1 \lambda_1 e_1 + y_2 \lambda_2 e_2 + \dots + y_n \lambda_n e_n, y_1 e_1 + y_2 e_2 + \dots + y_n e_n) \\ &= \lambda_1 y_1^2 + \lambda_2 y_2^2 + \dots + \lambda_n y_n^2 \\ &\geq \lambda_1 y_1^2 + \lambda_1 y_2^2 + \dots + \lambda_1 y_n^2 \\ &\geq \lambda_1 (y_1^2 + y_2^2 + \dots + y_n^2) \\ &\geq \lambda_1 \|y\|^2. \end{aligned}$$

Et le lemme est prouvé.

On en déduit, en utilisant l'inégalité de Cauchy-Schwarz pour le terme (b, y) :

$$f(y) \geq \frac{1}{2} \lambda_1 \|y\|^2 - \|b\| \cdot \|y\|,$$

et la coercivité en résulte. On a même, en utilisant astucieusement l'inégalité " $|cd| \leq \frac{1}{2}(c^2 + d^2)$ " ⁽⁹⁾ :

$$f(y) \geq \frac{1}{8} \lambda_1 \|y\|^2 - \frac{2}{\lambda_1} \|b\|^2,$$

ce qui donne une estimation de la taille de la (ou des) solutions en utilisant que $f(x) \leq f(0) = 0$.

Unicité : On pourrait conclure à l'unicité plus rapidement mais raisonnons de manière générale : l'unicité est (très) souvent reliée à la stricte convexité de la fonction à minimiser. Cette stricte convexité (à vérifier) s'écrit de la manière suivante sur une fonction $g : \mathbb{R}^n \rightarrow \mathbb{R}$ générale : pour tous $x_1 \neq x_2 \in \mathbb{R}^n$, pour tout $\alpha \in]0, 1[$, on a :

$$g(\alpha x_1 + (1 - \alpha)x_2) < \alpha g(x_1) + (1 - \alpha)g(x_2).$$

Pour la fonction f qui nous intéresse, on a le :

Lemme 1.5. *Pour tous $x_1 \neq x_2 \in \mathbb{R}^n$, pour tout $\alpha \in [0, 1]$, on a :*

$$f(\alpha x_1 + (1 - \alpha)x_2) = \alpha f(x_1) + (1 - \alpha)f(x_2) - \alpha(1 - \alpha)(A(x_1 - x_2), x_1 - x_2).$$

Nous laissons la preuve (fastidieuse mais sans mystère!) de ce résultat au lecteur : il suffit de calculer (et quand on connaît le résultat...). La stricte convexité en découle immédiatement puisque $(A(x_1 - x_2), x_1 - x_2) \geq \lambda_1 \|x_1 - x_2\|^2 > 0$.

Rappelons enfin l'argument classique prouvant l'unicité à partir de la stricte convexité. Si $m = \min_{\mathbb{R}^n} f$ et si on a deux solutions x_1, x_2 du

(9). cette inégalité fait partie des outils les plus utilisés lorsque l'on fait des estimations en analyse ; le lecteur est invité à en donner au moins deux démonstrations élémentaires.

problème d'optimisation, on a donc $f(x_1) = f(x_2) = m$ et la stricte convexité implique avec le choix $\alpha = \frac{1}{2}$:

$$\begin{aligned} f\left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right) &< \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2) \\ &< \frac{1}{2}m + \frac{1}{2}m = m, \end{aligned}$$

ce qui contredit la définition de m .

Propriété du point de minimum : la fonction f étant polynomiale, elle est dérivable et on peut écrire qu'en un point de minimum (local ou global) : $f'(x) = 0$.

Nous profitons de cette occasion pour rappeler les trois formes que prend la dérivation pour une fonction $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Par définition, f est dérivable au point x s'il existe une application linéaire, notée $f'(x) : \mathbb{R}^n \rightarrow \mathbb{R}$, telle que :

$$f(x+h) = f(x) + f'(x)(h) + o(h),$$

où $o(h)$ désigne une fonction qui a la propriété : $\frac{o(h)}{\|h\|} \rightarrow 0$ quand $h \rightarrow 0$.

Même si cette remarque peut sembler évidente, nous notons que la dérivation est ainsi définie sous la forme d'un développement limité et que (peut-être ?) faire un développement limité pour calculer une dérivée n'est pas totalement absurde⁽¹⁰⁾.

Pour passer à une formulation (plus rassurante) avec des dérivées partielles, il suffit de choisir une base de \mathbb{R}^n que nous noterons $(e_i)_i$ et qui n'est peut-être pas la base canonique (sacrilège !) : en écrivant $h = h_1e_1 + h_2e_2 + \dots + h_n e_n$, le terme correspondant à la dérivée s'écrit :

$$f'(x)(h) = h_1 f'(x)(e_1) + h_2 f'(x)(e_2) + \dots + h_n f'(x)(e_n),$$

et on voit facilement que, pour tout i , $f'(x)(e_i) = \frac{\partial f}{\partial x_i}(x)$ et donc :

$$f'(x)(h) = h_1 \frac{\partial f}{\partial x_1}(x) + h_2 \frac{\partial f}{\partial x_2}(x) + \dots + h_n \frac{\partial f}{\partial x_n}(x).$$

Hélas, si vous êtes réfractaire au Calcul Différentiel, vous n'êtes pas tout à fait au bout de vos peines car cette égalité peut être vue de deux façons ; on peut d'abord l'interpréter sous une version matricielle car $f'(x)$ est une application linéaire de \mathbb{R}^n dans \mathbb{R} et on peut exhiber la matrice dans la base $(e_i)_i$, que l'on notera $Df(x)$. C'est une matrice 1-ligne – n colonnes qui s'écrit :

$$Df(x) = \left(\frac{\partial f}{\partial x_1}(x) \quad \dots \quad \frac{\partial f}{\partial x_i}(x) \quad \dots \quad \frac{\partial f}{\partial x_n}(x) \right).$$

(10). même si malheureusement, ce réflexe "naturel" s'observe relativement rarement chez les étudiants ...

Si on introduit le vecteur colonne des coordonnées de h :

$$H = \begin{pmatrix} h_1 \\ \vdots \\ h_i \\ \vdots \\ h_n \end{pmatrix},$$

on a :

$$f'(x)(h) = Df(x).H,$$

le membre de gauche étant la version intrinsèque (indépendante du choix d'une base), le membre de droite étant la version non intrinsèque (on a choisi une base) et le “.” désigne un produit de matrices.

La deuxième interprétation que l'on peut donner de cette égalité est celle d'un produit scalaire, ce qui nous conduit à la notion de “gradient”. On note :

$$\nabla f(x) = \frac{\partial f}{\partial x_1}(x)e_1 + \cdots + \frac{\partial f}{\partial x_i}(x)e_i + \cdots + \frac{\partial f}{\partial x_n}(x),$$

et on a, pour tout h :

$$f'(x)(h) = (\nabla f(x), h).$$

Cette écriture renvoie au théorème de représentation des formes linéaires (trivial en dimension finie). Il est à noter que le gradient est un vecteur de \mathbb{R}^n : il “vit” donc dans le même espace que x , ce qui sera fondamental dans les méthodes de gradient. Il dépend aussi de la base choisie et nous utiliserons numériquement cette propriété pour choisir (si possible) le “meilleur” gradient.

L'intérêt de notre exemple va être d'illustrer ces notions, en tout cas deux d'entre elles. Par un simple calcul, on a :

$$f(x+h) = f(x) + \frac{1}{2}(Ax, h) + \frac{1}{2}(Ah, x) - (b, h) + \frac{1}{2}(Ah, h).$$

On a d'abord, en utilisant Cauchy-Schwarz :

$$0 \leq \frac{(Ah, h)}{\|h\|} \leq \|Ah\|$$

et comme $\|Ah\| \rightarrow 0$ quand h tend vers 0 par continuité de l'application linéaire $h \mapsto Ah$, on voit que le dernier terme est le $o(h)$ apparaissant dans la définition de la dérivée.

L'application $h \mapsto \frac{1}{2}(Ax, h) + \frac{1}{2}(Ah, x) - (b, h)$ est linéaire et c'est le $f'(x)(h)$ cherché car on sait que f est dérivable puisqu'elle est polynomiale. Nous insistons sur la forme de ce terme qui apparaît naturellement sous la forme d'une application linéaire.

Pour passer au gradient, il faut écrire cette expression sous la forme du produit scalaire d'un certain vecteur avec h et c'est une opération non triviale qui nécessite l'introduction de la matrice transposée de A , notée A^* . On a ainsi :

$$f'(x)(h) = \frac{1}{2}(Ax, h) + \frac{1}{2}(Ah, x) - (b, h) = \left(\frac{1}{2}Ax + \frac{1}{2}A^*x - b, h\right),$$

et, sans supposer A symétrique, on aboutit à $\nabla f(x) = \frac{1}{2}Ax + \frac{1}{2}A^*x - b$ qui devient $\nabla f(x) = Ax - b$ quand A est symétrique.

Si x est le point de minimum de f sur \mathbb{R}^n , on a donc $\nabla f(x) = 0$ et donc $Ax - b = 0$. L'unique solution du problème d'optimisation est donc solution du système linéaire.

La réciproque est aisée car, en rassemblant les informations ci-dessus, on s'aperçoit que, pour tout h :

$$f(x+h) = f(x) + (Ax - b, h) + \frac{1}{2}(Ah, h).$$

Si $Ax = b$, comme $(Ah, h) > 0$ si $h \neq 0$, on constate que $f(x+h) > f(x)$ pour tout $h \neq 0$ et donc x est l'unique point de minimum global strict de f sur \mathbb{R}^n .

Cette preuve "ad hoc" cache, en fait, une propriété plus générale : notre argument resterait valable si f était une fonction strictement convexe (et dérivable). Nous renvoyons le lecteur aux exercices sur ce thème.

Nous insistons sur le caractère fondamental du résultat de cette section pour l'Analyse Numérique : quand A est symétrique avec toutes ses valeurs propres positives, on peut choisir de résoudre directement le système linéaire OU de changer complètement d'approche et d'utiliser une méthode de minimisation comme les méthodes de gradient décrites dans la section suivante. Cette souplesse est intéressante car elle permet, suivant les cas, de se tourner vers la méthode la plus performante (ou, de manière moins avouable, vers sa méthode préférée...).

Exercice 9.

1. Reprendre le parallèle de l'exercice 4 au niveau numérique : discrétiser le problème variationnel et étudier les liens avec le schéma numérique pour l'équation associée.

2. Même question pour l'exercice 4.

1.5 Méthodes de gradient pour la résolution de problème d'optimisation

Les méthodes de gradient sont des algorithmes itératifs permettant de résoudre numériquement des problèmes d'optimisation dans \mathbb{R}^N ; ils utilisent

de manière fondamentale le gradient de la fonction à minimiser, d'où leur nom.

Pour être plus précis, on veut résoudre le problème général :

$$\text{Trouver } x_\infty \in \mathbb{R}^N \text{ tel que } J(x_\infty) = \min_{x \in \mathbb{R}^N} J(x)$$

où J est une fonction de classe C^1 sur \mathbb{R}^N .

1.5.1 Premier essai : méthode du gradient à pas constants

Pour démarrer le processus de minimisation, on se donne un point de départ x_0 et on cherche simplement x_1 tel que $J(x_1) < J(x_0)$. On suppose que $\nabla J(x_0) \neq 0$ (sinon on est peut-être au point de minimum ? en tout, cas ce serait vrai dans le cadre de la section précédente.).

Si on se limite à chercher x_1 proche de x_0 , on peut poser $x_1 = x_0 + h$ avec l'idée que, pour h est petit, on peut utiliser la différentiabilité de J . On a donc :

$$J(x_0 + h) = J(x_0) + (\nabla J(x_0), h) + o(h).$$

Si le "o" est vraiment négligeable, trouver le plus petit $J(x_0 + h)$ possible revient à trouver un produit scalaire minimal et, comme Cauchy-Schwarz nous dit que :

$$(\nabla J(x_0), h) \geq -\|\nabla J(x_0)\| \cdot \|h\|,$$

avec égalité si et seulement si h est colinéaire à $\nabla J(x_0)$ et de direction opposée, on est conduit à envisager le choix : $h = -\rho \nabla J(x_0)$ avec $\rho > 0$ petit (car il faut s'assurer que le "o" ne joue pas de rôle).

Ceci est naturel du point de vue de la physique car il s'agit de la "direction de plus grande pente", direction qu'une bille posée sur le graphe de J emprunterait pour dévaler la pente.

D'où notre premier choix, **la méthode du gradient à pas constant** : on choisit *une fois pour toute* un $\rho > 0$ petit et à partir du point initial x_0 , on calcule les x_n par récurrence via la formule :

$$x_{n+1} = x_n - \rho \nabla J(x_n).$$

Cet algorithme est très simple (et donc tentant) mais il est hautement instable comme le montre dans \mathbb{R} , l'exemple suivant : si on considère $J(x) = x^4$, on a $J'(x) = 4x^3$ et donc $x_{n+1} = x_n - 4\rho x_n^3 = x_n(1 - 4\rho x_n^2)$. Si, par malheur, on a choisi ρ ou x_0 trop grand de telle sorte que $1 - 4\rho x_0^2 < -1$, on voit facilement que, pour tout n , $1 - 4\rho x_n^2 < -1$ et $|x_0| < |x_1| < \dots < |x_n|$. En particulier, $J(x_0) < J(x_1) < \dots < J(x_n)$, ce qui n'est pas vraiment le but recherché... et l'algorithme diverge.

1.5.2 Deuxième essai : méthode de plus grande pente

Pour pallier au défaut de la méthode du gradient à pas constants, on peut essayer d'optimiser le choix du ρ à chaque étape (autant que faire se peut), ce qui conduit à la méthode de plus grande pente⁽¹¹⁾ : on calcule les x_n par récurrence via la formule :

$$x_{n+1} = x_n - \rho_n \nabla J(x_n) ,$$

où (théoriquement) ρ_n est défini par :

$$J(x_n - \rho_n \nabla J(x_n)) = \min_{\rho \geq 0} J(x_n - \rho \nabla J(x_n)) .$$

On notera désormais $\tilde{J}(\rho) = J(x_n - \rho \nabla J(x_n))$ (on n'a pas indexé \tilde{J} par n pour avoir des notations plus légères ; en général, il n'y a pas d'ambiguïté car on travaillera sur \tilde{J} à n fixé).

Nous avons écrit "théoriquement" ci-dessus car il est clair que, numériquement, on ne tentera jamais de déterminer le vrai ρ_n de manière exacte. La petite procédure suivante (même si elle semble naïve au premier abord) se révèle très efficace en pratique.

On procède par dichotomie de la manière suivante : on se donne $\bar{\rho}_0 > 0$ et on pose :

$$\begin{aligned} \bar{\rho}_1 &= \bar{\rho}_0/2 \quad , \quad \bar{\rho}_2 = \bar{\rho}_0 \quad , \quad \bar{\rho}_3 = 2\bar{\rho}_0 \quad , \\ j_1 &= \tilde{J}(\bar{\rho}_1) \quad , \quad j_2 = \tilde{J}(\bar{\rho}_2) \quad , \quad j_3 = \tilde{J}(\bar{\rho}_3) . \end{aligned}$$

Plusieurs cas :

- si $j_1 \leq j_2$, on imagine que le minimum doit être atteint pour des ρ plus petits et on déplace les points vers la gauche comme suit :

$$\bar{\rho}_1^n = \bar{\rho}_1^a/2 \quad , \quad \bar{\rho}_2^n = \bar{\rho}_2^a/2 = \bar{\rho}_1^a \quad , \quad \bar{\rho}_3^n = \bar{\rho}_3^a/2 = \bar{\rho}_2^a \quad ,$$

où on a noté avec un n les nouvelles valeurs alors que le a fait référence aux anciennes valeurs. Et on reprend la procédure avec ces nouvelles valeurs.

On remarque qu'à cause de la forme de $\bar{\rho}_1, \bar{\rho}_2, \bar{\rho}_3$ et des changements ci-dessus, la nouvelle itération ne demandera que le calcul d'un seul j_i (ici j_1) donc elle sera peut coûteuse. Cette remarque restera valable dans tous les cas.

- si $j_1 > j_2 \geq j_3$, on imagine cette fois que le minimum doit être atteint pour des ρ plus grands et on déplace les points vers la droite comme suit :

$$\bar{\rho}_1^n = 2\bar{\rho}_1^a = \bar{\rho}_2^a \quad , \quad \bar{\rho}_2^n = 2\bar{\rho}_2^a = \bar{\rho}_3^a \quad , \quad \bar{\rho}_3^n = 2\bar{\rho}_3^a .$$

- si $j_1 > j_2$ et $j_2 < j_3$, la procédure s'arrête et, ou bien on considère que $\bar{\rho}_2$ est le point de minimum cherché, ou bien on utilise une interpolation par

(11). "steepest descent method" en anglais

un polynôme de degré 2 pour affiner (du moins, on l'espère) la valeur de ce point.

Munie de cette procédure de calcul du “ ρ optimal”, la méthode de plus grande pente est assez robuste et elle marche assez bien à condition d'initialiser convenablement le $\bar{\rho}_0$ (le ρ optimal de l'étape précédente est rarement un mauvais choix...).

Le théorème suivant montre que cette méthode converge si J est de classe \mathcal{C}^2 et si elle est convexe coercive.

Théorème 1.11. *On suppose que J est de classe \mathcal{C}^2 et que, pour tout $x \in \mathbb{R}^N$:*

$$D^2J(x) \geq \alpha Id \quad (\alpha > 0) .$$

Alors la méthode de plus grande pente converge vers l'unique point dce minimum (global) de J sur \mathbb{R}^N .

Preuve : On rappelle que l'hypothèse sur J a deux conséquences : pour tous $x, y \in \mathbb{R}^N$, on a :

- (i) $J(y) \geq J(x) + (\nabla J(x), y - x) + \frac{\alpha}{2} \|y - x\|^2$,
- (ii) $(\nabla J(y) - \nabla J(x), y - x) \geq \alpha \|y - x\|^2$.

La preuve de ces deux propriétés est laissée en exercice (mais on peut suggérer au lecteur de reviser la formule de Taylor avec reste intégral...).

De la première propriété, on déduit que J est coercive (en prenant par exemple $x = 0$) et donc J atteint son minimum en un point x_∞ qui vérifie $\nabla J(x_\infty) = 0$. Il est difficile de loucher l'unicité qui est une conséquence de (i) ou (ii) au choix ; par exemple (i) nous dit :

$$J(y) \geq J(x_\infty) + \frac{\alpha}{2} \|y - x_\infty\|^2 > J(x_\infty) \quad \text{si } y \neq x_\infty ,$$

et donc x_∞ est clairement le seul point de minimum. De (i) ou (ii), on tire aussi que x_∞ est le seul point où le gradient de J s'annule (propriété remarquable des fonctions convexes coercives).

Pour prouver la convergence de la méthode de plus grande pente, on examine, pour chaque n , le problème d'optimisation de \tilde{J} . Il est à noter que \tilde{J} est coercive, donc le minimum est atteint (ρ_n existe), et que \tilde{J} est de classe \mathcal{C}^1 avec pour tout ρ :

$$\tilde{J}'(\rho) = - \left(\nabla J(x_n - \rho \nabla J(x_n)), \nabla J(x_n) \right) .$$

Cette formule montre que $\tilde{J}'(0) = -\|\nabla J(x_n)\|^2$ et donc 0 ne peut être point de minimum que si $\nabla J(x_n) = 0$ et donc on aurait $x_n = x_\infty$ [et donc $x_k = x_\infty$ si $k \geq n$] et la convergence.

Sinon $\rho_n > 0$ et on peut réécrire l'égalité $\tilde{J}'(\rho_n) = 0$ sous la forme :

$$(\nabla J(x_{n+1}), \nabla J(x_n)) = 0 .$$

Ce qui s'interprète comme le fait que deux directions de descente consécutives sont orthogonales.

On s'intéresse maintenant à la valeur des $J(x_n)$: par le choix de $\rho_1, \rho_2, \dots, \rho_n$, on a :

$$J(x_n) \leq J(x_{n-1}) \leq \dots \leq J(x_1) \leq J(x_0) .$$

En particulier, tous les points x_n appartiennent au convexe compact :

$$K := \{x \in \mathbb{R}^N; J(x) \leq J(x_0)\} .$$

On utilise alors l'inégalité (i) :

$$J(x_n) \geq J(x_{n+1}) + (\nabla J(x_{n+1}), x_{n+1} - x_n) + \frac{\alpha}{2} \|x_{n+1} - x_n\|^2 .$$

Mais $x_{n+1} - x_n = -\rho_n \nabla J(x_n)$ et donc :

$$(\nabla J(x_{n+1}), x_{n+1} - x_n) = -\rho_n (\nabla J(x_{n+1}), \nabla J(x_n)) = 0 .$$

Il en résulte :

$$\frac{\alpha}{2} \|x_{n+1} - x_n\|^2 \leq J(x_n) - J(x_{n+1}) ,$$

ce qui, compte tenu du fait que la suite $(J(x_n))_n$ est décroissante, minorée donc convergente, montre que $\|x_{n+1} - x_n\| \rightarrow 0$.

Pour conclure, on utilise le caractère \mathcal{C}^2 de J et le théorème des accroissements finis : D^2J étant borné sur le convexe compact K , on a :

$$\begin{aligned} \|\nabla J(x_n)\|^2 &= (\nabla J(x_n), \nabla J(x_n) - \nabla J(x_{n+1})) \\ &\leq \|\nabla J(x_n)\| \|\nabla J(x_n) - \nabla J(x_{n+1})\| \\ &\leq C \|\nabla J(x_n)\| \|x_{n+1} - x_n\| . \end{aligned}$$

Et on en déduit que $\|\nabla J(x_n)\| \rightarrow 0$.

Enfin, par la propriété (ii) et le fait que $\nabla J(x_\infty) = 0$:

$$\begin{aligned} \frac{\alpha}{2} \|x_n - x_\infty\|^2 &\leq (\nabla J(x_n) - \nabla J(x_\infty), x_n - x_\infty) \\ &\leq \|\nabla J(x_n)\| \|x_n - x_\infty\| , \end{aligned}$$

et donc $\alpha \|x_n - x_\infty\| \leq \|\nabla J(x_n)\| \rightarrow 0$, ce qui conclut la démonstration.

Un inconvénient de la méthode de plus grande pente :

Si on utilise cette méthode pour minimiser dans \mathbb{R}^2 , la fonction :

$$f(x, y) = \frac{1}{2}(x^2 + 2y^2) ,$$

à chaque étape, on a :

$$x_{n+1} = x_n - \rho_n x_n ,$$

$$y_{n+1} = y_n - 2\rho_n y_n ,$$

avec $x_{n+1}x_n + 4y_{n+1}y_n = 0$ qui traduit $(\nabla J(x_{n+1}), \nabla J(x_n)) = 0$. Ce qui donne :

$$0 < \rho_n = \frac{x_n^2 + 2y_n^2}{x_n^2 + 4y_n^2} \leq 1 .$$

En examinant d'un peu plus près le processus de minimisation, on constate qu'on tourne autour du point de minimum $(0,0)$ sans jamais l'atteindre, chaque étape étant relativement inefficace...

Cet exemple très simple suggère deux remarques :

1. Peut-être a-t-on mal choisi le gradient ! On rappelle que le lien entre la différentielle de J (objet intrinsèque) et un gradient $\tilde{\nabla}J$ se fait via le choix d'un produit scalaire $\langle \cdot, \cdot \rangle$ par la formule de représentation :

$$J'(x)(h) = \langle \tilde{\nabla}J(x), h \rangle \quad \text{pour tout } h \in \mathbb{R}^N .$$

Donc le choix $\nabla J(x) := \left(\frac{\partial J}{\partial x_1}(x), \dots, \frac{\partial J}{\partial x_N}(x) \right)$ lié à celui du produit scalaire standard dans \mathbb{R}^N n'est pas le seul possible.

Choisir un produit scalaire revient à choisir une matrice A symétrique définie positive car on sait que, si $\langle \cdot, \cdot \rangle$ est un produit scalaire, il existe une telle matrice A pour laquelle on a, pour tous $x, y \in \mathbb{R}^N$:

$$\langle x, y \rangle = (Ax, y) .$$

Donc, pour tout $h \in \mathbb{R}^N$:

$$J'(x)(h) = \langle \tilde{\nabla}J(x), h \rangle = (A\tilde{\nabla}J(x), h) ,$$

et on en déduit que $\nabla J(x) = A\tilde{\nabla}J(x)$.

Tous les gradients possibles se déduisent du gradient standard en appliquant une matrice symétrique définie positive. Le problème du choix du gradient s'appelle *problème de préconditionnement*; un bon choix de gradient peut accélérer la convergence.

Sur l'exemple ci dessus, avec le produit scalaire :

$$\langle (x, y), (x', y') \rangle = xx' + 2yy' ,$$

il est facile de voir que $\tilde{\nabla}f(x, y) = (x, y)$ et la méthode converge en 1 itération. Donc on a effectivement accéléré la convergence !

Le préconditionnement peut jouer un rôle par exemple pour des fonctions du type :

$$J(x) = \frac{1}{2}(Ax, x) + F(x) ,$$

quand A est une matrice symétrique définie positive, et surtout quand on pense que la fonction $F \in \mathcal{C}^1(\mathbb{R}^N)$ a un moindre effet sur la minimisation ("petite" perturbation). Évidemment utiliser le produit scalaire

donné par A est tentant mais il ne faut pas oublier que le calcul du nouveau gradient nécessitera à chaque étape la résolution du système linéaire $A\tilde{\nabla}J(x) = \nabla J(x)$ et souvent on préfère utiliser une matrice plus simple (par exemple, la diagonale de A ou une sous-matrice qui rend la résolution du système linéaire plus simple).

2. Le préconditionnement n'étant pas tout à fait évident et étant parfois coûteux, on peut se dire qu'on devrait mieux utiliser les informations "passées" pour mieux choisir la direction de descente ; par exemple, pourquoi ne pas utiliser $\nabla J(x_{n-1})$ en plus de $\nabla J(x_n)$ pour déterminer une meilleure direction de descente à partir de x_n ?

Cette idée conduit à la **méthode du gradient conjugué**, décrite dans la section suivante, qui consiste à utiliser des directions de descente différentes, notées w_n .

1.5.3 Troisième essai : méthode du gradient conjugué

Cette méthode peut être décrite comme suit :

Initialisation : On choisit un point de départ $x_0 \in \mathbb{R}^N$ et on pose $r_0 = w_0 = \nabla J(x_0)$ et on calcule x_1 par la méthode de plus grande pente.

Itération : x_n, r_{n-1}, w_{n-1} étant connus, on pose :

$$r_n = \nabla J(x_n)$$

$$w_n = r_n + \theta_n w_{n-1}$$

avec :

$$\theta_n = \frac{(r_n, r_n - r_{n-1})}{\|r_{n-1}\|^2}$$

$$x_{n+1} = x_n - \rho_n w_n$$

où (théoriquement ici aussi) ρ_n est solution du problème d'optimisation :

$$\tilde{J}(\rho_n) = \min_{\rho \geq 0} \tilde{J}(\rho) ,$$

où $\tilde{J}(\rho) = J(x_n - \rho w_n)$.

La méthode du gradient conjugué ressemble donc à la méthode de plus grande pente, à part qu'ici on utilise une direction de descente qui n'est plus le gradient. On va maintenant justifier l'utilisation de telles directions w_n avec des formules aussi curieuses.

1.5.4 Justification de la méthode du gradient conjugué

Si on suppose que l'on démarre proche de la solution x_∞ cherchée et si J est assez régulière, la formule de Taylor (en se souvenant que $\nabla J(x_\infty) = 0$) montre que :

$$J(x_\infty + h) = J(x_\infty) + \frac{1}{2}(D^2J(x_\infty)h, h) + o(\|h\|^2).$$

Au $o(\|h\|^2)$ près, on est donc essentiellement ramené à la minimisation d'une fonction quadratique avec une matrice $D^2J(x_\infty)$ positive puisque l'on est en un point de minimum.

On va donc se restreindre au cas des fonction J du type :

$$J(x) = \frac{1}{2}(Ax, x),$$

où la matrice A est symétrique définie positive.

On se fixe alors l'objectif suivant : connaissant x_n et w_{n-1} (direction de descente qui a servi à calculer x_n), on veut déterminer la meilleure direction de descente w_n à utiliser à partir de x_n , c'est-à-dire celle qui donne la valeur minimale de $J(x_{n+1})$, avec la restriction que w_n est dans le plan engendré par $r_n = Ax_n$ et w_{n-1} . En d'autres termes, on cherche θ_n tel que, si :

$$w_n = r_n + \theta_n w_{n-1},$$

on ait :

$$\min_{\rho \geq 0} J(x_n - \rho w_n) = \min_{\theta \in \mathbb{R}} \min_{\rho \geq 0} J(x_n - \rho(r_n + \theta w_{n-1})).$$

On rappelle tout d'abord que la minimisation de $\tilde{J}(\rho) = J(x_n - \rho w_n)$ qui a conduit au point x_n [détermination du ρ optimal] et l'égalité $\tilde{J}'(\rho_{n-1}) = 0$ se traduit par $(r_n, w_{n-1}) = 0$.

Ensuite on calcule $J(x_n - \rho w)$ pour $w = r_n + \theta w_{n-1}$:

$$J(x_n - \rho w) = J(x_n) - \rho(Ax_n, w) + \frac{1}{2}\rho^2(Aw, w).$$

Comme $Ax_n = r_n$, la propriété rappelée ci-dessus implique que $(Ax_n, w) = \|r_n\|^2$ et :

$$J(x_n - \rho w) = J(x_n) - \rho\|r_n\|^2 + \frac{1}{2}\rho^2(Aw, w).$$

Le minimum en ρ se calcule facilement pour ce polynôme de degré 2 et un calcul élémentaire permet de voir que pour rendre ce minimum en ρ minimal par rapport à θ , il suffit de minimiser par rapport à θ la quantité (Aw, w) . Pour cela, on la réécrit :

$$(Aw, w) = (Ar_n, r_n) + 2\theta(Ar_n, w_{n-1}) + \theta^2(Aw_{n-1}, w_{n-1}).$$

On s'intéresse d'abord à (Ar_n, w_{n-1}) ; comme $x_n = x_{n-1} - \rho_{n-1}w_{n-1}$ et que l'on peut supposer que $\rho_{n-1} \neq 0$, on a :

$$w_{n-1} = \frac{1}{\rho_{n-1}}(x_{n-1} - x_n) \quad \text{et donc} \quad Aw_{n-1} = \frac{1}{\rho_{n-1}}(r_{n-1} - r_n).$$

Il en résulte :

$$(Ar_n, w_{n-1}) = (r_n, Aw_{n-1}) = \frac{1}{\rho_{n-1}}(r_n, r_{n-1} - r_n).$$

D'autre part :

$$(Aw_{n-1}, w_{n-1}) = \frac{1}{\rho_{n-1}}(r_{n-1} - r_n, w_{n-1}) = \frac{1}{\rho_{n-1}}(r_{n-1}, w_{n-1}).$$

Mais comme $w_{n-1} = r_{n-1} + \theta_{n-1}w_{n-2}$ par l'étape précédente et que l'on a $(r_{n-1}, w_{n-2}) = 0$, on a finalement :

$$(r_{n-1}, w_{n-1}) = \|r_{n-1}\|^2 + \theta_{n-1}(r_{n-1}, w_{n-2}) = \|r_{n-1}\|^2.$$

Finalement :

$$(Aw, w) = (Ar_n, r_n) + \frac{1}{\rho_{n-1}}(2\theta(r_n, r_{n-1} - r_n) + \theta^2\|r_{n-1}\|^2),$$

et la minimisation en θ donne un optimum pour :

$$\theta_n = \frac{(r_n, r_n - r_{n-1})}{\|r_{n-1}\|^2}.$$

Ce qui justifie bien la formule annoncée.

L'efficacité de la méthode du gradient conjugué est démontrée par le résultat suivant :

Théorème 1.12. *Pour des fonctions du type :*

$$J(x) = \frac{1}{2}(Ax, x) - (b, x),$$

dans \mathbb{R}^N où A est une matrice symétrique définie positive et $b \in \mathbb{R}^N$, la méthode du gradient conjugué converge en moins de N itérations, quelle que soit la donnée initiale.

Remarque : Comme on sait que minimiser une telle fonction est équivalent (sous les hypothèses du théorème) à résoudre $Ax = b$, on a une méthode de résolution d'un système linéaire qui converge en N itérations.

Nous laisserons la preuve en exercice : elle est basée sur deux ingrédients ; d'abord, grâce à une simple translation en x , on se ramène au cas où b est nul (on utilise une translation par la solution unique de $Ax = b$) ; le point

de minimum devient donc 0. Puis on applique le lemme suivant dans lequel on utilise le produit scalaire naturel :

$$\langle x, y \rangle = (Ax, y) \quad \text{pour } x, y \in \mathbb{R}^N .$$

Les propriétés d'orthogonalité du lemme suivant font donc référence à ce produit scalaire.

Lemme 1.6. *Pour tout $n \geq 0$, x_{n+1} est le vecteur de l'espace engendré par $x_n, w_n, w_{n-1}, \dots, w_0$ qui est orthogonal à $E_{n+1} = \text{Vect}(w_n, w_{n-1}, \dots, w_0)$ et tel que $x_{n+1} - x_n \in E_{n+1}$. De même, w_{n+1} est le vecteur de l'espace engendré par $r_{n+1}, w_n, w_{n-1}, \dots, w_0$ qui est orthogonal à E_{n+1} et tel que $w_{n+1} - r_{n+1} \in E_{n+1}$.*

La méthode du gradient conjugué s'apparente donc à un procédé d'orthogonalisation de Gram-Schmidt pour le produit scalaire adapté $\langle \cdot, \cdot \rangle$. Le résultat découle immédiatement du lemme puisque les w_k pour $0 \leq k \leq N-1$ sont linéairement indépendants et donc $E_{N-1} = \mathbb{R}^N$ et x_{N-1} qui est orthogonal à E_{N-1} est donc nul (c'est donc le point de minimum de J).

1.6 Un résultat de convergence pour le schéma numérique

Nous avons décrit plusieurs méthodes pour calculer numériquement une approximation U (qui dépend de Δx) de la solution du problème de Dirichlet. Il est maintenant temps de se demander en quel sens U approche la solution de (1)-(2). La dépendance en Δx devenant ici importante, on rappelle que l'on note aussi h la quantité Δx et on notera $U^h = (u_j^h)_j$ la solution discrète associée.

On pose $E_h := \max_{1 \leq j \leq N} |u(jh) - u_j^h|$; E_h est erreur commise avec un pas de discrétisation h , mesurée avec la norme du sup.

On a le résultat de convergence suivant :

Théorème 1.13. *On a :*

- Si $f \in \mathcal{C}^2([0, 1])$, $E_h = O(h^2)$.
- Si $f \in \mathcal{C}^1([0, 1])$, $E_h = O(h)$.
- Si $f \in \mathcal{C}([0, 1])$, $E_h \rightarrow 0$ quand $h \rightarrow 0$.

Le sens de ce résultat est très clair : on a un schéma d'ordre 2 et la convergence est effectivement d'ordre 2 (en $O(h^2)$) si la solution est suffisamment régulière. Car l'énoncé devrait être (on le verra mieux dans la preuve) :

- Si $u \in \mathcal{C}^4([0, 1])$, $E_h = O(h^2)$.
- Si $u \in \mathcal{C}^3([0, 1])$, $E_h = O(h)$.
- Si $u \in \mathcal{C}^2([0, 1])$, $E_h \rightarrow 0$ quand $h \rightarrow 0$.

Cette régularité de u intervient via la consistance : contrairement à ce que nous avons fait dans la partie sur les schémas numériques où la consistance était “testée” à l’aide de fonctions \mathcal{C}^∞ quelconques, nous allons ici injecter la solution elle-même dans le schéma et, suivant sa régularité, nous obtiendrons une évaluation de l’erreur de consistance assez différente.

Preuve : On procède en plusieurs étapes.

Etape 1 : estimation de l’erreur dans le schémas. On se place dans le cas où $f \in \mathcal{C}^2([0, 1])$. On va poser $\tilde{U} := (\tilde{u}_j)_j$ où $\tilde{u}_j = u(jh)$ pour $1 \leq j \leq N$, en utilisant aussi la convention $\tilde{u}_0 = \tilde{u}_{N+1} = 0$. \tilde{U} désigne donc l’élément de \mathbb{R}^N correspondant à la solution exacte.

Nous allons évaluer l’erreur commise dans le schéma en remplaçant la “vraie” solution $U = U^h$ par \tilde{U} , i.e. on va évaluer :

$$-\frac{\tilde{u}_{j+1} + \tilde{u}_{j-1} - 2\tilde{u}_j}{(\Delta x)^2} - f_j.$$

On va faire un calcul de type consistance mais d’une manière un peu plus précise. Si $D_j := \tilde{u}_{j+1} + \tilde{u}_{j-1} - 2\tilde{u}_j$, on a :

$$D_j = (u(x_{j+1}) - u(x_j)) - (u(x_j) - u(x_{j-1})) = \int_{x_j}^{x_{j+1}} u'(t) dt - \int_{x_{j-1}}^{x_j} u'(t) dt.$$

On va maintenant intégrer par parties en se souvenant que u est de classe \mathcal{C}^4 puisque f est de classe \mathcal{C}^2 .

$$\begin{aligned} D_j &= [(t - x_{j+1})u'(t)]_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} (t - x_{j+1})u''(t) dt \\ &\quad - [(t - x_{j-1})u'(t)]_{x_{j-1}}^{x_j} + \int_{x_{j-1}}^{x_j} (t - x_{j-1})u''(t) dt. \end{aligned}$$

On remarque que les termes tous intégrés s’éliminent et on recommence :

$$\begin{aligned} D_j &= -\left[\frac{(t - x_{j+1})}{2}u''(t)\right]_{x_j}^{x_{j+1}} + \int_{x_j}^{x_{j+1}} \frac{(t - x_{j+1})^2}{2}u'''(t) dt \\ &\quad + \left[\frac{(t - x_{j-1})^2}{2}u''(t)\right]_{x_{j-1}}^{x_j} - \int_{x_{j-1}}^{x_j} \frac{(t - x_{j-1})^2}{2}u'''(t) dt. \end{aligned}$$

Un calcul des termes tous intégrés et une dernière intégration par parties donne :

$$\begin{aligned} D_j &= h^2 u''(x_j) - \int_{x_j}^{x_{j+1}} \frac{(t - x_{j+1})^3}{6} u^{(4)}(t) dt \\ &\quad + \int_{x_{j-1}}^{x_j} \frac{(t - x_{j-1})^3}{6} u^{(4)}(t) dt. \end{aligned}$$

En majorant les deux intégrales par l'inégalité $|\int \cdots| \leq \int |\cdots|$ puis $|u^{(4)}(t)|$ par $\|u^{(4)}\|_\infty = \|f''\|_\infty$, on trouve finalement que :

$$D_j = h^2 u''(x_j) + \tilde{e}_j^h,$$

avec :

$$|\tilde{e}_j^h| \leq \frac{h^4}{12} \|f''\|_\infty.$$

Et donc :

$$(1.17) \quad -\frac{1}{h^2} A\tilde{U} = F + e^h,$$

avec $e^h = (e_j^h)_j$ où $e_j^h = \frac{1}{h^2} \tilde{e}_j^h$. On a donc :

$$(1.18) \quad \max_{1 \leq j \leq N} |e_j^h| \leq \frac{h^2}{12} \|f''\|_\infty.$$

On notera λ^h la quantité $\frac{h^2}{12} \|f''\|_\infty$.

Etape 2 : erreur de consistance et estimation d'erreur. On veut déduire de (1.17)-(1.18) une estimation de $\tilde{U} - U$ pour la norme du sup (et pas seulement en norme euclidienne). On procède comme pour la stabilité du schéma en introduisant W qui sera ici la solution associée à $G = (1)_j$ (toutes les coordonnées de G sont les mêmes et elles sont égales à 1). Ce schéma est associé au problème de Dirichlet :

$$-w'' = 1 \quad \text{avec } w(0) = w(1) = 0,$$

dont la solution se calcule facilement :

$$w(x) = \frac{1}{2} x(1-x).$$

Des calculs assez simples montrent qu'en fait, on a une résolution exacte du système linéaire dans ce cas particulier et que $W = (w(x_j))_j$.

Un simple calcul montre que :

$$-\frac{1}{h^2} A(\tilde{U} - U + \lambda^h W) \geq 0,$$

où nous rappelons que cette inégalité signifie que toutes les coordonnées du vecteur $-\frac{1}{h^2} A(\tilde{U} - U + \lambda^h W)$ sont positives ou nulles. De même,

$$-\frac{1}{h^2} A(U - \tilde{U} + \lambda^h W) \geq 0.$$

En utilisant le principe du maximum discret, on en déduit :

$$\tilde{U} - U + \lambda^h W \geq 0,$$

et :

$$U - \tilde{U} + \lambda^h W \geq 0 ,$$

ce qui conduit à :

$$\max_{1 \leq j \leq N} |u_j^h - \tilde{u}_j| \leq \lambda^h \max_{1 \leq j \leq N} |w(x_j)| \leq \frac{\lambda^h}{8} .$$

Ceci termine la preuve dans le cas \mathcal{C}^2 où l'on a une estimation complètement explicite.

Étape 3 : les cas $f \in \mathcal{C}([0, 1])$ et $f \in \mathcal{C}^1([0, 1])$. Plusieurs preuves sont possibles : soit une preuve directe en modifiant la manière d'estimer $-\frac{1}{h^2}A\tilde{U}-F$, soit par approximation.

Nous choisissons la deuxième méthode qui consiste à approcher f par une suite $(f_\varepsilon)_\varepsilon$ de fonctions de classe \mathcal{C}^2 ; on note u_ε la solution du problème de Dirichlet associée à f_ε et U_ε la solution du schéma associée à F_ε avec des notations naturelles. De même on introduit \tilde{U}_ε et E_h^ε l'erreur analogue à E_h mais entre la solution u_ε et U_ε .

Par l'inégalité triangulaire, on a :

$$E_h \leq \|\tilde{U} - \tilde{U}_\varepsilon\|_\infty + E_h^\varepsilon + \|U - U_\varepsilon\|_\infty ,$$

et on examine les termes du membre de droite : par le principe du maximum pour l'équation continue, on a :

$$\|\tilde{U} - \tilde{U}_\varepsilon\|_\infty \leq \|u - u_\varepsilon\|_\infty \leq \frac{1}{8}\|f - f_\varepsilon\|_\infty .$$

De même, en recopiant l'argument de l'étape 2 ci-dessus :

$$\|U - U_\varepsilon\|_\infty \leq \frac{1}{8}\|F - F_\varepsilon\|_\infty \leq \frac{1}{8}\|f - f_\varepsilon\|_\infty .$$

Enfin, par le résultat des étapes 1 et 2, on a :

$$E_h^\varepsilon \leq \frac{h^2}{96}\|f_\varepsilon''\|_\infty .$$

D'où :

$$E_h \leq \frac{1}{4}\|f - f_\varepsilon\|_\infty + \frac{h^2}{96}\|f_\varepsilon''\|_\infty .$$

Une suite $(f_\varepsilon)_\varepsilon$ étant choisie, il s'agira ensuite de déterminer le meilleur ε possible, c'est-à-dire celui qui donne la meilleure estimation, donc le membre de droite le plus petit.

Pour cela, précisons le choix de la suite $(f_\varepsilon)_\varepsilon$. On peut d'abord prolonger la fonction f à \mathbb{R} tout entier en une fonction continue ou \mathcal{C}^1 qui est à support compact. On régularise alors f par convolution :

$$f_\varepsilon(x) = \int_{\mathbb{R}} f(y)\rho_\varepsilon(x-y)dy ,$$

où $(\rho_\varepsilon)_\varepsilon$ est une suite d'approximation de l'unité, typiquement :

$$\rho_\varepsilon(t) = \frac{1}{\varepsilon} \rho\left(\frac{t}{\varepsilon}\right) \quad \text{pour tout } t \in \mathbb{R},$$

où ρ est une fonction \mathcal{C}^∞ positive, paire, à support compact dans $] -1, 1[$ et telle que $\int_{\mathbb{R}} \rho(y) dy = 1$.

Il est ⁽¹²⁾ bien connu que, si f est de classe \mathcal{C}^1 alors

$$\|f - f_\varepsilon\|_\infty \leq C\varepsilon \quad \text{et} \quad \|f'_\varepsilon\|_\infty \leq \frac{C}{\varepsilon},$$

pour une certaine constante C dépendant de f et de ρ , alors que si f est seulement continue :

$$\|f - f_\varepsilon\|_\infty \leq \chi(\varepsilon) \quad \text{et} \quad \|f'_\varepsilon\|_\infty \leq \frac{C}{\varepsilon^2},$$

où χ est un module de continuité de f , donné par exemple par :

$$\chi(t) = \sup_{|x-y| \leq t} \|f(x) - f(y)\|.$$

Si f est de classe \mathcal{C}^1 , on a donc :

$$E_h \leq \frac{1}{4} C\varepsilon + \frac{h^2}{96} \frac{C}{\varepsilon},$$

et le minimum en ε est atteint pour $\varepsilon = ch$, c étant une constante explicite, et on a bien $E_h = O(h)$.

Si f est seulement continu, on a :

$$E_h \leq \frac{1}{4} \chi(\varepsilon) + \frac{h^2}{96} \frac{C}{\varepsilon^2};$$

le calcul du ε réalisant le minimum n'est plus possible mais en prenant $\varepsilon = h^{1/2}$, on voit bien que le second membre tend vers 0.

1.7 Annexes : Rappels d'Analyse Hilbertienne

On note H un espace de Hilbert, (\cdot, \cdot) son produit scalaire et $\|\cdot\|$ la norme associée. On rappelle la propriété caractéristique de la norme, *l'égalité de la médiane* :

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2) \quad \text{pour tous } x, y \in H.$$

Le résultat fondamental est le suivant :

(12). ou devrait être ... le lecteur est vivement encouragé à redémontrer ces estimations en guise d'exercice.

Théorème 1.14. (*Théorème de projection*) Soit K un sous-ensemble convexe fermé, non vide de H et soit $x \in H$. Il existe un unique point $y \in K$ tel que :

$$\|x - y\| = \min_{z \in K} \|x - z\|.$$

De plus, y est l'unique solution dans K de l'inéquation variationnelle :

$$(x - y, z - y) \leq 0 \quad \text{pour tout } z \in K.$$

Preuve : $\{\|x - z\|; z \in K\}$ est un ensemble non vide minoré de \mathbb{R} , il admet donc une borne inférieure $m = \inf_{z \in K} \|x - z\|$ et par définition de la borne inférieure, il existe une suite minimisante $(z_n)_n$ d'éléments de K , c'est-à-dire telle que $\|x - z_n\| \rightarrow m$.

En appliquant l'égalité de la médiane à " x " = $x - z_n$ et " y " = $x - z_p$ pour $n, p \in \mathbb{N}$, on obtient :

$$\|x_n - x_p\|^2 = 2(\|x - z_n\|^2 + \|x - z_p\|^2) - \|2x - z_n - z_p\|^2.$$

Mais :

$$\|2x - z_n - z_p\|^2 = 4\left\|x - \frac{z_n + z_p}{2}\right\|^2,$$

et, comme K est convexe, $\frac{z_n + z_p}{2} \in K$ donc $\|x - \frac{z_n + z_p}{2}\| \geq m$. Il en résulte :

$$\|x_n - x_p\|^2 \leq 2(\|x - z_n\|^2 + \|x - z_p\|^2) - 4m^2; .$$

En utilisant le fait que $\|x - z_n\|^2, \|x - z_p\|^2 \rightarrow m^2$, on déduit de cette inégalité que la suite $(z_n)_n$ est de Cauchy donc elle converge vers un élément y qui est dans K car les z_n sont dans K et K est fermé. Par passage à la limite, en utilisant la continuité de la norme, on obtient $\|x - y\| = m$ puisque $\|x - z_n\|$ converge à la fois vers m (par définition) et vers $\|x - y\|$ (par continuité de la norme). Une superbe application de l'unicité de la limite!

Pour l'unicité, on raisonne par l'absurde : s'il existe $y, y' \in K$ tels que $\|x - y\| = \|x - y'\| = m$, on réutilise l'égalité de la médiane avec " x " = $x - y$ et " y " = $x - y'$; par le même argument que ci-dessus, on déduit :

$$\|y - y'\|^2 \leq 2(\|x - y\|^2 + \|x - y'\|^2) - 4m^2 = 0.$$

D'où l'unicité.

Pour l'inéquation variationnelle, on "teste" l'inégalité $\|x - y\|^2 \leq \|x - z\|^2$ pour tout $z \in K$, en changeant z en $tz + (1 - t)y$ pour $t \in]0, 1[$. Cette combinaison convexe reste dans K puisque $y, z \in K$ et, en écrivant :

$$\|x - (tz + (1 - t)y)\|^2 = \|x - y + t(z - y)\|^2 = \|x - y\|^2 + 2t(x - y, z - y) + t^2\|z - y\|^2,$$

on déduit :

$$2t(x - y, z - y) + t^2\|z - y\|^2 \geq 0.$$

Il suffit de diviser par $t > 0$ et de faire tendre t vers 0 pour obtenir l'inéquation variationnelle.

Enfin, pour montrer que y est l'unique solution de cette inéquation variationnelle (I.V.), on procède par l'absurde en considérant un autre élément y' de K qui la satisfait. En prenant $z = y'$ dans l'I.V. de y et $z = y$ dans celle de y' et en les sommant, on a :

$$(x - y, y' - y) + (x - y', y - y') \geq 0 .$$

Mais le membre de gauche vaut $-||y - y'||^2$. Contradiction. □

Remarque : Si F est un sous-espace fermé de H , on peut projeter sur F et si on note $y = p_F(x)$, on montre facilement que p_F est linéaire, continue et que $||p_F(x_1) - p_F(x_2)|| \leq ||x_1 - x_2||$, pour tous $x_1, x_2 \in H$, cette dernière inégalité étant aussi vraie quand on projète sur un convexe fermé quelconque.

Théorème 1.15. (*Théorème de représentation de Riesz*) Si L est une forme linéaire continue sur H , il existe un unique élément $a \in H$ tel que :

$$L(x) = (a, x) \quad \text{pour tout } x \in H .$$

Preuve : On note $F = \ker(L)$. F est un sous-espace fermé car L est continu. Si L n'est pas identiquement nul (sinon $a = 0$ convient), il existe $x \in H$ tel que $L(x) \neq 0$. On note y la projection de x sur F . L'inéquation variationnelle implique que :

$$(x - y, z - y) \geq 0 \quad \text{pour tout } z \in F ,$$

mais F étant un sous-espace vectoriel, on peut prendre $z = y + h$, h décrivant F puis on peut changer h en $-h$. On aboutit donc à la relation d'orthogonalité :

$$(x - y, h) \geq 0 \quad \text{pour tout } h \in F .$$

On note $b = x - y$; comme $L(b) = L(x) \neq 0$, $b \neq 0$ et on va maintenant prouver le :

Lemme 1.7. $H = F \oplus \mathbb{R}b$.

Preuve : Si $z \in H$, un calcul immédiat montre que $z - \frac{L(z)}{L(b)}b$ est dans F et donc $z = \frac{L(z)}{L(b)}b + \bar{z}$ avec $\bar{z} \in F$ et cette décomposition est unique car le coefficient de b est imposé par le calcul de $L(z)$. □

On pose alors $a = \frac{L(x)}{||b||^2}b$ et on utilise la décomposition d'un z quelconque de H comme ci-dessus. Comme $(a, \bar{z}) = 0$ puisque $\bar{z} \in F$ et donc $(b, \bar{z}) = 0$, on déduit :

$$(a, z) = \frac{L(z)}{L(b)}(a, b) = \frac{L(z)}{L(b)} \frac{L(x)}{||b||^2}(b, b) = L(z) ,$$

puisque $L(b) = L(x)$. \square

Remarque : Ce théorème sert à résoudre des équations : par exemple, c'est le théorème de Lax-Milgram. Exemple : si $f \in L^2(]0, 1[)$ et si $H = H_0^1(]0, 1[)$, en utilisant le Théorème de Riesz pour le produit scalaire :

$$(u, v) = \int_0^1 (u'(t)v'(t) + u(t)v(t)) dt ,$$

et la forme linéaire continue $L(u) = \int_0^1 f(t)u(t) dt$, on résout formellement $-u'' + u = f$ avec $u(0) = u(1) = 0$.

1.7.1 Convergence faible

Définition 1.6. On dit que la suite $(x_n)_n$ d'éléments de H converge faiblement vers $l \in H$ si, pour toute forme linéaire continue L , $(L(x_n))_n$ converge vers $L(l)$ ou de manière équivalente si, pour tout $a \in H$, $(a, x_n) \rightarrow (a, l)$. Si $(x_n)_n$ converge faiblement vers l , on note $x_n \rightharpoonup l$.

Exemples (sympathiques ou moins sympathiques) :

1. En dimension finie, on peut utiliser pour L les formes linéaires coordonnées, $x = (x_1, \dots, x_N) \mapsto x_i$ pour $1 \leq i \leq N$ et donc la convergence faible implique la convergence de toutes les coordonnées donc la convergence classique. *Nihil novi sub sole!*⁽¹³⁾

2. En dimension infinie, ça se gâte : si $H = l^2(\mathbb{N}) := \{y = (y_n)_n ; \sum_{i=0}^{+\infty} y_n^2 < +\infty\}$, muni de la norme :

$$\|y\|^2 := \sum_{i=0}^{+\infty} y_n^2 ,$$

on note $e_n = (0, \dots, 0, 1, 0 \dots 0, \dots)$ la suite qui ne contient que des 0 sauf un 1 à la nième place. $(e_n)_n$ est une base hilbertienne de H . On a simultanément :

- $\|e_n\| = 1$ pour tout n ,
- $e_n \rightharpoonup 0$: en effet, si $a = (a_n)_n \in H$, $(a, e_n) = a_n \rightarrow 0$ car la convergence de la série implique que a_n tend vers 0.

Cet exemple où l'on a à la fois $\|e_n\| \equiv 1$ et $e_n \rightharpoonup 0$ montre bien la différence entre convergence faible et convergence forte (i.e. en norme), et la difficulté de manipuler une telle notion.

3. Autre exemple, dans $H = L^2(]0, 1[)$, la suite de fonctions $u_n(t) = \sin(2\pi nt)$ converge faiblement vers 0 (Lemme de Riemann-Lebesgue) et pourtant sa norme L^2 reste constante (exercice!).

(13). rien de nouveau sous le soleil.

Remarque : Il faut faire bien attention à la définition : on doit considérer les suites (a, x_n) pour a FIXÉ. Si $(x_n)_n$ et $(y_n)_n$ convergent faiblement, on ne sait rien a priori de (x_n, y_n) : on en a vu des exemples ci-dessus avec $y_n = x_n$. De même toute opération non linéaire sur les suites convergeant faiblement sont problématiques (cf. ci-dessus $\sin(2\pi nt)$ qui converge faiblement vers 0 mais pas $\sin^2(2\pi nt)$!).

Nous passons maintenant au résultat de compacité qui fait l'intérêt de l'introduction de la notion de convergence faible.

Théorème 1.16. *Dans un espace de Hilbert H , de toute suite bornée, on peut extraire une sous-suite qui converge faiblement. En d'autres termes, les bornés sont précompacts pour la topologie faible.*

Ce théorème montre qu'il y a une différence fondamentale entre la topologie forte (la topologie usuelle de la convergence au sens de la norme) et la topologie faible car un autre théorème de Riesz nous dit que les bornés sont précompacts pour la topologie forte si et seulement si l'espace est de dimension finie. Bien sûr, le théorème 1.16 est admis dans le cadre de ce cours.

Dans le cadre de la résolution de problèmes d'optimisation, ce résultat nous donne (dans les bons cas...) de la compacité pour les suites minimisantes mais, compte tenu de la remarque ci-dessus, il n'est pas clair a priori que l'on sera capable de passer à la limite dans la fonctionnelle pour prouver que la limite faible d'une suite minimisante est le point de minimum recherché. Le paragraphe suivant fournit des outils utiles dans cette direction.

Quelques résultats utiles :

Proposition 1.9. *Soit $J : H \rightarrow \mathbb{R}$ une fonction convexe de classe \mathcal{C}^1 et $(x_n)_n$ une suite d'éléments de H qui converge faiblement vers l . Alors :*

$$\liminf_n J(x_n) \geq J(l).$$

En d'autres termes, J est sci pour la topologie faible.

Preuve : La convexité et le caractère \mathcal{C}^1 de J implique :

$$J(y) \geq J(x) + J'(x)(y - x) \quad \text{pour tous } x, y \in H,$$

et donc :

$$J(x_n) \geq J(l) + J'(l)(x_n - l),$$

et comme $J'(l)$ est une forme linéaire continue, $J'(l)(x_n - l) \rightarrow 0$ par la convergence faible. Il suffit donc de passer à la liminf. \square

Remarque : $x \mapsto \|x\|^2$ est convexe et de classe \mathcal{C}^1 , donc, si $x_n \rightharpoonup l$:

$$\liminf_n \|x_n\|^2 \geq \|l\|^2.$$

On en a vu deux exemples plus haut avec des inégalités strictes...

Proposition 1.10. *Si $x_n \rightharpoonup l$ et si $\|x_n\| \rightarrow \|l\|$ alors $x_n \rightarrow l$.*

En d'autres termes : convergence faible + convergence de la norme = convergence forte.

Preuve : On écrit simplement :

$$\|x_n - l\|^2 = \|x_n\|^2 - 2(l, x_n) + \|l\|^2$$

et on remarque que, dans le membre de droite $(l, x_n) \rightarrow (l, l) = \|l\|^2$ par la convergence faible alors que les deux autres termes convergent vers $\|l\|^2$. \square

Remarque : Ce critère de convergence forte est souvent utile quand les propriétés de compacité sont simplement suffisantes pour passer à la limite a sens de la convergence faible (via des sous-suites). On récupère la convergence forte a posteriori.

Exercice 10. *Soit $a : H \times H \rightarrow \mathbb{R}$ une forme bilinéaire continue, i.e. il existe une constante $M > 0$ telle que :*

$$|a(u, v)| \leq M\|u\| \cdot \|v\| \quad \text{pour tous } u, v \in H .$$

Montrer qu'il existe une application linéaire continue $A : H \rightarrow H$ telle que :

$$a(u, v) = (Au, v) , \quad \text{pour tous } u, v \in H .$$