Université catholique de Louvain

Analyse numérique 2

MATH2180
2005-2006

Alphonse Magnus,
Institut de Mathématique Pure et Appliquée,
Université Catholique de Louvain,
Chemin du Cyclotron,2,
B-1348 Louvain-la-Neuve
(Belgium)
(0)(10)473157 , magnus@inma.ucl.ac.be, http ://www.math.ucl.ac.be/~magnus/

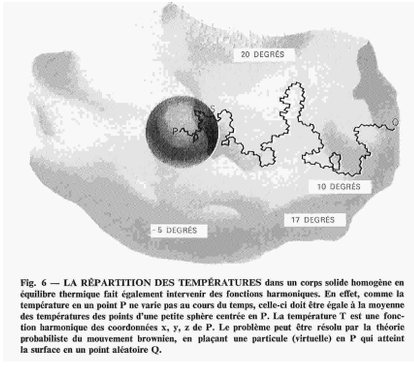


Fig. 6 — LA RÉPARTITION DES TEMPÉRATURES dans un corps solide homogène en équilibre thermique fait également intervenir des fonctions harmoniques. En effet, comme la température en un point P ne varie pas au cours du temps, celle-ci doit être égale à la moyenne des températures des points d'une petite sphère centrée en P. La température T est une fonction harmonique des coordonnées x, y, z de P. Le problème peut être résolu par la théorie probabiliste du mouvement brownien, en plaçant une particule (virtuelle) en P qui atteint la surface en un point aléatoire Q.

Table des matières

Quelques références. 7
0.1. Livres et articles 7
0.2. Ressources réseau 8
0.2.1. Usenet, FAQ, KuLeuven, Matlab, Netlib. 8
0.2.2. Packages. 9

Introduction. 17
1. Solutions d'EDP, caractéristiques. [Courant & Hilbert] 17
1.1. Premiers exemples 17
1.2. Problème de Cauchy 20
1.3. Un exemple exemplaire : équation de la membrane 21
1.4. EDP d'ordre 2 : caractéristiques 22

Chapitre 1. Équations elliptiques, formulations variationnelles. 27
1. Équations elliptiques et problèmes aux limites. 27
1.1. Solution donnée sur une frontière 27
1.2. Formulations variationnelles : introduction 28
1.3. Méthode numérique associée à une formulation variationnelle 30
1.4. Espaces de fonctions continûment dérivables par morceaux 30
1.5. Traitement complet d'un problème à une dimension 33
1.5.1. Formulation classique. 33
1.5.2. Formulation variationnelle forte. 33
1.5.3. Formulation variationnelle semi-faible. 33
1.5.4. Formulation variationnelle faible, ou distributionnelle. 34
1.6. EDP elliptiques et formulations variationnelles en dimension > 1 34
1.6.1. Où se trouve la solution d'une EDP elliptique ? 36
2. Formes coercives, problème de minimisation, méthode de Ritz-Galerkin. 37
2.1. Formulation variationnelle dans un espace vectoriel 37
2.2. Méthode de Ritz-Galerkin 42
3. Formes bilinéaires et opérateurs dans des espaces de Hilbert. 44
3.1. Espaces de Banach et de Hilbert 45
3.2. Le problème de l'existence de la solution. Théorème de Lax-Milgram. 47
4. Relations entre méthodes de projection : Galerkin, etc. 53
4.1. Galerkin 54
4.2. Ritz 54
4.3. Galerkin-Petrov 54
4.4. Moindres carrés 57
4.5. Collocation 57

Chapitre 2. Méthode des éléments finis. 58
1. Introduction et définition. 58
2. Ensembles unisolvants, interpolation. 59
3. Eléments unidimensionnels. 61
3.1. Interpolation linéaire par morceaux 61
3.2. Interpolation polynomiale de Lagrange 62
3.3. Interpolation polynomiale d'Hermite 63
3.4. Interpolation cubique d'Hermite par morceaux 63
4. Éléments bidimensionels. 76
4.1. Éléments rectangulaires 79
4.1.1. Éléments produits 79
4.1.2. L'élément "**serendipity**". 80
4.1.3. Hermite bicubique 81
4.2. Éléments triangulaires 82
4.2.1. Élément linéaire (Courant) 82
4.2.2. Triangle à six points 85
4.2.3. Autres éléments triangulaires d'interpolation 85
4.2.4. Triangle d'Hermite 86
4.2.5. Triangle d'Argyris 87
5. Résumé de ce chapitre. 87

Chapitre 3. Espaces de Sobolev, convergence. 89
1. Introduction et définitions. 89
1.1. Produit scalaire de Sobolev 89
1.2. Espaces de Sobolev 90
1.2.1. **Définition** 90
1.2.2. Identification à un espace de fonctions 90
2. Propriétés des espaces de Sobolev. 91
2.1. Formes définies sur $H^m(\Omega)$ 91
2.2. Formes bilinéaires et opérateurs ; dérivées faibles 95
2.3. Traces 97
3. Coercivité de $\displaystyle\int_\Omega \mathbf{grad}\ u \cdot\ \mathbf{grad}\ v\ d\boldsymbol{x}$ dans $H_0^1(\Omega)$. 100
4. Erreur d'approximation de la méthode de Ritz. 102
5. Méthodes d'éléments finis non conformes, "crimes variationnels". 104

Chapitre 4. Méthodes numériques d'obtention de l'approximation de Ritz. 110
1. Élaboration et résolution des équations. Conditionnement. 110
2. Un exemple de traitement en matlab. 113

Chapitre 5. Schémas de différences finies : problèmes elliptiques. 122
1. Opérateurs de prolongement et de restriction. 122
1.1. Normes 122
2. Approximation d'opérateurs. Consistance. 122
2.1. Définition 123
2.2. Discrétisations du laplacien 124
3. Solubilité des équations discrètes. Stabilité numérique. 125
3.1. Spectres de laplaciens discrétisés 126

3.2. Déterminant, constante de Catalan. 127
3.3. Consistance et stabilité numérique ⇒ convergence 127
4. Méthodes itératives de résolution numérique. Méthodes multigrilles. 128
4.1. Méthode de Jacobi 128
4.2. Méthodes multigrilles 129
5. Matrices d'inverses positives. Convergence. 138
5.1. Matrices positives, $M-$matrices 138
5.2. Application au laplacien 139
5.3. Processus stochastique 140
6. Autres méthodes de traitement du laplacien. 142
6.1. Transformation conforme (2D) 142
6.2. Equations intégrales sur la frontière, fonction de Green 143
6.3. Développements multipolaires 144
7. Décomposition de domaine, complément de Schur, substructuring. 146
8. Conditionnement et méthodes itératives, préconditionnement. 147

Chapitre 6. Schémas de différences finies : problèmes d'évolution. 151
1. Equation de la chaleur ; de la diffusion. 151
1.1. Equation de la chaleur 151
1.2. Solution par noyau de Poisson 152
1.3. Equation de la diffusion ; diffusion des euros 152
1.3.1. Effect of adding fresh coins. 153
1.4. Modes et séries de Fourier 155
1.5. Exemples de stabilité et instabilité numérique 155
2. Consistance et stabilité pour problèmes d'évolution $\partial u/\partial t + Mu = f$. 157
2.1. Problèmes bien posés. Opérateur solution. 157
2.2. Consistance et stabilité de discrétisations de problèmes d'évolution 159
3. Théorème d'équivalence de Lax. 159
4. Classe des équations paraboliques. 160
4.1. Examen de quelques schémas 161
4.2. Schémas à deux niveaux de temps 161
4.3. Schémas à plus de deux niveaux de temps 164
5. Equations hyperboliques. 168
5.1. Caractéristiques, domaine d'influence 168
5.2. Théorème 172
5.3. Stabilité numérique 174

Chapitre 7. Problèmes d'évolution : conditions de stabilité numérique. 180
1. Norme matricielle. 181
2. Quelques conditions suffisantes. 182
2.1. Norme $\leqslant 1 + \text{const.}\Delta t$ 182
2.2. Matrices symétriques, normales 182
2.3. Formes de Jordan et de Schur 182
3. Condition nécessaire de von Neumann. 183
4. Théorème de Kreiss. 183
4.1. Préparation 183
4.2. Le théorème de Kreiss 184

4.3. Preuve 184

Chapitre 8. Méthodes (pseudo) spectrales. 188
1. Fonctions propres d'opérateurs autoadjoints. 188
2. Calcul en représentation spectrale. 189
2.1. Ritz-Galerkin 189
2.2. Un problème de Trefethen 191
2.3. Méthode des tau 194
3. Calcul en représentation ponctuelle. 194
3.1. Méthode de collocation 195
3.2. Représentation matricielle des opérateurs différentiels 195
4. Exemples de conditions de stabilité numérique : 197

Index 198

Le cours passe en revue les principales méthodes de résolution numérique des équations aux dérivées partielles. Il se situe entre des cours consacrés à la théorie de ces équations et de leurs solutions

cf. `http ://www.ucl.ac.be/etudes/programmes/`

INMA 2325 Equations différentielles ordinaires [30-15], Q2, 4 créd., P. Habets J. Mawhin,
INMA 2345 Méthodes variationnelles, semi-groupes , [30-0], Q1, 3 créd., P. Habets,
MATH 2111 Analyse fonctionnelle, [30-15], Q1 , 5 créd., M. Willem,
MATH 2421 Analyse convexe et méthodes variationnelles [30-0], Q1, 3 créd., M. Willem,
MATH 2490 Problèmes aux limites pour les EDO et EDP [45-0], Q1, 4,5 créd., J. Mawhin,
MAPA 3037 Méthodes topologiques et variationnelles en analyse [30], Q1 + Q2, 2 créd., P. Habets, M. Willem.

et des cours orientés vers des applications spécifiques

MECA 2120 Introduction aux méthodes d'éléments finis† [30-30], Q1, 5 créd., V. Legat,
MECA 2170 Conception assistée par ordinateur en génie mécanique [30-30] , Q1, 5 créd., V. Legat,
MECA 2620 Simulation des phénomènes de transfert dans les procédés industriels [30-10], Q1 , 4 créd., F. Dupret,
MECA 2660 Méthodes numériques en mécanique des fluides [30-22,5], Q2 , 5 créd., G. Winckelmans,

Cela ne veut pas dire qu'il faut avoir suivi un ou des cours théoriques (cependant vivement recommandés, bien entendu), on établira (ou rappellera, plus ou moins bien) l'essentiel des bases théoriques nécessaires.

Au fait, le présent cours vaut 4,5 créd.

† Voir, à ce sujet,

`http ://labo.icampus.ucl.ac.be/MECA2120/` aussi
`http ://www.mema.ucl.ac.be/numlab/index.html`

extrait de
`http ://www.meca.ucl.ac.be/memawww/members/vl/figure.gif`

Ce cours a été créé par le Professeur ***Jean Meinguet*** qui assura son enseignement jusqu'en 1995. Au fil des années, l'accent fut mis sur les recherches contemporaines en méthodes de résolution de divers problèmes d'analyse fonctionnelle appliquée.

En 1973, le Professeur ***Jean Descloux***, de l'École Polytechnique Fédérale de Lausanne, vint donner à Louvain-la-Neuve un cours sur l'analyse mathématique de la méthode des ***éléments finis***, méthode qui commençait alors à être convenablement formalisée††. Les notes du Professeur Descloux forment encore l'essentiel de la première partie du présent cours.

†† Parmi les pionniers de la méthode, citons ***Fraeijs de Veubeke***, professeur à Liège et Louvain. Par ailleurs, notre Université reçut également les visites de ***P. Ciarlet*** et ***P. Raviart*** (Paris).

Quelques références.

0.1. Livres et articles.

R.A. Adams, *Sobolev Spaces*, Acad. Press, 1975.

K. Atkinson, W. Han, *Theoretical Numerical Analysis, A Functional Analysis Framework*, Springer-Verlag, 2001. Lectures in Mathematics - ETH Zürich

W. Bangerth, R. Rannacher, *Adaptive Finite Element Methods for Differential Equations*, Birkhäuser Verlag AG, 2003.

John P. Boyd, *Chebyshev and Fourier Spectral Methods*, First edition (out of print), Springer-Verlag (1989), 792 pp. Second edition, Dover, New York (2001), 688 pp. The online version is split into two files, *in* `http ://www-personal.engin.umich.edu/~jpboyd/BOOK_Spectral2000.html`

D. Braess, *Finite Elements. Theory, Fast Solvers and Applications in Solid Mechanics* , 2nd edition, Cambridge University Press 2001.

C. Canuto, M.Y. Hussaini, A. Quarteroni, T.A. Zang, *Spectral Methods in Fluid Dynamics*, Springer, 1988.

P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, 1978.

L. Collatz, *The Numerical Treatment of Differential Equations*, 3ème éd., Springer, 1966.

R. Courant, D. Hilbert, *Methods of Mathematical Physics, II, Partial Differential Equations*, Interscience, 1962.

C. Cuvelier, J. Descloux, J. Rappaz, C. Stuart, B. Zwahlen, assistés par G. Caloz : *Éléments d'équations aux dérivées partielles pour ingénieurs*, 2 vol., Presses polytechniques romandes, Lausanne, 1988.

R. Dautray, J.L. Lions, *Analyse mathématique et calcul numérique*, vol. 2 : *Opérateur de Laplace*, vol. 3 : *Transformations, Sobolev, opérateurs*, vol. 4 : *Méthodes variationnelles*, vol. 6 : *Méthodes intégrales et numériques* (chap. XII, Méthodes numériques pour les problèmes stationnaires, chap. XIII, Approximation des équations intégrales par éléments finis), vol. 9 : *Évolution : numérique, transport* (chap. XX, Méthodes numériques pour les problèmes d'évolution), Masson, 1987, 1988.

C. de Boor, *Numerical Functional Analysis*, cours CS717, University of Wisconsin-Madison, `http ://www.cs.wisc.edu/~deboor/cs717.html`

Brenner, Susanne C., Scott, L. Ridgeway, *The Mathematical Theory of Finite Element Methods* (Texts in Applied Mathematics Vol 15), 2nd Ed 385 pages, Springer-Verlag New York Inc., Jul. 2002 [1].

Kenneth Eriksson, Donald Estep, Peter Hansbo, and Claes Johnson, *Computational Differential Equations*, Cambridge U.P., 1996.

S.K. Godunov, V.S. Ryabenki, *Theory of Difference Schemes*, North-Holland, 1964.

D. Gottlieb, S.A. Orszag, *Numerical Analysis of Spectral Methods : Theory and Applications*, SIAM Reg. Conf. Appl. Math., 1977.

B. Gustafsson, H.-O. Kreiss, J. Oliger, *Time Dependent Problems and Difference Methods*, Wiley, 1995.

[1] Réf. communiquée par C. Delfosse (MAP23, 2001-2002).

E. Hairer, Polycopiés du cours "Analyse II, Partie B", fichiers postscript dans
`http ://www.unige.ch/math/folks/hairer/polycop.html`
en particulier le chap. 2 : Maxima et minima relatifs et calcul de variations.

A. Iserles, *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, 1996.

C. Johnson, *Numerical solution of partial differential equations by the finite element method*, Cambridge University Press, 1987[2].

J.M.-S. Lubuma, M.K. Luhandjula, D.B. Reddy, éditeurs : *Problèmes variationnels en mathématiques appliquées, Annales de la Faculté des Sciences de l'Université de Kinshasa*, Numéro Spécial 3, 1997.

A. MacKinnon, Imperial College, Computational Physics – 3rd/4th Year Option,
`http ://www.sst.ph.ic.ac.uk/angus/Lectures/compphys/`

Serge Nicaise, *Analyse numérique et équations aux dérivées partielles. Cours et problèmes résolus*, Dunod, Paris, 2000.

P. Rabier, J.M. Thomas, *Exercices d'analyse numérique des équations aux dérivées partielles*, Masson, Paris, 1985.

J. Rappaz, M. Picasso, *Introduction à l'analyse numérique*, PPUR (Presses Polytechniques et Universitaires Romandes), Lausanne, 1998.

P.A. Raviart, J.M. Thomas, *Introduction à l'analyse numérique des équations aux dérivées partielles*, Masson, Paris, 1988.

B.D. Reddy, *Introductory Functional Analysis, With Applications to Boundary Value Problems and Finite Elements*, Springer, 1998.

R.D. Richtmyer, *Difference Methods for Initial-Value Problems*, Interscience, New York, 1957 ($2^{\text{ème}}$ édition : 1967 : Richtmyer & Morton, Wiley ; réimpression par Krieger en 1994).

G. Strang, G.J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, 1973. Réimpression récente chez Wellesley-Cambridge Press.

R.Temam, *Analyse numérique, résolution approchée d'équations aux dérivées partielles*, PUF, Paris, 1970.

V. Thomée, From finite differences to finite elements. A short history of numerical analysis of partial differential equations, *J. Comp. Appl. Math.* **128** (2001) 1-54.

Lloyd N. Trefethen, *Finite Difference and Spectral Methods for Ordinary and Partial Differential Equations*, unpublished text, 1996, available at **http ://web.comlab.ox.ac.uk/oucl/work/nick.trefethen/p**

Lloyd N. Trefethen, *Spectral Methods in MATLAB*, SIAM, 2000.

O.C. Zienkiewicz, *The Finite Element Method in Engineering Science*, McGraw-Hill, 1971.

0.2. Ressources réseau.

0.2.1. *Usenet, FAQ, KuLeuven, Matlab, Netlib.*

[Usenet] `news :sci.math.num-analysis`

[FAQ] Numerical Analysis and Associated Fields Resource Guide, by Steve Sullivan (Mathcom, Inc.), `http ://www.mathcom.com/corpdir/techinfo.mdir/scifaq/index.html`

où on regardera :

Partial Differential Equations (PDEs) and Finite Element Modeling (FEM)
PDE and FEM Web Sites
Newsgroups for PDE and FEM
Books and References for PDE and FEM

[2]Merci à S. Legrand (PHYS22, 1998-1999) pour cette référence.

Comparison of Meshing Software Packages
PDE and FEM Software on the net
Commercial Packages for FEM

Steve Sullivan Copyright 1995-2002 Mathcom info@mathcom.com Updated : July 12, 2002

[KULeuven] Fast and Robust Solvers for partial Differential Equations,
`http ://www.cs.kuleuven.ac.be/cwis/research/twr/research/topics/`

[Matlab] `http ://www.mathworks.com`, `ftp ://ftp.mathworks.com`, `http ://www.mathtools.net`

[Netlib] : très nombreux programmes, `http ://www.netlib.org/`

0.2.2. *Packages.*

[Java] `http ://www.steven.pop.net.tw/javamesh/`

From : William Schiesser <wes1@lehigh.edu>
Date : Sun, 08 Feb 2004 10 :16 :10 -0500
Subject : Java ODE/PDE Routines

The Java ODE/PDE routines in the book "Ordinary and Partial Differential Equation Routines in C, C++, Fortran, Java, Maple and Matlab", CRC Press, 2004, have been slightly revised and reorganized so that the applications are now in single, self-contained subdirectories.

Information for acquiring these routines and all of the others in the book (gratis) is available from :

`http ://www.lehigh.edu/~wes1/wes1.html`

[AFEPack] From : W. B. Liu <W.B.Liu@ukc.ac.uk>
Date : Mon, 12 May 2003 07 :46 :55 +0100 (BST)
Subject : Availability of Adaptive Finite Element Library
We are delighted to announce the availability of the finite element library AFEPack 1.6 to the numerical analysis community. AFEPack is a generic C++ adaptive finite element library. Comparing with the existing finite element libraries, its contribution is twofold :
First, it introduces a framework of multi-meshes. That means you can approximate multi-variables using very different adaptive meshes, and this could save significant computational work. This feature is particularly important when these variables have very different computational difficulties.
Second, you can easily build very flexible finite element spaces with AFEPack.
The package has been tested on elliptic problems, parabolic problems, flow problems, and several optimal control problems. Detailed documents are included with the package. AFEPack 1.6 now can be downloaded from
`http ://www.kent.ac.uk/cbs/staff/homepage/wbl/data/soft.htm` or obtained by sending an e-mail to w.b.liu@ukc.ac.uk. The document now can browsing at `http ://162.105.68.168/AFEPack`

Dr. R Li School of Mathematics Peking University, China rli@math.pku.edu.cn

Prof. WB Liu CBS & IMS University of Kent Canterbury, CT2 7PE UK w.b.liu@ukc.ac.uk

[AMGToolbox] From : Jane Cullum <cullumj@lanl.gov>
Date : Fri, 11 Jul 2003 17 :56 :45 -0600 (MDT)
Subject : Algebraic Multigrid MATLAB Code

The following Software is available for informal distribution. The code assumes some familiarity with basic algebraic multigrid algorithms. It is being made available as is.

Name of Software : AMGToolbox

Description of Software :

AMGToolbox is a research tool for experimenting with various ideas for algebraic multigrid algorithms. It provides the infrastructure for testing various ideas for the different components of a typical algebraic multigrid (AMG) algorithm :

1. the Coarse/Fine split of variables
2. the construction of Prolongation and Restriction Operators
3. the construction of the Coarse Problems at each level of the AMG hierarchy.
4. the use of different Smoothers

It provides simple software tools for automatically collecting the results of sets of experiments over user-specified choices of parameters, and for generating corresponding tables of results as Tex files for use in talks and papers.

AMGToolbox is written in the prototyping language, MATLAB.

Authors : Menno Verbeek Jane Cullum Wayne Joubert

[AWFD] Date : Wed, 6 Aug 2003 19 :06 :18 +0200

Summary : A new C++ class library for wavelet based solvers for PDEs and integral equations.

We are announcing the availibility of our AWFD (Adaptivity, Wavelets & Finite Differences) software package. It is a C++ class library for wavelet/interpolet-based solvers for PDEs and integral equations.

The main features of AWFD are :
* Petrov-Galerkin discretizations of linear and non-linear elliptic and parabolic PDE (scalar as well as systems)
* Adaptive sparse grid strategy for a higher order interpolet multiscale basis
* Adaptivity control via thresholding of wavelet coefficients
* Multilevel lifting-preconditioner for linear systems
* Dirichlet and Neumann boundary conditions.

It consists of :
1. MATLAB functions for the generation of wavelet filter masks
2. Data structures for uniform, level-adaptive and fully adaptive trial spaces (i.e. grids)
3. Algorithms for the initialization and refinement of adaptive grids
4. Algorithms for (adaptive) wavelet transforms, finite difference- /collocation-/Galerkin-discretizations
5. Linear algebra
6. Solvers / preconditioners
7. IO functions with interfaces to e.g. MATLAB or VTK

The software can be downloaded from :
`http ://wissrech.iam.uni-bonn.de/research/projects/AWFD/index.html`

Department of Scientific Computing and Numerical Simulation University of Bonn Germany

[Chalmers] `http ://www.md.chalmers.se/Centres/Phi/research/` Chalmers finite element centre

[CONCEPTS] Date : Fri, 27 Aug 2004 22 :44 :18 +0200 Subject : Concepts 2.0.0 Released We proudly announce the initial public release of CONCEPTS [1] under the Open Source license GPL. CONCEPTS is a C++ class library for solving elliptic partial differential equations (PDEs) numerically. `http ://www.concepts.math.ethz.ch/`

[Deal] From : Guido Kanschat <guido.kanschat@iwr.uni-heidelberg.de> Date : Fri, 07 Jan 2005 18 :37 :40 +0100 Subject : Deal, Finite Element Library

Version 5.1 of the deal.II object-oriented finite element library has been released. It is available from the deal.II home-page at

http ://www.dealii.org

All main features of the previous versions have been continued and improved : - Support for dimension-independent programming - Extensive documentation and working example programs - Locally refined grids - A zoo of different finite elements - Fast linear algebra - Built-in support for symmetric multi-processing (SMP) - Output for a variety of visualization platforms.

deal.II can be downloaded for free and is distributed under an Open Source license.

[FEMLAB] `http ://www.femlab.com` , `http ://www.comsol.se` FEMLAB

Date : Thu, 17 Jul 2003 09 :47 :32 +0200

notre nouveau "Tour CD 2003" est disponible, qui compile les derniers modéles disponibles avec FEMLAB dans les domaines suivants :

- Acoustique, - CFD, - Matériaux, - Diffusion-réaction, - Electromagnétisme, - Electrostatique, - Couplage multiphysique, - Génie électrochimique, - Transport réactifs, - Dynamique des fluides couplée (électrostatique, transferts de chaleur...), - Propagation d'ondes électromagnéiques, - Quasi-statique en électromagnétisme, - Poutres et plaques, - Couplages multiphysiques en mcanique (PAC, piezo).

Si vous souhaitez recevoir ce CD (gratuitement), merci de laisser vos coordonnes précises sur : `http ://www.comsol.fr/tourcd/`

FEMLAB est un environnement de modélisation multiphysique de problèmes 1D, 2D et 3D, complètement intégré à MATLAB, utilisé en recherche, ingénierie et enseignement. Des démonstrations animées sont accessibles sur `http ://www.comsol.fr/showroom/tutorials/femlab.php`

besoin d'une solution rapidement ? `http ://www.femlab.com/support/knowledgebase/`

Date : Mon, 22 Dec 2003 16 :27 :54 +0100 Subject : New Version of FEMLAB

Today we announce the release of FEMLAB 3 – the next generation of the popular multiphysics modeling software. Visit our web pages for more about FEMLAB 3. Take this opportunity to order a free copy of the new product catalog and a free copy of the Spring Tour CD :

`http ://www.comsol.com/contact/request_form.php`

FEMLAB is used in research, product development and teaching, in such fields as : - Acoustics - Antennas - Bioscience - Bioengineering - Chemical reactions - Diffusion - Ecology - Electromagnetics - Environmental science - Fluid dynamics - Fuel cells - Geophysics - Heat transfer - Math/Applied PDEs - MEMS - Microwave engineering - Nanotechnology - Optics and photonics - Physics - Porous media flow - Quantum mechanics - Radio frequency components - Semiconductor devices - Structural mechanics - Transport phenomena - Wave propagation - Any combination of the above

Best regards,

Seema Khan COMSOL, Inc. 781-273-3322 www.comsol.com

[FIDISOL/CADSOL] `http ://www.uni-karlsruhe.de/~numerik/`,

FIDISOL/CADSOL is a program package for the solution of partial differential equations. 2- and 3-dimensional systems of elliptic (stationary) and parabolic (time-dependent) equations can be solved. The boundary conditions may be arbitrary. The solution method is the finite difference method.

More information is available on the web site : `http ://www.rz.uni-karlsruhe.de/~fidisol/`

[GeoFEM] From : Kengo Nakajima **nakajima@tokyo.rist.or.jp**
Date : Thu, 12 Jun 2003 09 :58 :21 +0900 Subject : GeoFEM Parallel FEM Platform

Dear Colleagues,

RIST (Research Organization for Information Science & Technology, Japan) has released the GeoFEM ver.6.0, a parallel finite element platform.

GeoFEM has been developed as the part of the Japanese national project, "Parallel Platform for Large Scale Solid Earth Simulation" funded by "Special Promoting Funds of Science & Technology" of the Ministry of Education, Science and Technology (MEXT), Japan which has been conducted since 1998.

The project just completed in the end of March, 2003. Ver.6.0 is the final version of GeoFEM platform.

GeoFEM includes parallel finite element codes for linear/nonlinear solid mechanics and thermal fluid simulations, parallel iterative linear solver library, partitioning subsystem, parallel visualization subsystem and utilities for parallel I/O and coupling of multiple codes. GeoFEM is originally developed for solid earth simulation but applicable for various types of engineering and science applications. Users also plug-in their own FEM codes to GeoFEM platform and can develop parallel FEM code easily.

GeoFEM is written in Fortran 90 and MPI (part of visualization/utility subsystems in C and C++) and can work on various types of platforms from LINUX clusters with Intel Fortran to massively parallel computers.

Researchers and engineers interested in GeoFEM can get information at the following website :

http ://geofem.tokyo.rist.or.jp/

Source files of the entire codes, example input/output files and documents can be downloaded from :

http ://geofem.tokyo.rist.or.jp/download_en/

GeoFEM will be inherited by a new project, "HPC Middleware (HPC-MW, http ://hpcmw.tokyo.rist.or.jp/)" under a national project, "Frontier Simulation Software for Industrial Science (http ://www.fsis.iis.u-tokyo.ac.jp/en/)" by MEXT, Japan. Various kinds of technologies developed in GeoFEM project will be available in HPC-MW.

[GETFEM] From : Yves Renard <renard@insa-toulouse.fr>
Date : Tue, 4 Mar 2003 15 :11 :37 +0100

GETFEM++ 1.4 Available for Download with the MATLAB interface.

The GETFEM++ project is centred around the development of a generic and efficient C++ library for elementary computations for finite element methods. The goal is to build a library allowing the computation of any elementary matrix (even for mixed methods) on the largest class

of methods and elements and for arbitrary dimension ($1 \leqslant n \leqslant 255$). It offers a complete separation between integration methods (exact or approximated), geometric transformations (linear or not) and finite element methods of arbitrary degrees.

Examples of available finite element methods : Pk on simplices in arbitray degrees and dimensions, Qk on parallelepipeds, P1, P2 with bubble functions, hierarchical elements ...

To define a new finite element method, it is sufficient to describe it on the reference element (mainly, to define base functions). Extensions are provided for Hermite elements, non polynomials and vectorial elements.

Moreover, the library includes other tools for computation with finite element such as assembly methods for classical PDEs, interpolation methods, computation of norms, mesh operations, boundary conditions ... This library allows to build finite elements codes which are completely independent of the dimension, the particular method or element. Examples are provided.

It is possible to obtain mass matrices (or other type of matrices) for finite element method defined on two different meshes, allowing mortar methods or certain mixed methods.

A MATLAB interface to this library is available and fully documented.

The release of GETFEM++ is under the LGPL license and may be downloaded at :
`http ://www-gmm.insa-toulouse.fr/getfem`

[GMM] Date : Mon, 15 Sep 2003 18 :18 :10 +0200 Subject : GMM++, A Generic C++ Matrix Library

GMM++ 1.5 Available for Download

This generic matrix library is built as an interface for already existing class of vectors and matrices. It provides also predefined type of dense, skyline and sparse matrices. It is largely inspired from MTL/ITL but we hope a little bit simpler to use. matrix x matrix multiplication works for any type of matrices (even with mixed types).

An important extension is the possibility to access to sub-matrices for read or write operations with any kind of interfaced matrix.

Linear solvers has been imported from ITL (CG, BICGSTAB, GMRES, QMR, dense LU, ...) with classical preconditionners (Incomplete cholesky, ILU, ILUT ...)

The performance is comparable with the one of MTL, slightly better on some solvers we optimized.

Inpout / output with standard formats is also provided.

This library is available under LGPL licenseand may be downloaded at :

`http ://www-gmm.insa-toulouse.fr/getfem/gmm_intro`

A short documentation exists.

Yves Renard (renard@insa-toulouse.fr) Dept de Mathematiques, INSA de Toulouse Complexe Scientifique de Rangueil 31077 Toulouse Cedex, FRANCE `http ://www-gmm.insa-toulouse.fr/~renard`

[GrAL] From : Guntram Berti `berti@ccrl-nece.de`
Date : Wed, 27 Feb 2002 14 :46 :19 +0100
Subject : Grid Algorithms Library

GrAL – Grid Algorithms Library version 0.2 now available at

`http ://www.math.tu-cottbus.de/~berti/gral`

GrAL is a generic library for grid (mesh) data structures and algorithms operating on them. Its contribution is twofold : First, it introduces a framework for decoupling algorithms from grid data structures, much like the C++ STL does for linear sequences. Second, it offers implementations of generic grid-related algorithms and data structures.

GrAL is open source. It is written in standard C++, and has been tested with g++ 2.95, 2.96 and 3.0.x on Linux platforms.

Enjoy !

– Guntram Berti

[IFISS] From : Alison Ramage <alison@maths.strath.ac.uk>
Date : Wed, 4 May 2005 17 :13 :47 +0100 (BST)
Subject : IFISS, Incompressible Flow and Iterative Solver Software

We are pleased to announce the availability of

Incompressible Flow & Iterative Solver Software (IFISS) Version 2.0.

This is an open-source Matlab software package that is associated with the forthcoming book *Finite Elements and Fast Iterative Solvers with applications in incompressible fluid dynamics* by Howard C. Elman, David J. Silvester, and Andrew J. Wathen (see `http ://www.oup.co.uk/isbn/0-19-852868-X` for further details).

The IFISS software can be used to generate typical linear systems arising from finite element discretisations of four important PDE applications : diffusion, convection-diffusion, Stokes flow and Navier-Stokes flow problems. It has built-in multigrid and Krylov subspace solvers and includes a variety of appropriate preconditioning strategies for each problem. We have used early versions of the software to suport technical workshops we have given in the last decade on fast solvers for incompressible flow problems.

Key features include

- implementation of a variety of mixed finite element approximation methods - automatic calculation of stabilization parameters where appropriate - a posteriori error estimation - a range of preconditioned Krylov subspace solvers (including MINRES and BICGSTAB(ell)) - a built-in geometric multigrid solver/preconditioner - an interface to the algebraic multigrid solver of FEMLAB (see `http ://www.comsol.com/`) - useful visualisation tools.

IFISS has been tested under Matlab Versions 5.3 to 7.0 and can be run under Windows, Unix and Mac architectures. The library is free software which can be redistributed and/or modified under the terms of the GNU Lesser General Public License as published by the Free Software Foundation. It can be downloaded from `http ://www.cs.umd.edu/~elman/ifiss.html`, `http ://www.manchester.ac.uk/ifiss`

David Silvester, Howard Elman and Alison Ramage.

[NetSolve] NA Digest Sunday, August 5, 2001 Volume 01 : Issue 29

From : Jack Dongarra `dongarra@cs.utk.edu` Date : Fri, 03 Aug 2001 13 :52 :33 -0400 Subject : NetSolve Version 1.4 Available

NetSolve grid based software system, version 1.4 is now available. The software can be downloaded from : `http ://icl.cs.utk.edu/netsolve`

The NetSolve project is being developed at the University of Tennessee's Innovative Computing Laboratory. The system provides NetSolve-enabled programs with transparent network access to computational resources via its many client application programming interfaces or APIs. In this version, client programs implemented in C, Fortran, Matlab and Mathematica can access

the NetSolve system and the hardware and software services it provides. These services include sophisticated numerical solvers from libraries like LAPACK, ScaLAPACK, PETSc, Aztec, SuperLU, etc. Facilities are available for creating new NetSolve services. All components have been tested on a variety of UNIX operating systems, including Linux, AIX, Irix, OSF, and Solaris. A client interface is available for the Microsoft Windows 2000 platform. There is a test grid deployed worldwide for experimentation that is accessible by anyone who has installed the new client interface. See `http ://icl.cs.utk.edu/netsolve` for further details and documentation.

From : Jack Dongarra <dongarra@cs.utk.edu>
Date : Tue, 14 Oct 2003 10 :08 :38 -0400
Subject : New Version of NetSolve

New Release of NetSolve/GridSolve Version 2.0

A new and more powerful version of the well-known grid middleware NetSolve/GridSolve has just been released. NetSolve 2.0, like its predecessors, is a client-server-agent system that enables users to solve complex scientific problems remotely using distributed resources on a computational grid. The system provides users access to both hardware and software computational resources distributed across a network. When a user submits a problem to the NetSolve agent, the agent searches the network of computational resources that has registered with it, chooses the best one (or set) available, solves the problem, and then returns the solution to the user. Load balancing for good performance and retry for fault-tolerance are handled automatically by the system. Version 2.0, which builds on NetSolve's traditional strengths, adds capabilities and robustness both to NetSolve's native environment and to its integration with other grid middleware, such as Condor-G. NetSolve's traditional strengths are its ease-of-use and its broad support for critical numerical software packages. For ease-of-use, NetSolve bindings exist for the most common scientific programming environments ; Fortran, C, Matlab, Mathematica. Moreover, many of the most important numerical and scientific libraries (e.g. Linear Algebra, Optimization, Fast Fourier Transforms) are easily integrated into the NetSolve server software, and users can also add their own special libraries. Finally, NetSolve has been designed to avoid restrictions on the type of software components that can be integrated into the system, and users with various types of applications have been able to incorporate their own packages into NetSolve with relative ease. Version 2.0 enhances these basic capabilities in two directions. There are enhancements to NetSolve's native environment to make it even easier to use, more capable, and more secure. And there are additions to NetSolve that enable it to more easily interoperate with, and leverage the power of, other leading grid environments. The most notable feature of the latter is the incorporation of GridRPC into NetSolve ; this expanded version of the NetSolve environment is called GridSolve. Grid RPC is an evolving standard from the Global Grid Forum (GGF).
NetSolve 2.0 is an open source package and can be freely downloaded from the ICL Web site `http ://icl.cs.utk.edu/netsolve/` Funding for the NetSolve effort is provided by grants from the National Science Foundation (NSF) through the NPACI and the NMI programs and the Department of Energy (DOE) through the MICS effort.

[OFELI] From : Rachid Touzani `Rachid.Touzani@math.univ-bpclermont.fr` Date : Mon, 14 May 2001 08 :59 :37 +0200 Subject : Object Finite Element LIbrary

A stabilized version of the OFELI library is now available.

OFELI (Object Finite Element LIbrary) is a library (or a toolkit) of C++ classes to build up finite element codes. The package contains numerous simple and less simple examples of finite element codes using OFELI. A significant documentation (html and pdf formats) is provided.

OFELI is free but Copyrighted software. It is distributed under the terms of the GNU General Public License (GPL).

The home page for the library is `http ://ofeli.sourceforge.net` A UNIX or Windows version can be downloaded from there.

To obtain further information contact : Rachid.Touzani@math.univ-bpclermont.fr

[OpenFEM] From : Dominique Chapelle ¡Dominique.Chapelle@inria.fr¿ Date : Wed, 05 Nov 2003 10 :03 :28 +0100 Subject : New Version of OpenFEM

This is to announce the new version (1.1) of OpenFEM, an opensource finite element toolbox for Matlab and Scilab. This is the first release of the toolbox in which a Scilab version is included. More detailed information (and downloads) at http ://www.openfem.net

D. Chapelle (INRIA-Rocquencourt)

[PDE Portal] From : Are Magnus Bruaset mailto :amb@nobjects.com Date : Mon, 20 Aug 2001 11 :43 :12 +0200 Subject : Web Portal for PDE-related Sites

Numerical Objects is pleased to announce the first release of its PDE Portal at

http ://www.pdeportal.com

This is a searchable and categorized index to several hundreds of high quality web sites focusing on technologies for partial differential equations. We hope that this complimentary service will become a valuable tool for the PDE community, and we invite you to check it out.

In order to continuously improve the contents of the PDE Portal, we encourage your submissions for non-listed PDE sites that might of general interest to the public. Procedures for submission is available at www.pdeportal.com.

Best regards

Are Magnus Bruaset Numerical Objects AS

[PETSc] From : Barry Smith bsmith@mcs.anl.gov Date : Fri, 13 Apr 2001 15 :36 :34 -0500 (CDT) Subject : Release of PETSc 2.1.0

Portable, Extensible Toolkit for Scientific Computation (PETSc) http ://www.mcs.anl.gov/petsc

We are pleased to announce the release of the PETSc 2.1.0 parallel software libraries for the implicit solution of PDEs and related problems.

As always, please send bug reports, questions, and requests for new features to petsc-maint@mcs.anl.gov.

Thanks for your continued support.

The PETSc developers, Satish, Kris, Bill, Dinesh, Lois, and Barry

[PLTMG] From : Randy Bank ¡reb@sdna1.ucsd.edu¿ Date : Wed, 31 Mar 2004 10 :31 :53 -0800 Subject : PLTMG 9.0 Available

PLTMG 9.0 is a package for solving elliptic partial differential equations in general regions of the plane. It is based on continuous piecewise linear triangular finite elements. PLTMG features several adaptive meshing options and an algebraic multilevel solver for the resulting systems of linear equations. PLTMG provides a suite of continuation options to handle PDEs with parameter dependencies. It also provides options for solving several classes of optimal control and obstacle problems. The package includes an initial mesh generator and several graphics packages. Support for the Bank-Holst parallel adaptive meshing paradigm is also provided.

PLTMG is provided as Fortran (and a little C) source code, in both single and double precision versions. The code has interfaces to X-Windows, MPI, and Michael Holst's OpenGL display tool SG. The X-Windows, MPI, and SG interfaces require libraries that are NOT provided as part of the PLTMG package. PLTMG is available from uetlib, Mgnet, and my homepage : http ://scicomp.ucsd.edu/~reb/

Randy Bank University of California, San Diego

Introduction.

1. Solutions d'EDP, caractéristiques. [Courant & Hilbert]

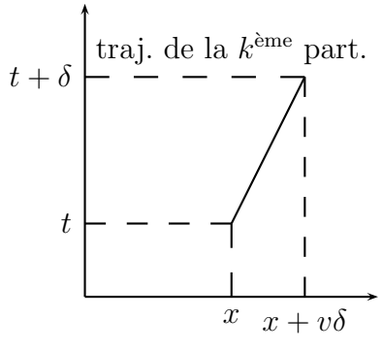
1.1. Premiers exemples. Si l'évolution d'une masse ponctuelle est susceptible d'être décrite par des équations différentielles ordinaires, un milieu étendu demande des équations aux dérivées partielles.

Comme exemple de passage d'un type de description à l'autre, considérons une étendue fluide soumise à des forces diverses. Tout d'abord, on voit le fluide comme une collection de particules dont on étudie le mouvement (Lagrange) : la $k^{\text{ème}}$ particule a une vitesse v_k qui vérifie

$$\rho_k \frac{dv_k}{dt} = f_k,$$

f_k étant la force par unité d'étendue qui s'exerce sur la particule (gravité, pression, autres particules,...). On a donc un grand système d'équations différentielles ordinaires.

Mais si on désire examiner ce qui se passe en un point (x,t),



La valeur de v en un point $(x+h, t+\delta)$ proche de (x,t) est $v + h\dfrac{\partial v}{\partial x} + \delta\dfrac{\partial v}{\partial t}$. Nous ne pouvons apprécier que l'accélération de particules sur leurs propres trajectoires, donc prendre $h = v\delta$, et nous savons que v augmente alors de $\delta\dfrac{dv}{dt}$, avec l'expression de $\dfrac{dv}{dt}$ vue plus haut, d'où

$$\rho\left(v\frac{\partial v}{\partial x} + \frac{\partial v}{\partial t}\right) = f$$

(Euler). A plus d'une variable spatiale, on a

$$\rho\left(\sum_j v_j \frac{\partial v_i}{\partial x_j} + \frac{\partial v_i}{\partial t}\right) = f_i$$

pour chaque composante v_i de la vitesse.

Autre exemple (J.M. Cooper, *Introduction to Partial Differential Equations with MATLAB*, Birkhauser, 1998) : le $k^{\text{ème}}$ automobiliste adapte sa vitesse $\dot{x}_k$ selon une fonction décroissante de la densité (nombre d'autos par km) d'autos sur la route : $\dot{x}_k = f(\rho_k)$. Comment la densité ρ va-t-elle évoluer dans le temps à partir d'une configuration initiale donnée $\rho(x,0)$ sur un segment de route ?

La variation de ρ est $d\rho = \dfrac{\partial\rho}{\partial t}\,dt + \dfrac{\partial\rho}{\partial x}\,dx = \left[\dfrac{\partial\rho}{\partial t} + f(\rho)\dfrac{\partial\rho}{\partial x}\right]dt$ en suivant la $k^{\text{ème}}$ voiture ; d'autre part, $\rho_k = \dfrac{1}{x_{k+1}-x_k}$ est devenu $\dfrac{1}{x_{k+1}+f(\rho_{k+1})dt - x_k - f(\rho_k)dt}$ après un temps dt, ou encore

$\dfrac{1}{\dfrac{1}{\rho_k} + dt\, f'(\rho_k)\,\Delta\rho_k} \sim \rho_k - dt\,\rho_k^2\, f'(\rho_k)\Delta\rho_k$, où $\Delta\rho_k = \rho_{k+1} - \rho_k \sim \rho_k^{-1}\partial\rho/\partial x$ (dérivée spatiale de ρ multipliée par la distance entre deux voitures). Il reste

$$\frac{\partial\rho}{\partial t} + F'(\rho)\frac{\partial\rho}{\partial x} = 0, \tag{1}$$

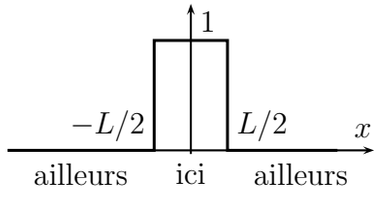
où $F(\rho) = \rho f(\rho)$ est le débit (nombre de voitures par heure) en un point (x,t).

Diffusion des euros.

(Voir `http ://www.math.ucl.ac.be/~magnus/eurodiff.htm`)

Un modèle très simple.

Des acteurs a_1, a_2,..., a_N répartis sur la droite réelle constatent la décroissance avec le temps de la proportion de monnaie indigène (ou autochtone) qu'ils détiennent. On note $y_i(t)$ la proportion détenue par a_i au temps t. Soit t_0 le temps moyen d'échange de toute la monnaie d'un acteur avec ses deux voisins, donc,



$$y_i(t+t_0) = \frac{y_{i-1}(t) + y_{i+1}(t)}{2} \tag{2}$$

avec initialement $y_i(0) = 1$ sur un intervalle de longueur L et $y_i(0) = 0$ ailleurs (il ne sagit que de la proportion de monnaie du pays considéré, la quantité d'argent détenue par les gens ne change pas[3]).

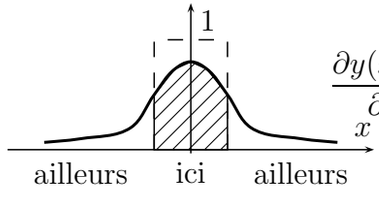
On a aussi

$$y_i(t+t_0) - y_i(t) = \frac{y_{i-1}(t) - 2y_i(t) + y_{i+1}(t)}{2}$$

ou encore

$$\frac{y_i(t+t_0) - y_i(t)}{t_0} = \frac{d_0^2}{2t_0}\frac{y_{i-1}(t) - 2y_i(t) + y_{i+1}(t)}{d_0^2}$$

si d_0 est la distance moyenne entre deux acteurs. Les quotients ainsi mis en évidence sont proches de dérivées :

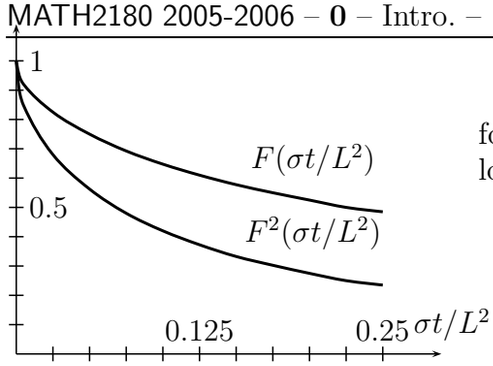


$$\frac{\partial y(x,t)}{\partial t} = \sigma\frac{\partial^2 y(x,t)}{\partial x^2}\,, \qquad \sigma = \frac{d_0^2}{2t_0} \tag{3}$$

Au temps $t > 0$, la distribution uniforme confinée au pays d'origine s'est affaisée et a quelque peu envahi les pays voisins. La proportion encore disponible dans le pays d'origine (c'est la valeur la plus facile à mesurer) est donnée par la partie hachurée

$$Y(t) = \frac{1}{L}\int_{-L/2}^{L/2} y(x,t)\,dx$$

[3]Mais alors, à quoi sert l'argent ? ? ?



Cette moyenne nationale décroît avec t selon une fonction universelle F de $(\sigma/L^2)t$ que nous verrons plus loin, au chapitre 6, page 152.

$(\sigma/L^2)t$	0	0.01	0.02	0.03	0.04	0.05	0.1	0.15	0.2	0.25
$F((\sigma/L^2)t)$	1.000	0.887	0.840	0.805	0.774	0.748	0.647	0.578	0.526	0.486
$F^2((\sigma/L^2)t)$	1.000	0.787	0.706	0.647	0.600	0.559	0.419	0.334	0.277	0.236

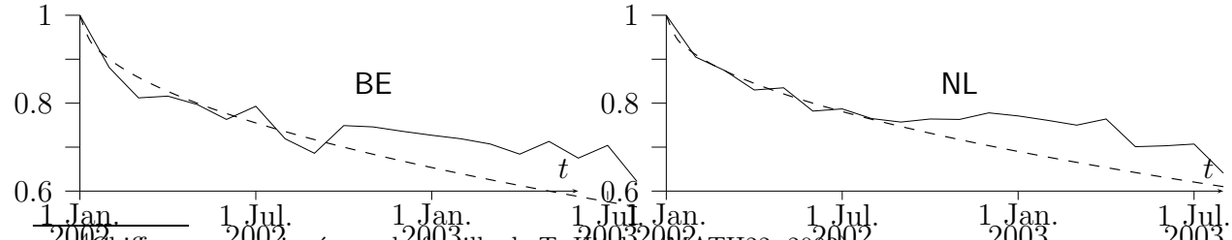
(Au fait, en appliquant directement (2), on obtient $Y = 1 - \frac{3}{2}\frac{5}{4}\cdots\frac{2k-1}{2k-2}\,\frac{d_0}{L}$ après $2k$ pas de temps, si $d_0 << L$ Cela donne $\approx 1 - \sqrt{\frac{4k}{\pi}}\,\frac{d_0}{L} = 1 - \sqrt{\frac{2t}{t_0\pi}}\,\frac{d_0}{L} = 1 - \sqrt{\frac{4\sigma\, t}{L^2\pi}}$ pour $(\sigma/L^2)t$ petit).

Pour un **rectangle**, on obtient $Y_1(t)Y_2(t) = F((\sigma/L_1^2)t)F((\sigma/L_2^2)t)$, avec les longueurs L_1 et L_2 des deux côtés. Avec un rectangle, on peut déjà respecter le rapport superficie/longueur de frontière d'un pays (N.B. : frontière commune avec l'Euroland !). Et tenir compte du nombre moyen de connexions (monétaires) entre les agents et leurs voisins.
Avec $d_0 \approx 1$ km (distance moyenne au supermarché ou à la pompe la plus proche, $L \approx$ une centaine de km, et $t_0 \approx 1$ mois, on s'attend à quelques millièmes pour σ/L^2.

date	BEL.	NED.	LUX
1 fév. 2002	0.881	0.905	0.750
1 mar. 2002	0.812	0.874	0.810
1 avr. 2002	0.816	0.830	0.660
1 mai 2002	0.797	0.835	0.660
1 juin 2002	0.763	0.782	0.510
1 juil. 2002	0.793	0.787	0.570
1 aoû. 2002	0.719	0.765	0.590
1 sep. 2002	0.686	0.757	0.510
1 oct. 2002	0.749	0.764	
1 nov. 2002	0.746	0.763	
1 dec. 2003	0.736	0.778	
1 jan. 2003	0.727	0.771	
1 feb. 2003	0.719	0.761	
1 mar. 2003	0.707	0.750	
1 avr. 2003	0.684	0.764	
1 mai 2003	0.713	0.701	
1 juin 2003	0.675	0.703	
1 juil. 2003	0.704	0.707	
1 août 2003	0.622	0.641	

Une énorme documentation portant sur des mesures effectuées par des centaines de gens aux Pays-Bas et en Belgique (du nord...) est rassemblée à **http ://www.eurodiffusie.nl**)
On y voit, parmi d'autres informations disponibles, la répartition des monnaies en Belgique, aux Pays-Bas et au Luxembourg[4] en fonction du temps.

On en estime très approximativement le σ/L^2 de la Belgique et des Pays-Bas à ≈ 0.0022 et 0.0018 mois^{-1}, et à ≈ 0.01 mois^{-1} pour le Luxembourg.
Suite p. 152



[4] Chiffres communiqués par la famille de T. Hitzky (MATH22, 2002).

1.2. Problème de Cauchy.

Une équation aux dérivées partielles (EDP), ou un système de telles équations, étant une ou plusieurs relations entre une ou des fonctions de plusieurs variables et certaines de ses (ou leurs) dérivées partielles, on peut chercher à se ramener à un problème d'équation différentielle ordinaire.

Pour fixer les idées, soit une équation du premier ordre pour une fonction de deux variables :

$$F(u,p,q,x,y)=0, \qquad p=\frac{\partial u}{\partial x},\ q=\frac{\partial u}{\partial y}. \tag{4}$$

Peut-on déterminer $u(x_1,y_1)$ à partir d'une valeur de départ $u(x_0,y_0)$?

Normalement non : soit $L=\{x=\lambda(s), y=\mu(s), s_0\leqslant s\leqslant s_1\}$ un arc d'extrémités (x_0,y_0) et (x_1,y_1). Sur L, la dérivée $\dfrac{du}{ds}=p\lambda'+q\mu'$ n'est pas déterminée puisque (4) ne donne qu'une relation entre p et q au lieu de donner p et q en fonction de x, y et u.

Cependant, si (4) est linéaire en p et q (équation *quasi-linéaire*)

$$\alpha(x,y,u)p+\beta(x,y,u)q+\gamma(x,y,u)=0, \tag{5}$$

on voit que l'on a une équation différentielle ordinaire

$$\frac{du}{ds}=-\kappa(s)\,\gamma(\lambda(s),\mu(s),u)$$

sur toute courbe $C=\{x=\lambda(s), y=\mu(s)\}$, de tangente alignée sur (α,β) : $\lambda'(s)=\kappa(s)\,\alpha(x(s),y(s),u)$, $\mu'(s)=\kappa(s)\,\beta(x(s),y(s),u)$ sur C (*courbe caractéristique*), où κ est une fonction arbitraire.

On peut donc déterminer $u(x,y)$ à partir de la valeur de u en un point de la courbe caractéristique passant par (x,y). Pour déterminer u dans une région, il faut partir des valeurs de u en différents points situés sur des courbes caractéristiques différentes $\Rightarrow$ donner u sur une courbe C_0 n'ayant nulle part de tangente commune avec une courbe caractéristique (**problème de Cauchy**).

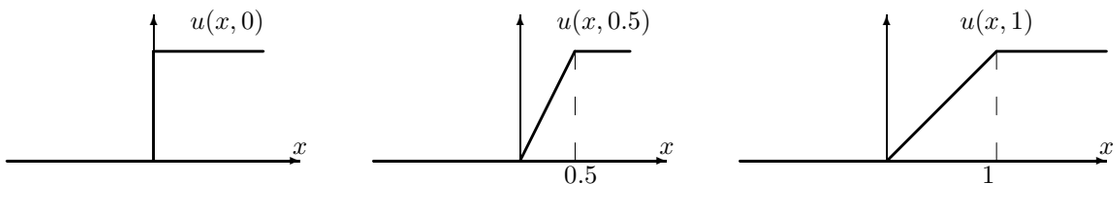
Une courbe caractéristique est donc une courbe ne pouvant servir en aucun point de courbe de départ pour le problème de Cauchy.

On dit aussi que les courbes caractéristiques sont les lieux de propagation des discontinuités (physiques, numériques) des solutions.

Exercices. Résoudre $2\partial u/\partial x+3\partial u/\partial y=0$ sachant que $u(x,0)=0$ sur $x<0$ et $u(x,0)=1$ sur $x>0$ (fonction échelon).

La solution *générale* de $\alpha\partial u/\partial x+\beta\partial u/\partial y=0$ est $u(x,y)=\varphi(\beta x-\alpha y)$, où φ est une fonction *arbitraire* d'une variable.

Résoudre l'équation de Burgers sans viscosité $\partial u/\partial t+u\,\partial u/\partial x=0$ avec $u(x,0)=0$ sur $x\leqslant 0$, $u(x,0)=x/\varepsilon$ sur $[0,\varepsilon]$ et $u(x,0)=1$ sur $x\geqslant\varepsilon$. Remarquez que, comme $\gamma=0$, u est constante sur chaque caractéristique. Que se passe-t-il quand $\varepsilon\to 0$?



On voit donc apparaître très naturellement des fonctions non dérivables (***solutions généralisées***), on reviendra sur la nécessaire extension du cadre fonctionnel au delà de $\mathscr{C}^1$...

Pour l'équation (1), on a $\rho =$ constante sur les droites caractéristiques $x = t\,F'(\rho)+$ constante... tant que ces caractéristiques ne se rencontrent pas !

Par exemple, vitesse de propagation du front d'un bouchon : si $\rho = \rho_{\max}$ en $x = x_0$ lorsque $t = 0$, où $f(\rho_{\max}) = 0$, $\rho = \rho_{\max}$ en $x = x_0 + F'(\rho_{\max})\,t$ au temps t.

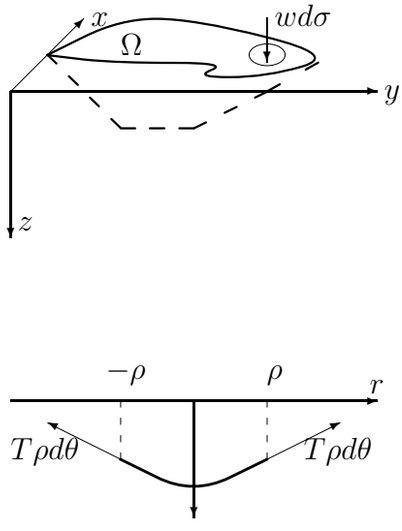
Il existe des méthodes numériques de résolution de (4) ou (5) utilisant les courbes caractéristiques, mais on étudiera plutôt ici des méthodes de détermination de u sur des points d'une grille donnée.

Une équation du premier ordre à n variables : on a toujours des *courbes* caractéristiques, le problème de Cauchy consiste à donner u sur une (hyper)*surface* (variété de dimension $n-1$).

Plusieurs équations du premier ordre à n variables : si $n > 2$, on peut avoir des surfaces caractéristiques, cf. chap. 5.

Pour des EDP d'ordre **deux**, le problème de Cauchy consiste à déterminer u à partir de ses valeurs et des valeurs de sa dérivée normale sur une courbe C_0 (la dérivée tangentielle sur C_0 étant connue si u est connue). Mais les problèmes de physique mathématique se ramènent-ils toujours à un problème de Cauchy ?

1.3. Un exemple exemplaire : équation de la membrane. Voici un exemple de dérivation (pas très propre) d'une EDP du second ordre en physique mathématique, qui ne se résout pas par propagation le long de caractéristiques.



Soit une membrane initialement tendue sur un domaine Ω de $\mathbb{R}^2$, maintenue sur le bord $\partial\Omega$, soumise à une tension T par unité de longueur. On impose une charge de densité w tournée vers le bas. Sous l'effet de cette charge, la membrane se déforme en se courbant, ses points se placent sur une surface d'équation $z = u(x, y)$. On suppose w suffisamment petite pour que u et ses dérivées partielles restent petites (*élasticité infinitésimale*).

Du fait de la courbure de la membrane, les tensions s'exerçant sur un fragment d'aire $d\sigma$ prennent une composante verticale devant compenser exactement la charge $w(x, y)d\sigma$. Exprimons ce fait :

la section formant un angle θ avec l'axe des x d'un petit disque de rayon ρ a pour équation $z = \tilde{u}(r) := u(x + r\cos\theta, y + r\sin\theta)$. Les tensions en $r = \pm\rho$ ont des composantes verticales $T\rho d\theta\tilde{u}'(\pm\rho)$, d'où une résultante verticale $T\rho d\theta(\tilde{u}'(\rho) - \tilde{u}'(-\rho)) \approx 2T\rho^2 d\theta\tilde{u}''(0)$.

Voyons en termes des dérivées partielles de u : comparons les développements de Taylor jusqu'au second ordre

$$\tilde{u}(r) = \tilde{u}(0) + r\tilde{u}'(0) + \frac{1}{2}r^2\tilde{u}''(0) =$$

$$u(x + r\cos\theta, y + r\sin\theta) = u(x,y) + \begin{bmatrix} r\cos\theta & r\sin\theta \end{bmatrix} \begin{bmatrix} \dfrac{\partial u}{\partial x} \\ \dfrac{\partial u}{\partial y} \end{bmatrix} + \frac{1}{2}\begin{bmatrix} r\cos\theta & r\sin\theta \end{bmatrix} \begin{bmatrix} \dfrac{\partial^2 u}{\partial x^2} & \dfrac{\partial^2 u}{\partial x \partial y} \\ \dfrac{\partial^2 u}{\partial x \partial y} & \dfrac{\partial^2 u}{\partial y^2} \end{bmatrix} \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix}$$

d'où la résultante verticale

$$2T\rho^2 d\theta \tilde{u}''(0) = 2T\rho^2 d\theta \left[\cos^2\theta \frac{\partial^2 u}{\partial x^2} + 2\sin\theta\cos\theta \frac{\partial^2 u}{\partial x \partial y} + \sin^2\theta \frac{\partial^2 u}{\partial y^2}\right]$$

et intégrons en θ de 0 à π : on obtient $T\pi\rho^2 \left[\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}\right]$ qui doit compenser $w(x,y)d\sigma = w(x,y)\pi\rho^2$, d'où

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -w(x,y)/T. \tag{6}$$

La physique suggère que l'on peut déterminer u dans Ω à partir des valeurs de u sur la courbe $C_0 = \partial\Omega$, ce n'est donc pas un problème de Cauchy. Que se passe-t-il ?

1.4. EDP d'ordre 2 : caractéristiques.

L'EDP générale d'ordre 2 est

$$F(u,p,q,r,s,t,x,y) = 0, \qquad r = \frac{\partial^2 u}{\partial x^2}, s = \frac{\partial^2 u}{\partial x \partial y}, t = \frac{\partial^2 u}{\partial y^2},$$

mais limitons nous d'emblée à l'équation quasi-linéaire :

$$\alpha(x,y,p,q,u)r + \beta(x,y,p,q,u)s + \gamma(x,y,p,q,u)t + \delta(x,y,p,q,u) = 0. \tag{7}$$

Montrons dans quelles conditions on peut trouver les trois dérivées secondes r, s et t sur $C_0 = \{x = \lambda_0(\tau), y = \mu_0(\tau)\}$ à partir de u et de sa dérivée normale $\partial u/\partial n$:
connaissant u sur C_0, on en déduit sa dérivée tangentielle $\partial u/\partial\tau$, et, comme on donne aussi la dérivée normale de u le long de C_0, nous disposons du gradient de $u = [\partial u/\partial x, \partial u/\partial y] = [p,q]$ sur C_0 :

$$\frac{\partial u}{\partial \tau} = \text{dérivée directionnelle selon } [\lambda_0', \mu_0'] = p\lambda_0' + q\mu_0',$$
$$\frac{\partial u}{\partial n} = \text{dérivée directionnelle selon } [-\mu_0', \lambda_0'] = -p\mu_0' + q\lambda_0',$$

ce qui donne bien $p = \partial u/\partial x$ et $q = \partial u/\partial y$ comme fonctions de τ sur C_0. Dérivons maintenant ces fonctions le long de C_0 :

$$\left.\begin{aligned} \frac{\partial p}{\partial \tau} &= \operatorname{grad} p \cdot [\lambda_0', \mu_0'] = \lambda_0' \partial p/\partial x + \mu_0' \partial p/\partial y = \lambda_0' r + \mu_0' s \\ \frac{\partial q}{\partial \tau} &= \operatorname{grad} q \cdot [\lambda_0', \mu_0'] = \lambda_0' \partial q/\partial x + \mu_0' \partial q/\partial y = \lambda_0' s + \mu_0' t \\ &\text{équation (7) : } -\delta = \alpha r + \beta s + \gamma t, \end{aligned}\right\}$$

Système de 3 équations linéaires à 3 inconnues r, s et t de déterminant

$$D = \alpha {\mu'_0}^2 - \beta \lambda'_0 \mu'_0 + \gamma {\mu'_0}^2.$$

Les *directions caractéristiques* sont les (λ'_0, μ'_0) qui annulent D. On dit que l'EDP (7) est

- **Hyperbolique** s'il y a deux directions caractéristiques réelles distinctes : $\beta^2 - 4\alpha\gamma > 0$,
- **Parabolique** s'il y a deux directions caractéristiques confondues : $\beta^2 - 4\alpha\gamma = 0$,
- **Elliptique** s'il n'y a pas de direction caractéristique réelle : $\beta^2 - 4\alpha\gamma < 0$.

On reviendra sur ces notions et on comparera avec d'autres définitions. Le type de l'EDP commande le mode de traitement numérique, mais les choses ne sont pas toujours simples : ainsi, les équations de mécanique des fluides visqueux sont paraboliques mais "presque" hyperboliques si la viscosité est faible. Autre exemple intéressant, l'équation de propagation sur les lignes électriques

$$\frac{\partial U}{\partial x} = rI + \ell \frac{\partial I}{\partial t}, \frac{\partial I}{\partial x} = gU + \gamma \frac{\partial U}{\partial t} \quad \Rightarrow \frac{\partial^2 u}{\partial x^2} = rg\, u + (\gamma r + \ell g)\frac{\partial u}{\partial t} + \gamma\ell \frac{\partial^2 u}{\partial t^2},$$

$(r, g, \ell, \gamma =$ coefficients de résistance, conductance, inductance, capacité$)$, $u = U$ ou I, devient, si g est négligeable, *l'équation des télégraphistes*

$$\frac{\partial^2 u}{\partial x^2} = \gamma r \frac{\partial u}{\partial t} + \gamma\ell \frac{\partial^2 u}{\partial t^2}, \quad \text{hyperbolique,} \tag{8}$$

si, de plus, ℓ est négligeable, *l'équation de la chaleur*

$$\frac{\partial^2 u}{\partial x^2} = \gamma r \frac{\partial u}{\partial t}, \quad \text{parabolique,} \tag{9}$$

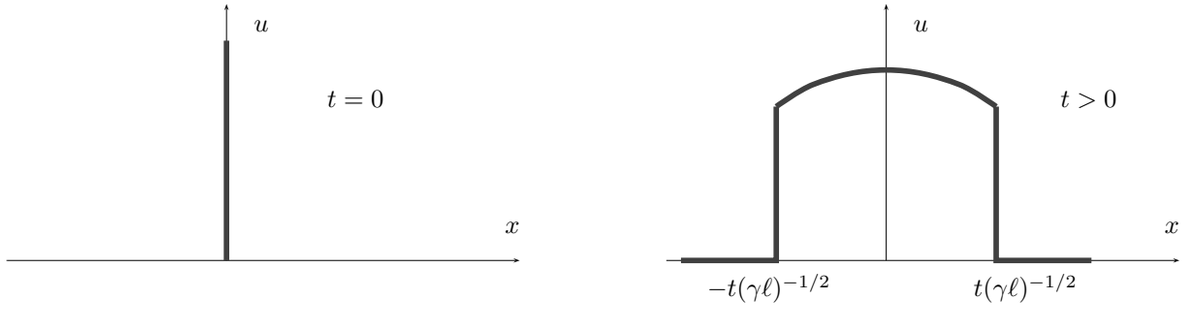
si g et r sont négligeables, *l'équation des ondes* ou des *cordes vibrantes*

$$\frac{\partial^2 u}{\partial x^2} = \gamma\ell \frac{\partial^2 u}{\partial t^2}, \quad \text{hyperbolique.}$$

Exercice. Constater que la solution *générale* de $\partial^2 u/\partial x^2 = c^{-2}\partial^2 u/\partial t^2$ est $u(x,t) = \varphi(x - ct) + \psi(x + ct)$, où φ et ψ sont deux fonctions *arbitraires*.

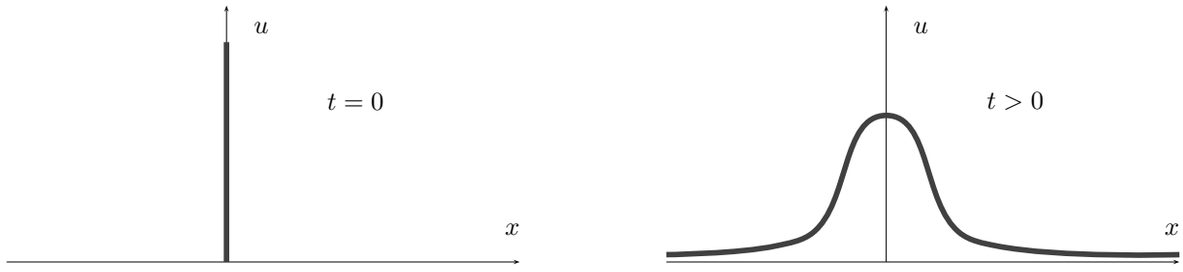
Exemples de solutions du problème de Cauchy. On considère une équation du second ordre à deux variables x et t, et on cherche à déterminer $u(x,t)$ pour $t > 0$, connaissant les fonctions $u(x,0)$ et $(\partial u/\partial t)_{t=0}$.

(1) **Équation hyperbolique**, équation des télégraphistes (8), et partons d'une impulsion concentrée en $x = 0$:



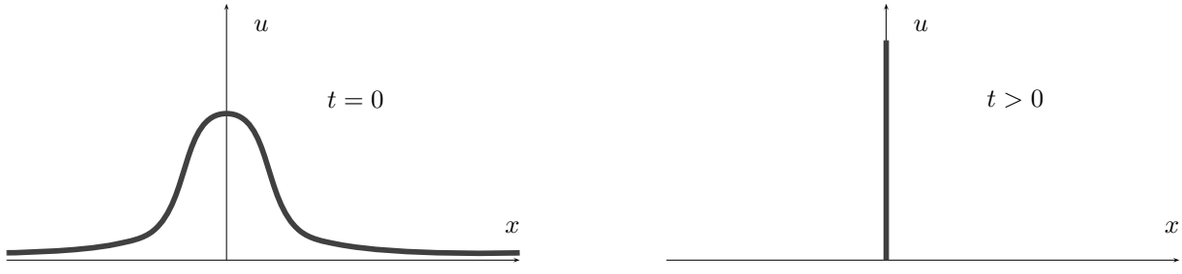
Au temps $t > 0$, le signal s'étend sur le support $[-t(\gamma\ell)^{-1/2}, t(\gamma\ell)^{-1/2}]$: vitesse de propagation finie $(\gamma\ell)^{-1/2}$. La formule du cas particulier examiné ici est $u(x,t) = e^{-rt/(2\ell)} I_0\left(\frac{r}{2\ell}\sqrt{t^2 - \gamma\ell x^2}\right)$ sur $x^2 < \gamma\ell t^2$, où I_0 (fonction de Bessel modifiée) est solution de $I_0''(X) + X^{-1} I_0'(X) = I_0(X)$ [Courant & Hilbert, vol. 2].

(2) **Équation parabolique**, équation de la chaleur (9) :



cette fois, la solution s'étend sur toutes les valeurs réelles de x dès que $t > 0$: vitesse de propagation infinie. De plus, la solution s'amortit et devient de plus en plus lisse à mesure que t augmente. La formule de cet exemple est $\sqrt{\dfrac{\gamma r}{4\pi t}}e^{-\gamma r x^2/(4t)}$.

(3) **Équation elliptique**, équation de Laplace $\partial^2 u/\partial x^2 + \partial^2 u/\partial t^2 = 0$.
Phénomène inverse du précédent : des singularités peuvent apparaître en tout $t > 0$!



Une condition initiale peut être infiniment amplifiée en n'importe quelle valeur > 0 de t : ***Le problème de Cauchy est mal posé***. Exemple : $u(x,t) =$ partie réelle de $\dfrac{1}{(x+it)^2+\varepsilon}$, $\varepsilon > 0$.

Nom	Équation	Applications	Type
Laplace Poisson Helmholtz plaque	$\Delta u = 0$ $-\Delta u = 4\pi\rho$ $(\Delta + k^2)u = 0$ $\Delta^2 u = f$	milieux continus etc.	E
Chaleur, diffusion	$\dfrac{\partial u}{\partial t} = C\Delta u + \Phi$	$u =$ température ou concentration	P
ondes (d'Alembert)	$\Delta u - \dfrac{1}{c^2}\dfrac{\partial^2 u}{\partial t^2} = \Box u = 0$	propagation	H
Maxwell potentiels : $\boldsymbol{E} = -\nabla V - \dfrac{\partial \boldsymbol{A}}{\partial t}$ $\boldsymbol{B} = \nabla \times \boldsymbol{A}$	$\nabla\cdot\boldsymbol{D} = \rho$ $\nabla\cdot\boldsymbol{B} = 0$ $\nabla \times \boldsymbol{E} = -\dfrac{\partial \boldsymbol{B}}{\partial t}$ $\nabla \times \boldsymbol{H} = \boldsymbol{J} + \dfrac{\partial \boldsymbol{D}}{\partial t}$ $\Box V = -\dfrac{\rho}{\varepsilon_0}$ $\Box \boldsymbol{A} = -\mu_0 \boldsymbol{J}$	électro- magnétisme	H
continuité (conservation masse)	$\dfrac{\partial \rho}{\partial t} + \nabla \cdot (\rho \boldsymbol{v}) = 0$	milieux continus	
Advection-diffusion Burgers Stokes Navier-Stokes (incompressible)	$\dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = \nu\dfrac{\partial^2 u}{\partial x^2}$ $\dfrac{\partial u}{\partial t} + u\dfrac{\partial u}{\partial x} = \nu\dfrac{\partial^2 u}{\partial x^2}$ $\nabla\cdot\boldsymbol{v} = 0, \rho\dfrac{\partial \boldsymbol{v}}{\partial t} = \rho\boldsymbol{g} - \nabla p + \eta\Delta\boldsymbol{v}$ $\nabla\cdot\boldsymbol{v} = 0,\ \rho\dfrac{\partial \boldsymbol{v}}{\partial t} + \rho(\boldsymbol{v}\cdot\nabla)\boldsymbol{v} = \rho\boldsymbol{g} - \nabla p + \eta\Delta\boldsymbol{v}$	fluides visqueux	P
Korteweg-De Vries Boussinesq	$\dfrac{\partial u}{\partial t} + au\dfrac{\partial u}{\partial x} + b\dfrac{\partial^3 u}{\partial x^3} = 0$ $\dfrac{\partial^2 u}{\partial t^2} - \dfrac{\partial^2 u}{\partial x^2} - \dfrac{\partial^4 u}{\partial x^4} - 3\dfrac{\partial^2 u^2}{\partial x^2} = 0$	solitons	

Nom	Équation	Applications	Type
Schrödinger	$i\hbar\dfrac{\partial\psi}{\partial t} = -\dfrac{\hbar^2}{2m}\Delta\psi + U\psi$		P
Klein-Gordon	$\left(\Delta - \dfrac{1}{c^2}\dfrac{\partial^2}{\partial t^2} - \dfrac{m_0^2c^2}{\hbar^2}\right)\psi = 0$	quantique	H
Dirac	$\left(\gamma_\lambda\dfrac{\partial}{\partial x_\lambda} + \dfrac{m_0c^2}{\hbar}\right)\boldsymbol{\psi} = 0$		H
sinus-Gordon	$\dfrac{\partial^2\psi}{\partial x\partial t} - \sin\psi = 0$		
Fokker-Planck*	$\partial p/\partial t = -\sum\partial(b_i p)/\partial x_i + (1/2)\sum_{i,j}\partial^2(a_{i,j}p)/\partial x_i\partial x_j$	distributions de particules et mathématiques financières.	P
Black-Scholes	$\partial c/\partial t = (\nu^2/2)S^2\partial^2 c/\partial S^2 + rS\partial c/\partial S - rc$		
Hamilton-Jacobi	$\dfrac{\partial S}{\partial t} + H(\mathbf{grad}_{\boldsymbol{q}}S, \boldsymbol{q}) = 0$	mécanique analytique	

Quelques équations

* G.W. Wei, D.S. Zhang, D.J. Kouri and D.K. Hoffman, Distributed approximating functional approach to the Fokker-Planck equation : Time propagation, *J. Chem. Phys.* , **107**, 3239-3246 (1997). `http ://www.math.msu.edu/~wei/PAPER/p20.pdf`

G.W. Wei, A unified approach for the solution of the Fokker-Planck equation, *J. Phys. A, Mathematical and General*, **33**, 4935-4953 (2000). `http ://www.math.msu.edu/~wei/PAPER/p46.pdf`

Chapitre 1

Équations elliptiques, formulations variationnelles.

1. Équations elliptiques et problèmes aux limites.

1.1. Solution donnée sur une frontière.

Nous venons de voir dans un exercice que les solutions de $c^2\partial^2 u/\partial x^2 - \partial^2 u/\partial t^2 = 0$ se représentent comme fonctions de $x - ct$ et $x + ct$.

Plus généralement, l'équation (7) à coefficients constants, c'est-à-dire

$$\alpha\frac{\partial^2 u}{\partial x^2} + \beta\frac{\partial^2 u}{\partial x \partial y} + \gamma\frac{\partial^2 u}{\partial y^2} + \delta = 0, \tag{10}$$

avec α, β et γ constants, et $\delta = 0$, admet des solutions $\varphi(\mathsf{A}x + \mathsf{B}y) \Rightarrow r = \mathsf{A}^2\varphi'', s = \mathsf{AB}\varphi'', t = \mathsf{B}^2\varphi''$, si le trinôme

$$\alpha\mathsf{A}^2 + \beta\mathsf{AB} + \gamma\mathsf{B}^2 = 0$$

a des racines réelles, ce qui n'est *pas* le cas pour des équations elliptiques. On ne peut donc alors représenter des valeurs de la solution par transport, ou propagation, le long de caractéristiques.

***Cependant**, la constance du signe de $\alpha\mathsf{A}^2 + \beta\mathsf{AB} + \gamma\mathsf{B}^2$ dans le cas des équations elliptiques permet de justifier un autre type de conditions aux limites.*

En effet, il est alors possible de déterminer la solution u de (10) dans un domaine borné Ω de $\mathbb{R}^2$ à partir des valeurs de u sur la courbe fermée $\partial\Omega$ qui sert de frontière à Ω.

Sans prouver ce fait (qui est tout un chapitre des mathématiques, et que nous ne ferons qu'effleurer), montrons quand même l'unicité de la solution : si u_1 et u_2 sont solutions de (10) en prenant les mêmes valeurs sur $\partial\Omega$, $v := u_2 - u_1$ vérifie donc

$$\alpha\frac{\partial^2 v}{\partial x^2} + \beta\frac{\partial^2 v}{\partial x \partial y} + \gamma\frac{\partial^2 v}{\partial y^2} = 0,$$

(α, β, γ et δ étant toujours supposés constants). Multiplions par v et intégrons sur Ω :

$$\int_\Omega v\left[\alpha\frac{\partial^2 v}{\partial x^2} + \beta\frac{\partial^2 v}{\partial x \partial y} + \gamma\frac{\partial^2 v}{\partial y^2}\right] dxdy = 0.$$

Intégrons par parties selon x et/ou y :

$$\begin{aligned}
&\alpha \int_{\partial\Omega} \pm v \frac{\partial v}{\partial x}\, dy - \alpha \int_\Omega \left(\frac{\partial v}{\partial x}\right)^2 dxdy \\
+&\frac{\beta}{2} \int_{\partial\Omega} \pm v \frac{\partial v}{\partial y}\, dy - \frac{\beta}{2} \int_\Omega \frac{\partial v}{\partial x}\frac{\partial v}{\partial y}\, dxdy \\
+&\frac{\beta}{2} \int_{\partial\Omega} \pm v \frac{\partial v}{\partial x}\, dx - \frac{\beta}{2} \int_\Omega \frac{\partial v}{\partial y}\frac{\partial v}{\partial x}\, dxdy \\
+&\gamma \int_{\partial\Omega} \pm v \frac{\partial v}{\partial y}\, dx - \gamma \int_\Omega \left(\frac{\partial v}{\partial y}\right)^2 dxdy = 0.
\end{aligned}$$

Comme $v = 0$ sur $\partial\Omega$, il reste bien un trinôme

$$-\int_\Omega [\alpha \mathsf{A}^2 + \beta \mathsf{A}\mathsf{B} + \gamma \mathsf{B}^2] dxdy = 0,$$

ce qui n'est possible que si $\mathsf{A} = \partial v/\partial x$ et $\mathsf{B} = \partial v/\partial y$ sont nuls dans tout $\Omega \Rightarrow v =$ constante $\Rightarrow v = 0$ puisque $v = 0$ sur la frontière de Ω.

Un exemple de détermination (Poisson). Soit $\Omega =$ le disque $x^2 + y^2 < 1$. On cherche u harmonique (c'est-à-dire $\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = 0$) à partir de ses valeurs $\mathfrak{u}(\varphi)$ sur la frontière. Solution : $u(r,\theta) = \sum_{k=-\infty}^{\infty} c_k r^{|k|} e^{ik\theta}$, où $\mathfrak{u}(\varphi) = \sum_{k=-\infty}^{\infty} c_k e^{ik\varphi}$ est la série de ***Fourier*** de $\mathfrak{u}$.

Comme chaque coefficient de Fourier c_k est une intégrale $c_k = (2\pi)^{-1} \int_0^{2\pi} \mathfrak{u}(\varphi) e^{-ik\varphi}\, d\varphi$, on a la représentation intégrale

$$u(r,\theta) = \frac{1}{2\pi} \int_0^{2\pi} \mathfrak{u}(\varphi) \frac{1 - r^2}{1 - 2r\cos(\theta - \varphi) + r^2}\, d\varphi\,, \qquad r < 1,$$

qui montre comment les valeurs intérieures de u dépendent d'une intégrale impliquant les valeurs sur la frontière.

1.2. Formulations variationnelles : introduction.

En nous efforçant de trouver une solution du problème fonctionnel $F(u) = 0$, nous sommes passés d'une formulation ***ponctuelle***, où F est une application définie sur un espace de fonctions suffisamment continûment dérivables, dite encore formulation ***classique***, à une formulation ***variationnelle***, où on apprécie des relations entre fonctions à travers des intégrales. Ainsi, on cherchera par exemple à établir $\int_\Omega (F(u))^2\, dxdy = 0$, d'où on tirera $F(u) = 0$ presque partout. Il faudra donc aussi s'attendre à rencontrer des espaces de fonctions comme L^1 et L^2 plutôt que $\mathscr{C}^1$, $\mathscr{C}^2$, etc.

Supposons disposer d'une approximation (nous sommes en analyse numérique, tout de même) $\tilde{u}$ de u. Cherchons une correction $\tilde{u} + w$. Prenons le cas où F est linéaire, en fait affine :

$$F(u) = F_0 + F_1 u,$$

et évaluons l'intégrale du carré de $F(\tilde{u} + w)$:

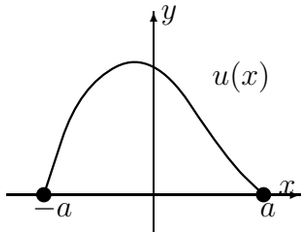
$$\int_\Omega [F(\tilde{u} + w)]^2 dxdy = \int_\Omega [F(\tilde{u})]^2 dxdy + 2\int_\Omega F(\tilde{u})\, v\, dxdy + \int_\Omega v^2\, dxdy,$$

où $v = F_1 w$. On voit que la meilleure fonction v ***minimise***

$$J(v) = 2\int_\Omega F(\tilde{u})v\,dxdy + \int_\Omega v^2\,dxdy.$$

On appelle ***calcul des variations*** l'art de minimiser ou de maximiser une forme, généralement une intégrale, sur un espace de fonctions, avec ou sans contrainte.

Problème le plus célèbre : trouver la courbe fermée de longueur fixée contenant le domaine d'aire maximale (problème de Didon [fondation de Carthage]).



Faisons semblant de ne pas connaître la solution du problème et adoptons la version suivante : parmi les arcs de courbe passant par deux points fixés, d'équation $y = u(x)$, avec $u(-a) = u(a) = 0$, de longueur $L = \int_{-a}^{a} \sqrt{1 + u'^2(x)}\,dx$ fixée, trouver l'arc enfermant une aire $s = \int_{-a}^{a} u(x)\,dx$ maximale $\iff$ minimiser $L = L(u)$ pour s fixé.

Perturbons u par une fonction de faible amplitude que l'on baptise εv, et développons $L(u + \varepsilon v)$ jusqu'au premier ordre en Taylor :

$$L(u+\varepsilon v) = \int_{-a}^{a} \sqrt{1 + (u' + \varepsilon v')^2(x)}\,dx = L(u) + \varepsilon \int_{-a}^{a} \frac{u'(x)v'(x)}{\sqrt{1+u'^2(x)}\,dx} + O(\varepsilon^2).$$

Le coefficient de ε (***variation première***) doit être nul (pour toute fonction v vérifiant $v(-a) = v(a) = 0$ et $\int_{-a}^{a} v(x)\,dx = 0$), sinon L ne serait pas extrémal en u. Après intégration par parties, on trouve $\int_{-a}^{a} \left(\frac{u'(x)}{\sqrt{1+u'^2(x)}}\right)' v(x)\,dx = 0$, ce qui sera réalisé si $[u'/\sqrt{1+u'^2}]'$ est une constante λ.

On finit par trouver, bien sûr, un arc de cercle $u(x) = \text{const.}\ \sqrt{1-\lambda^2 x^2} + \text{const.}$

On est donc passé ici d'une formulation variationnelle à une formulation classique.

Exercice. Si $J(u) = \int_a^b F(u, u', x)\,dx$, la variation première de J est $\int_a^b \left[\frac{\partial F}{\partial u} - \left(\frac{\partial F}{\partial u'}\right)'\right] v\,dx + \left[\frac{\partial F}{\partial u'}\,v\right]_a^b$ (Euler).

Le passage par une formulation variationnelle, sous forme de de minimisation ou de maximisation, donne l'impression que le problème de l'existence est naturellement et spontanément résolu (point de vue de Riemann), comme le penseront (presque) tous les physiciens et ingénieurs. Ce n'est pas (toujours) faux, à condition de recourir aux espaces de fonctions convenables (espaces de Hilbert)...

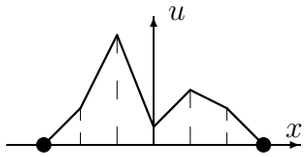
En théorie des surfaces minimales, surfaces d'aire minimum s'appuyant sur une (des) courbe(s) gauche(s) donnée(s) (***problème de Plateau***), le problème d'existence est parfois ***très*** difficile[1].

[1]Cf. R. Courant : *Dirichlet's Principle, Conformal Mapping, and Minimal Surfaces*, Interscience, New York, 1950, E. Giusti : *Minimal Surfaces and Functions of Bounded Variation*, Birkhauser 1984 ; F. Morgan : *Geometric Measure Theory : A beginner's guide*, 2d ed. Academic Press. 1995 ; Dierkes, Hildebrandt, Kuster, Wohlrab : *Minimal Surfaces I and II*, Springer-Verlag. 1992. Remerciements à J. Meinguet et Th. De Pauw.

1.3. Méthode numérique associée à une formulation variationnelle.

En tout cas, si le problème est de minimiser une forme $J(u)$, on imagine immédiatement la méthode numérique de résolution approchée consistant à minimiser $J(u_h)$ sur $u_h \in$ un espace de dimension ***finie***. On aboutira alors à un nombre ***fini*** d'équations pour les coefficients décrivant la solution approchée.

Par exemple, prenons une version simplifiée du problème abordé plus haut : minimiser $J(u) = \int_{-a}^{a} u'^2(x)\,dx$ sous contraintes $u(-a) = u(a) = 0$ et $\int_{-a}^{a} u(x)\,dx = s$.



Prenons d'abord des fonctions linéaires par morceaux. Une telle fonction est entièrement déterminée par ses ordonnées $u_k = u_h(-a+kh)$, $k = 1, \ldots, N-1$ (où $h = (2a/N)$).

Nous devons minimiser $hJ(u_h) = \sum_{k=0}^{N-1} (u_{k+1} - u_k)^2$ (où $u_0 = u_N = 0$) sous la contrainte $h\sum_1^{N-1} u_k = s$, ce qui donne les équations

$$2(u_k - u_{k-1}) - 2(u_{k+1} - u_k) = -2u_{k-1} + 4u_k - 2u_{k+1} = \lambda, \qquad k = 1, \ldots, N-1$$

système tridiagonal aisément soluble. C'était notre exemple le plus simple de résolution approchée par ***éléments finis***.

Un tout autre choix d'approximation est la somme de Fourier $u_h(x) = \sum_{k=1}^{N} c_k \sin \dfrac{k\pi(x+a)}{2a}$ (exemple d'approximation ***pseudo-spectrale***). On trouve $J(u_h) = \sum_1^N k^2\pi^2 c_k^2/(4a)$ et la contrainte $\dfrac{4a}{\pi} \sum\limits_{\substack{k=1 \\ k \text{ impair}}}^{N} \dfrac{c_k}{k} = s$. On trouve $c_k = \text{const.}/k^3$, $k = 1, 3, 5, \ldots$

1.4. Espaces de fonctions continûment dérivables par morceaux.

Abordons plus rigoureusement la suite de la discussion en précisant de premiers espaces de fonctions.

Soit Ω un ouvert (non vide) de $\mathbb{R}^n$, $\mathscr{C}^m(\Omega)$ est l'espace des fonctions (réelles) possédant des dérivées continues au moins jusque l'ordre m ($\mathscr{C}^0(\Omega)$ est l'espace des fonctions continues) dans Ω. Pour des fonctions de plusieurs variables, cela implique[2] que la valeur d'une dérivée partielle ne dépend pas de l'ordre des dérivations. Une dérivée partielle se note

$$D^s f(x) = \frac{\partial^{|s|}}{\partial x_1^{s_1} \cdots \partial x_n^{s_n}} f(x_1, \ldots, x_n),$$

où s est un $n-$uple d'entiers $\geqslant 0$, de somme $|s| = s_1 + \cdots + s_n \leqslant m$.

Donc, $D^r(D^s f) = D^{r+s} f$, tant que $|r+s| \leqslant m$. Une fonction de n variables possède $n^{|s|}$ dérivées partielles d'ordre $|s|$, dont seules $\dfrac{n(n+1)\ldots(n+|s|-1)}{|s|!} = \dfrac{[n]^{|s|}}{|s|!} = \dfrac{(n+|s|-1)!}{(n-1)!|s|!}$

[2] Théorie de Young et Schwarz.

sont susceptibles d'être distinctes ($|s|-$sélections d'un $n-$ensemble[3], ou combinaisons avec répétitions de n objets pris $|s|$ à $|s|$).

On définit aussi la $s^{\text{ème}}$ puissance d'un vecteur $h = (h_1, \ldots, h_n)$ de $\mathbb{R}^n$ comme étant

$$h^s := h_1^{s_1} h_2^{s_2} \ldots h_n^{s_n},$$

et la factorielle de s par $s! := s_1! s_2! \ldots s_n!$.

Il y a $\dfrac{|s|!}{s!}$ dérivées mixtes de même valeur $D^s f$.

Cette notation multi-indicielle[4] permet une grande économie d'écriture : ainsi, le ***développement de Taylor*** d'une fonction de classe $\mathscr{C}^{m+1}(\Omega)$ s'écrit simplement

$$f(a+h) = \sum_{|s| \leqslant m} \frac{D^s f(a)}{s!}\, h^s + \sum_{|s|=m+1} \frac{D^s f(a+\theta h)}{s!}\, h^s \quad 0 \leqslant \theta \leqslant 1, \tag{11}$$

si la boule de centre a et de rayon $|h|$ est incluse dans Ω.

En effet, pour a et h fixés dans $\mathbb{R}^n$, soit g la fonction d'*une* variable $g(t) := f(a+th)$, et considérons le développement de Taylor de g en $t = 1$:

$$f(a+h) = g(1) = \sum_{k=0}^{m} \frac{g^{(k)}(0)}{k!} + \frac{g^{(m+1)}(\theta)}{(m+1)!}.$$

On montre aisément[5] par récurrence sur k que $g^{(k)}(t) = \displaystyle\sum_{|s|=k} \frac{k!}{s!} D^s f(a+th) h^s$. □

Ω étant un ouvert, quel que soit m, une fonction de $\mathscr{C}^m(\Omega)$ n'a normalement aucune chance de pouvoir se prolonger à la frontière de Ω (par exemple, $1/x$ est $\mathscr{C}^\infty((0,1))$). Or, nous aurons besoin de valeurs sur la frontière pour traiter des conditions aux limites.

Soit Ω un domaine (= ouvert connexe) borné de $\mathbb{R}^n$, $\overline{\Omega}$ son adhérence, et $\Gamma = \partial\Omega$ sa frontière.

$\mathscr{C}^m(\overline{\Omega})$ est la partie de $\mathscr{C}^m(\Omega)$ constituée de fonctions de dérivées d'ordre $\leqslant m$ ***bornées*** et ***uniformément continues*** dans Ω. Une telle fonction et ses dérivées jusquà l'ordre m ont alors un et un seul prolongement continu sur $\overline{\Omega}$ [Adams, § 1.25-1.26].

(Par ***exemple***, avec $\Omega = (0,1)$, $\sqrt{x}$ est dans $\mathscr{C}^0([0,1])$, mais pas dans $\mathscr{C}^1([0,1])$;
$\sqrt{x}\,\sin x$ est dans $\mathscr{C}^1([0,1])$;
$\sin(1/x)$ n'est dans *aucune* classe $\mathscr{C}^m([0,1])$).

Pour pouvoir passer de formulations classiques à des formulations variationnelles, on demandera que Γ permette l'application du théorème de la divergence, ou de Gauss-Green :

$$\int_\Omega \frac{\partial f}{\partial x_i} d\boldsymbol{x} = \int_\Gamma f(x)\,(\vec{n} \cdot \vec{e}_i)\, dS\,, \qquad \int_\Omega \operatorname{div} \boldsymbol{F}\, d\boldsymbol{x} = \int_\Gamma \boldsymbol{F}\cdot\vec{n}\, dS, \tag{12}$$

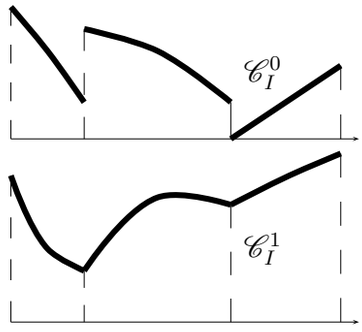
pour $\forall f \in \mathscr{C}^1(\overline{\Omega}), \boldsymbol{F} \in [\mathscr{C}^1(\overline{\Omega})]^n$, où $\vec{n}$ désigne le vecteur normal unitaire dirigé vers *l'extérieur* de la frontière Γ. Il suffit que Γ soit lisse par morceaux, c'est-à-dire constituée

[3] cf. cours FSA/MATH4

[4] introduite par H. Whitney (1934)

[5] En passant de $k-1$ à k, on obtient n contributions au terme contenant $D^s f$, à partir des n suites $s - (0,0,\ldots,0,1,0,\ldots,0)$ de poids $k-1$, donnant chacune lieu à un facteur $(k-1)!/[s_1! \ldots (s_p-1)! \ldots s_n!]$ pour $p = 1, 2, \ldots, n$, dont la somme donne bien $k!/|s|!$

d'un nombre fini de parties continûment différentiables (cf. par exemple, T.M. Apostol, *Calculus*, vol.II, chap. 11 et 12).



Définition. Fonctions dérivables par morceaux.
f est de classe $\mathscr{C}_I^0(\overline{\Omega})$ si on peut trouver un nombre fini d'ouverts disjoints, réguliers au sens (12), $\Omega_1, \cdots, \Omega_p$ tels que

– $\overline{\Omega} = \bigcup_{i=1}^p \overline{\Omega_i}$,
– la restriction de f à Ω_i admet une extension continue à $\mathbb{R}^n$, $i = 1, 2, \ldots, p$.

f est de classe $\mathscr{C}_I^m(\overline{\Omega})$, avec $m \geqslant 1$ si

– $f \in \mathscr{C}^{m-1}(\overline{\Omega})$;
– on peut trouver un nombre fini d'ouverts disjoints, réguliers au sens (12), $\Omega_1, \cdots, \Omega_p$ tels que
  – $\overline{\Omega} = \bigcup_{i=1}^p \overline{\Omega_i}$,
  – la restriction de f à Ω_i est de classe $\mathscr{C}^m(\overline{\Omega_i})$, $i = 1, 2, \ldots, p$.

Si $f \in \mathscr{C}_I^m(\overline{\Omega})$, $D^a f \in \mathscr{C}_I^{m-|a|}(\overline{\Omega})$, si $|a| \leqslant m$.
En effet, 1. dans chaque Ω_i, $D^b(D^a f) = D^{a+b} f$ est continue si $|a| + |b| \leqslant m$, donc si $|b| \leqslant m - |a|$, et se prolonge continûment jusqu'à la frontière de Ω_i. 2. Si $m > |a|$, $D^b(D^a f) = D^{a+b} f$ est continue dans tout Ω si $|a| + |b| \leqslant m - 1 \Rightarrow |b| \leqslant m - |a| - 1$.

On a $\ldots \subset \mathscr{C}^{m+1}(\overline{\Omega}) \subset \mathscr{C}_I^{m+1}(\overline{\Omega}) \subset \mathscr{C}^m(\overline{\Omega}) \subset \mathscr{C}_I^m(\overline{\Omega}) \ldots$.

Enfin, les ***intégrales sur*** Ω de fonctions définies par morceaux s'entendent comme la somme des intégrales sur les morceaux Ω_i, $i = 1, 2, \ldots, p$.

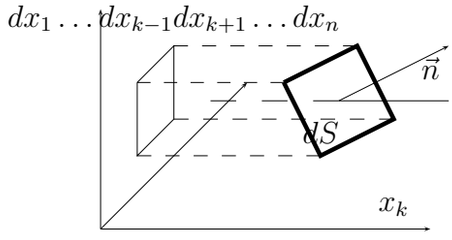
Intégration par parties. *Si $u \in \mathscr{C}_I^{|\alpha|}(\overline{\Omega})$, avec $\alpha_k > 0$, et $v \in \mathscr{C}_I^{|\beta|+1}(\overline{\Omega})$,*

$$\int_\Omega D^\alpha u(\boldsymbol{x})\ D^\beta v(\boldsymbol{x}) d\boldsymbol{x} = \int_\Gamma D^{\alpha-\vec{e}_k} u(\boldsymbol{x})\ D^\beta v(\boldsymbol{x})\,(\vec{n} \cdot \vec{e}_k)\, dS\ - \int_\Omega D^{\alpha-\vec{e}_k} u(\boldsymbol{x})\ D^{\beta+\vec{e}_k} v(\boldsymbol{x}) d\boldsymbol{x}. \tag{13}$$

On effectue une intégration par parties sur la seule variable x_k dans chaque Ω_i, où $D^\alpha u$ et $D^\beta v$ sont continues :

$$\int_\Omega D^\alpha u(\boldsymbol{x})\ D^\beta v(\boldsymbol{x}) d\boldsymbol{x} = \sum_{i=1}^p \int_{\Omega_i} \frac{\partial}{\partial x_k} D^{\alpha-\vec{e}_k} u(\boldsymbol{x})\ D^\beta v(\boldsymbol{x}) dx_1 \ldots dx_{k-1} dx_{k+1} \ldots dx_n\ dx_k$$

$$= \sum_{i=1}^p \left\{ \int D^{\alpha-\vec{e}_k} u\ D^\beta v\big|_{\partial\Omega_i}\ dx_1 \ldots dx_{k-1} dx_{k+1} \ldots dx_n - \int_{\Omega_i} D^{\alpha-\vec{e}_k} u(\boldsymbol{x})\ D^{\beta+\vec{e}_k} v(\boldsymbol{x}) d\boldsymbol{x} \right\},$$



où $\Big|_{\partial\Omega_i}$ désigne une somme signée ($-$ pour les points d'entrée ; $+$ pour les points de sortie) de contributions sur la frontière de Ω_i, pour chaque $(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)$ fixé. En sommant sur les Ω_i, Les contributions des frontières intérieures à Ω se réduisent par ***continuité*** de $D^{\alpha-\vec{e}_k}u$ et de $D^\beta v$ dans tout le domaine Ω.

1.5. Traitement complet d'un problème à une dimension.

On partira d'un problème différentiel aux limites très simple, et on en déduira quelques formulations variationnelles. On verra que ces formulations définissent la solution, tout aussi bien que le problème initial.

Le problème est de déterminer u telle que $-u'' = f$ donnée dans $\mathscr{C}^0_I([a,b])$, avec les conditions aux limites $u(a) = u(b) = 0$. On s'attend bien sûr à ce que $u \in \mathscr{C}^2_I([a,b])$. Appelons U_0 l'espace des fonctions continues vérifiant $u(a) = u(b) = 0$.

1.5.1. *Formulation classique.* Donc,

$$?u \in \mathscr{C}^2_I([a,b]) \bigcap U_0 : -u''(x) = f(x), x \in \Omega_i, i = 1, \dots, p \tag{14}$$

Solution : f n'a qu'un nombre fini de points de discontinuité, est donc intégrable, et

$$u'(x) = u'(a) - \int_a^x f(y)\, dy$$

(notons que $u'(a)$ est encore une inconnue), et

$$u(x) = u'(a)\,(x-a) - \int_a^x (x-y) f(y)\, dy.$$

u' est bien continue, avec une dérivée discontinue aux points de discontinuité de f, donc $u' \in \mathscr{C}^1_I((a,b))$, et on a bien $u \in \mathscr{C}^2_I((0,1))$.

Enfin, $u(b) = 0 \Rightarrow 0 = u'(a)(b-a) - \int_a^b (b-y) f(y)\, dy$, ce qui donne $u'(a)$.

Remarque : on condense le résultat dans une formule très élégante

$$u(x) = \int_a^b G(x,y) f(y)\, dy, \text{ avec } G(x,y) = \frac{(x-a)(b-y)}{b-a} \text{ si } x \leqslant y; G(x,y) = \frac{(y-a)(b-x)}{b-a} \text{ si } y \leqslant x. \tag{15}$$

(Cas très particulier d'une ***formule de Green***), ou encore

$$G(x,y) = \min\left(\frac{(x-a)(b-y)}{b-a}\,,\, \frac{(y-a)(b-x)}{b-a} \right).$$

1.5.2. *Formulation variationnelle forte.*

$$?u \in \mathscr{C}^2_I([a,b]) \bigcap U_0 : -\int_a^b u''(x) v(x)\, dx = \int_a^b f(x) v(x)\, dx, \qquad \forall v \in \mathscr{C}^0_I([a,b]). \tag{16}$$

1.5.3. *Formulation variationnelle semi-faible.*

$$?u \in \mathscr{C}^1_I([a,b]) \bigcap U_0 : \int_a^b u'(x) v'(x)\, dx = \int_a^b f(x) v(x)\, dx, \qquad \forall v \in \mathscr{C}^1_I([a,b]) \bigcap U_0. \tag{17}$$

1.5.4. *Formulation variationnelle faible, ou distributionnelle.*

$$?u \in \mathscr{C}^0_I([a,b]) : -\int_a^b u(x)v''(x)\,dx = \int_a^b f(x)v(x)\,dx, \qquad \forall v \in \mathscr{C}^2_I([a,b]) \bigcap U_0. \tag{18}$$

(Pourquoi "distributionnelle" ? Parce que l'on dit que $f = D^s g$ au sens des distributions dans Ω si $\displaystyle\int_\Omega f\varphi\,d\boldsymbol{x} = (-1)^{|s|}\int_\Omega gD^s\varphi\,d\boldsymbol{x}$ pour toute fonction-test $\varphi \in \mathscr{D}'(\Omega)$.)

Ces quatre problèmes ont tous une et une seule solution, la même dans les quatre cas !

(1) Constatons que la solution de la formulation classique est bien une solution de chaque formulation variationnelle :

   (a) $-u'' = f \Rightarrow -\displaystyle\int_a^b u''v = \int_a^b fv$, bien entendu,

   (b) ensuite, $-\displaystyle\int_a^b u''v = -[u'v]_a^b + \int_a^b u'v' = \int_a^b u'v'$, si $v \in U_0$,

   (c) enfin, $\displaystyle\int_a^b u'v' = [uv']_a^b - \int_a^b uv'' = \int_a^b uv''$, car u est la solution de la formulation classique, donc $u \in U_0$.

(2) Chaque formulation variationnelle n'a qu'une solution : si u_1 et u_2 sont solutions, on aurait, pour $w = u_1 - u_2$,

   (a) Dans (16), $\displaystyle\int_a^b w''v = 0$ pour tout $v \in \mathscr{C}^0_I([a,b])$, on peut donc choisir $v = w''$, d'où w''^2 d'intégrale nulle $\Rightarrow w'' = 0$ (par morceaux) $\Rightarrow w'$ = constante (la même constante partout, puisque $w' \in \mathscr{C}^1_I([a,b]) \subset \mathscr{C}^0([a,b])) \Rightarrow w(x) = \alpha x + \beta \Rightarrow w(x) \equiv 0$ puisque $w \in U_0$.

   (b) Dans (17), $\displaystyle\int_a^b w'v' = 0$ pour tout $v \in \mathscr{C}^1_I([a,b]) \cap U_0$, on peut donc choisir $v = w$, d'où w'^2 d'intégrale nulle $\Rightarrow w' = 0$ (par morceaux) $\Rightarrow w =$ constante (la même constante partout), $\Rightarrow w(x) \equiv 0$ puisque $w \in U_0$.

   (c) Dans (18), $\displaystyle\int_a^b wv''0$ pour tout $v \in \mathscr{C}^2_I([a,b]) \cap U_0$. Construisons $v \in \mathscr{C}^2_I([a,b]) \cap U_0$ tel que $v'' = w$ (en s'aidant de la formulation classique), on a donc w^2 d'intégrale nulle et w continue $\Rightarrow w = 0$.

1.6. EDP elliptiques et formulations variationnelles en dimension > 1. .

La simplicité trompeuse de l'exemple précédent se dissipe quand on passe à un problème à plus d'une variable : il n'y a plus de solution "évidente" de la formulation classique !

Prenons un problème aux limites, par exemple une équation de Poisson $-\Delta u = f$ dans un domaine Ω de $\mathbb{R}^n$, et des conditions aux limites, soit $u = 0$ sur la frontière $\Gamma = \partial\Omega$ de Ω. On demandera $u \in \mathscr{C}^2(\Omega)$, en fait $u \in \mathscr{C}^2_I(\overline{\Omega})$, c'est-à-dire $u \in \mathscr{C}^2$ par morceaux et $u \in \mathscr{C}^1(\overline{\Omega})$, et, bien entendu, $u|_\Gamma = 0$: appelons ça $u \in U$.

Alors, on passe de $-\int_\Omega v\Delta u\,d\boldsymbol{x} = -\int_\Omega v\ \text{div}\ \ \text{grad}\ u\,d\boldsymbol{x} = \int_\Omega vf\,d\boldsymbol{x}$ à la ***forme variationnelle semi faible***

$$\int_\Omega \text{grad}\ u\cdot\ \text{grad}\ v\,d\boldsymbol{x} = \int_\Omega vf\,d\boldsymbol{x}\ , \qquad \forall v\in X, \tag{19}$$

en appliquant deux fois (13), d'abord avec $\alpha=(2,0)$ et $\beta=(0,0)$, puis avec $\alpha=(0,2)$, et en sommant (théorème de la divergence, ou de Gauss-Green)

$$\int_\Omega G\,\text{div}\ \vec{F}\,d\boldsymbol{x} = \int_\Omega \text{div}\ (G\,\vec{F})\,d\boldsymbol{x} - \int_\Omega \vec{F}\cdot\ \text{grad}\ G\,d\boldsymbol{x} = \int_\Gamma G(\vec{F}\cdot\vec{n})\,dS - \int_\Omega \vec{F}\cdot\ \text{grad}\ G\,d\boldsymbol{x}. \tag{20}$$

où $\vec{n}$ désigne le vecteur normal unitaire dirigé vers *l'extérieur* de la frontière Γ (voir plus haut, p. 31).

Enfin, si le vecteur $\vec{F}$ est lui-même un gradient,

$$\int_\Omega v\Delta u\,d\boldsymbol{x} = \int_\Gamma v\frac{\partial u}{\partial n}\,dS\ - \int_\Omega \text{grad}\ u\cdot\ \text{grad}\ v\,d\boldsymbol{x},$$

puisque le laplacien Δ (souvent noté aussi ∇^2) est la divergence du gradient.

Donc,

(1) $u\in\mathscr{C}^2_I(\overline{\Omega})$, $u|_\Gamma=0$: $-\Delta u=f\in\mathscr{C}^0_I(\overline{\Omega})$ implique bien

(2) $u\in\mathscr{C}^1_I(\overline{\Omega})$, $u|_\Gamma=0$: $\int_\Omega \text{grad}\ u\cdot\ \text{grad}\ v\,d\boldsymbol{x} = \int_\Omega f\ v\,d\boldsymbol{x}$ pour tout $v\in\mathscr{C}^1_I(\overline{\Omega})$, $v|_\Gamma=0$,

mais on ne peut remonter aisément de (2) à (1). Remarquons cependant que (2) a au plus une solution : avec $v=u_2-u_1$, on aurait $\int_\Omega |\text{grad}(u_2-u_1)|^2\,d\boldsymbol{x}=0$.

Montrer convenablement l'équivalence entre une formulation différentielle et une formulation variationnelle n'est pas une chose simple. A plusieurs dimensions, il n'est généralement plus possible d'établir l'existence d'une solution de la formulation classique[6].

Cependant, le traitement des problèmes physiques peut aboutir naturellement à une formulation variationnelle [7], qui sera dès lors à considérer comme formulation de départ.

Ainsi, pour décrire la membrane chargée, appliquons le *principe des travaux virtuels* : modifions la fonction de déplacement vers le bas u en $u+\varepsilon v$, où ε est un (petit) nombre réel (positif ou négatif), et où v est une fonction $\in\mathscr{C}^1_I(\Omega)$. nulle sur la frontière $\Gamma=\partial\Omega$.

Le principe des travaux virtuels dit que l'énergie J du système est stationnaire en la solution u, c'est-à-dire que $J(u+\varepsilon v)-J(u)=o(\varepsilon)$. En fait, on va même montrer que J est *mimimum* en u : le système

(1) perd l'énergie potentielle due au travail des forces w se déplaçant vers le bas, soit une contribution $-\int_\Omega w(x,y)u(x,u)\,dxdy$,

[6] sauf pour des géométries simples et/ou des équations particulières : transformation conforme pour l'équation de Laplace à deux dimensions, etc.

[7] Cf. L. Bouckaert, *Théorie de l'élasticité*, Louvain. Chap.10 : les principes variationnels de l'élasticité ; R. Glowinski, J.L. Lions, R. Trémolières : *Analyse numérique des inéquations variationnelles*, 2 vol., Dunod, 1976

(2) acquiert l'énergie de déformation due au travail des forces de tension lors de l'étirement de la membrane, travail proportionnel à l'accroissement des éléments d'aire, l'élément d'aire dS devenant $dS/\cos Z$, où $\cos Z$ est le troisième cosinus directeur de la normale à la surface d'équation $-u(x,y)+z=0$, d'où la contribution

$$
\begin{aligned}
&T\int_\Omega\left(\frac{1}{\cos Z}-1\right)dS\\
&=T\int_\Omega\left(\sqrt{1+\|\text{grad }u\|^2}-1\right)dS\\
&\qquad\text{les cosinus directeurs de la normale à } -u+z=0 \text{ sont les composantes de}\\
&\qquad\qquad\text{grad }(-u+z)/\|\text{grad }-u+z\|)\\
&\approx\frac{T}{2}\int_\Omega\|\text{grad }u\|^2\,dxdy\qquad\text{élasticité infinitésimale).}
\end{aligned}
$$

Donc, avec $w=f$ et $T=1$,

$$J(u)=\frac{1}{2}\int_\Omega\|\text{grad }u\|^2\,dxdy-\int_\Omega uf\,dxdy.$$

On voit bien alors que

$$
\begin{aligned}
J(u+\varepsilon v)&=\frac{1}{2}\int_\Omega\text{grad }(u+\varepsilon v)\cdot\text{grad }(u+\varepsilon v)\,dxdy-\int_\Omega(u+\varepsilon v)f\,dxdy\\
&=\frac{1}{2}\int_\Omega\left[\text{grad }u\cdot\text{grad }u+2\varepsilon\text{grad }u\cdot\text{grad }v+\varepsilon^2\text{grad }v\cdot\text{grad }v\right]\,dxdy-\int_\Omega(u+\varepsilon v)f\,dxdy\\
&=J(u)+\varepsilon\int_\Omega\left[\text{grad }u\cdot\text{grad }v-fv\right]\,dxdy+\varepsilon^2\int_\Omega\|\text{grad }v\|^2\,dxdy
\end{aligned}
$$

est minimum en u si et seulement si le coefficient de ε

$$\int_\Omega\left[\text{grad }u\cdot\text{grad }v-fv\right]\,dxdy=0,\qquad\forall v\in X.$$

Les principes physiques ont donc abouti spontanément à la formulation mathématique variationnelle, sans passer par l'EDP du second ordre.

1.6.1. *Où se trouve la solution d'une EDP elliptique ?*

La solution d'un problème elliptique "raisonnable" sur $\Omega\subset\mathbb{R}^n$, $n>1$, sera-t-elle encore dans $\mathscr{C}^2_I$, ou seulement dans $\mathscr{C}^1_I$?

Considérons $-\Delta u=f=r^\mu$, $\mu\geqslant 0$ dans le secteur $\Omega: 0<r<R, 0<\theta<\alpha\pi$, avec $u=0$ sur $\Gamma=\partial\Omega$.

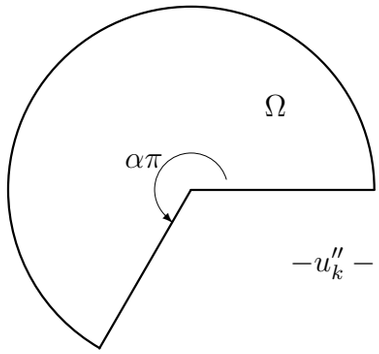
Sous réserve de justification ultérieure, on va essayer un développement

$$u=\sum_{k=1}^{\infty}u_k(r)\sin(k\theta/\alpha),$$

qui sera effectivement nul en $\theta=0$ et en $\theta=\alpha\pi$ si la série converge ponctuellement. En $r=R$, on demandera donc $u_k(R)=0$, $k=1,2,\ldots$

En fait, toute fonction de carré intégrable sur $(0, \alpha\pi)$ admet un développement de Fourier en $\sin(k\theta/\alpha)$, ainsi,

$$f = r^\mu = \frac{4}{\pi} \sum_{k\text{impair}} \frac{r^\mu}{k} \sin(k\theta/\alpha).$$



En coordonnées polaires, le laplacien à deux dimensions est

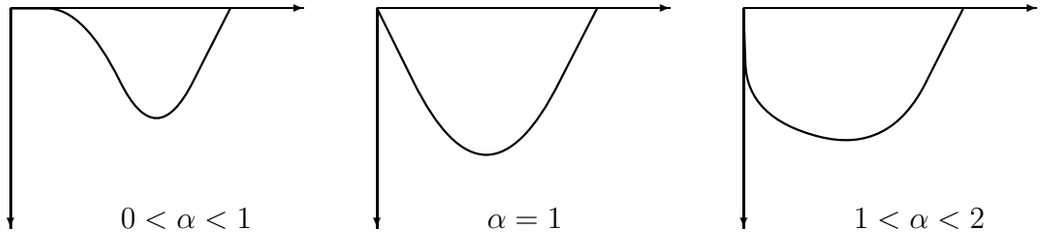
$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2},$$

d'où

$$-u_k'' - \frac{u_k'}{r} + \frac{k^2 u_k}{\alpha^2 r^2} = \frac{4r^\mu}{k\pi}, \qquad k = 1, 3, 5, \ldots$$

ce qui donne $u_k(r) = \dfrac{4\alpha^2 r^{\mu+2}}{k\pi[k^2 - (\mu+2)^2\alpha^2]}$ + une combinaison de $r^{k/\alpha}$ et $r^{-k/\alpha}$, et la seule condition limite $u_k(R) = 0$. Bien entendu, on rejettera les exposants négatifs, afin de garder u continue. Alors, si $\alpha > 1$ (α est entre 0 et 2), u contient des puissances $r^{1/\alpha} \notin \mathscr{C}^1(\overline{\Omega})\ldots$

Le profil de u le long de la bissectrice de Ω a cet aspect :



2. Formes coercives, problème de minimisation, méthode de Ritz-Galerkin.

Les formulations variationnelles les plus élégantes rencontrées jusqu'ici, faisant intervenir des intégrales comme $\displaystyle\int_a^b u'(x)v'(x)\,dx$, ou $\displaystyle\int_\Omega \mathbf{grad}\ u(\boldsymbol{x}) \cdot \mathbf{grad}\ v(\boldsymbol{x})\,d\boldsymbol{x}$, manifestent des traits communs que nous rassemblons ici.

2.1. Formulation variationnelle dans un espace vectoriel.

Soit U un espace vectoriel réel, a une forme

(1) ***bilinéaire*** : $a(\alpha u + \beta v, w) = \alpha a(u, w) + \beta a(v, w)$, $a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w)$, $\forall u, v \in U, \alpha, \beta \in \mathbb{R}$,

(2) ***symétrique*** : $a(u, v) = a(v, u)$, $\forall u, v \in U$.

Définition. *La forme bilinéaire symétrique a est* ***définie positive*** *sur U si $\forall u \in U, a(u,u) \geqslant 0$* ***et*** *$a(u,u) = 0 \Rightarrow u = 0$.*

Proposition. Si a est symétrique définie positive sur U, le problème

$$?u : \qquad a(u,v) = \varphi(v), \quad \forall v \in U, \tag{21}$$

où φ est une forme définie sur U, est équivalent au problème de minimisation

$$\min_{u \in U} J(u) = \frac{1}{2}\, a(u,u) - \varphi(u); \tag{22}$$

la solution est unique *si elle existe.*

En effet, (21) $\Rightarrow J(u+\varepsilon v) = \frac{1}{2}a(u+\varepsilon v, u+\varepsilon v) - \varphi(u+\varepsilon v) =$
$\frac{1}{2}a(u,u) + \varepsilon a(u,v) + \frac{\varepsilon^2}{2}a(v,v) - \varphi(u) - \varepsilon\varphi(v) > J(u)$ dès que $\varepsilon v \neq 0$.

D'autre part, si (22) $\Rightarrow J(u+\varepsilon v) \geqslant J(u)$ pour tout $\varepsilon v \in U$, donc, $\forall v \in U$, $\varepsilon[a(u,v) - \varphi(v)] + \varepsilon^2 a(v,v)/2 \geqslant 0$ pour tout ε réel $\Rightarrow$ (21). □

Par exemple, $a(u,v) = \displaystyle\int_\Omega \mathbf{grad}\ u \cdot\ \mathbf{grad}\ v\, d\boldsymbol{x}$ est bien définie positive sur $\mathscr{C}^1_I(\overline{\Omega}) \cap U_0$ où U_0 est l'espace des fonctions s'annulant en un point c de $\overline{\Omega}$: $a(u,u) = 0 \Rightarrow\ \mathbf{grad}\ u = 0$ dans chaque $\Omega_i \Rightarrow u =$ constante dans chaque Ω_i, donc la même constante dans tout $\overline{\Omega}$ puisque u est continue dans $\overline{\Omega}$, enfin, $u(c) = 0 \Rightarrow u(\boldsymbol{x}) \equiv 0$.

La définition positive de a ne suffit cependant pas toujours à assurer l'existence de la solution de (21) ou (22).

Ainsi, prenons l'exemple plus haut avec $\Omega =$ le disque unité de $\mathbb{R}^2$ et $U_0 = \{u : u(0) = 0\}$. Prenons enfin $\varphi(u) = \displaystyle\int_\Omega u(\boldsymbol{x})\, d\boldsymbol{x}$.

Alors, avec $u(\boldsymbol{x}) = K_\alpha \|\boldsymbol{x}\|^\alpha \in \mathscr{C}^1_I \bigcap U_0$ si $\alpha > 0$, (on prend la norme euclidienne usuelle de $\boldsymbol{x}$),

$$J(u) = \frac{1}{2}\int_\Omega \|\mathbf{grad}\ u\|^2 d\boldsymbol{x} - \int_\Omega u(\boldsymbol{x})\, d\boldsymbol{x} = \frac{K_\alpha^2}{2}\int_0^1\int_0^{2\pi} \alpha^2 r^{2\alpha-2}\, r dr d\theta - K_\alpha \int_0^1\int_0^{2\pi} r^\alpha\, r dr d\theta = \frac{\alpha K_\alpha^2 \pi}{2} - \frac{2\pi K_\alpha}{\alpha+2}$$

que l'on peut rendre aussi négatif que l'on veut en prenant α assez petit, et $K_\alpha = 1/\alpha$.

Par contre, à une dimension, le problème $\min (1/2)\int_{-1}^1 (u'(x))^2 dx - \int_{-1}^1 u(x)\, dx$ sur $u \in \mathscr{C}^1_I([-1,1])$ avec $u(0) = 0$ a une solution parfaitement bien définie qui est $u(x) = -x^2/2 + x$ sign x. Remarquons que la solution n'est pas de classe $\mathscr{C}^2$, ni même $\mathscr{C}^1$! Ceci confirme que les formulations variationnelles sont plus efficaces que les formulations classiques (qui serait ici : $u'' = -1$ avec $u(0) = 0$, $u'(-1) = u'(1) = 0$).

Comment s'assurer que $J(u)$ ne puisse pas tendre vers $-\infty$?

Pour toute direction v, donc, $\|v\| = 1$, pratiquons un "line search" à partir de l'origine dans cette direction :

$$J(\lambda v) = \frac{1}{2}a(\lambda v, \lambda v) - \varphi(\lambda v) = \frac{\lambda^2}{2}a(v,v) - \lambda\varphi(v)$$

prend une valeur minimale de $-\dfrac{(\varphi(v))^2}{2a(v,v)}$ en $\lambda = \dfrac{\varphi(v)}{a(v,v)}$. Si φ est continue, $\varphi(v)$ reste bornée sur la sphère unité $\|v\| = 1$, et il suffit de s'assurer que $a(v,v)$ ne devienne pas arbitrairement petit sur cette sphère.

Définition. *La forme bilinéaire symétrique a est* ***coercive*** *sur U si $\forall u \in U, a(u,u) \geqslant c\|u\|^2$ avec $c > 0$.* On dit aussi alors que a est $U-$***elliptique*** [Ciarlet]. Un problème sera dit **elliptique** s'il est associé à une forme coercive dans un espace approprié (espace de Hilbert, voir plus loin).

La coercivité de a est bien équivalente (a étant bilinéaire) à l'existence d'un infimum strictement positif de $a(v,v)$ sur la sphère unité $\|v\| = 1$.

Proposition. *Si a est coercive et φ bornée sur U, il suffit de rechercher le minimum de J dans $\|u\| \leqslant \|\varphi\|/c$. De plus, J est partout au moins égal à $-\|\varphi\|^2/(2c)$.*

En effet, le "line-search" ci-dessus donne $u = \lambda v,$ avec $\lambda = \dfrac{\varphi(v)}{a(v,v)}$. Donc, comme $\|v\| = 1$ et $a(v,v) \geqslant c > 0$, $\|u\| \leqslant \|\varphi\|/c$.

Ensuite, $J(u) = -\dfrac{(\varphi(v))^2}{2a(v,v)} \geqslant -\dfrac{\|\varphi\|^2}{2c}$. □

Sur un espace U de dimension finie, le caractère défini positif de a implique la coercivité puisque, la sphère étant ***compacte***, la fonction continue $a(v,v)$ atteint son minimum en un point v_0, et $c := a(v_0,v_0) = \min\limits_{\|v\|=1} a(v,v)$ doit être > 0 si a est définie positive. Mais si U est de dimension infinie, la sphère unité n'est plus un compact, et l'infimum d'une forme définie positive peut être nul !

C'est bien ce qui se passe avec l'exemple du disque dans $\mathbb{R}^2$ vu plus haut : $U = \mathscr{C}^1_I(\overline{\Omega}), u(0) = 0$ muni de la norme de L^2 (U n'est pas complet, mais ça ne fait rien ici), $v = Kr^\alpha, \alpha > 0$, α près de 0. Alors, $\|v\|^2_{L^2} = \displaystyle\int_\Omega K^2 r^{2\alpha} dxdy =$

$2\pi K^2 \displaystyle\int_0^1 r^{2\alpha+1} dr = 2\pi K^2/(2\alpha+2)$ vaut 1 si K est proche de $1/\sqrt{\pi}$.

$$a(v,v) = \int_\Omega |\mathbf{grad}\, v|^2 dxdy = 2\pi K^2 \int_0^1 \alpha^2 r^{2\alpha-1} dr = \pi K^2 \alpha \to 0$$

quand $\alpha \to 0$.

Par contre $\displaystyle\int_{-1}^1 u'v'\,dx$ est bien coercive sur à peu peu près tout espace de fonctions vérifiant $u(0) = 0 : \forall x : u(x) = \displaystyle\int_0^x u'(t)\,dt$, d'où, par Hölder ou Cauchy-Schwarz :

$$(u(x))^2 \leqslant |x| \int_0^x (u'(t))^2 |dt| \leqslant \int_{-1}^1 (u'(t))^2 dt.$$

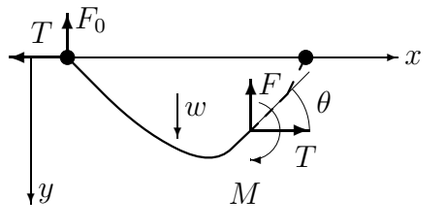
Plus difficile : soit $\Omega =$ le disque unité de $\mathbb{R}^2$. La forme $a(u,v) = \displaystyle\int_\Omega \Delta u \Delta v\; dxdy$, où Δ est le laplacien, est coercive sur l'ensemble des fonctions radiales dérivables s'annulant à l'origine (problème de la plaque circulaire maintenue en son centre). En effet, le laplacien à deux dimensions est $\dfrac{d^2}{dr^2} + \dfrac{d}{rdr} = \dfrac{1}{r}\dfrac{d}{dr}\left(r\dfrac{d}{dr}\right)$, d'où

$u(x) = \displaystyle\int_0^{|x|} u'dr = \int_0^{|x|} ru'\; r^{-1}\, dr = -\int_0^{|x|} (ru')' \log(r/|x|)\;\, dr = -\int_0^{|x|} \Delta u \log(r/|x|)\;\, rdr = -\frac{1}{2\pi}\int_\Omega \Delta u \log\frac{r}{|x|}\; dxdy$

$u^2(x) \leqslant$ const. $a(u,u)$.

Encore deux exemples :

1. Poutre en flexion.



Un segment de poutre posée sur deux appuis est soumis à la réaction $(-T, F_0)$ au point de gauche, à une force (T, F) au point courant $(x(s), y(s)$ [s est la longueur de la portion considérée], à des forces verticales de densité $w(s)$, et enfin à un moment fléchissant $M(s)$.

Équilibre des forces :

$$F + F_0 = \int_0^s w(\ell)\, d\ell, \qquad \text{ou } \frac{dF}{ds} = w(s),$$

Équilibre des moments :

$$\int_0^s x(\ell)w(\ell)\, d\ell - xF - yT + M = 0, \qquad \text{ou } M' = x'F + y'T = F\cos\theta + T\sin\theta.$$

Enfin, $M = EI$ fois la courbure signée $= EI\theta'$ (Euler-Bernoulli), d'où[8]

$$[EI\theta']' = F\cos\theta + T\sin\theta,$$

$$\left\{\frac{1}{\cos\theta}[EI\theta']'\right\}' = w + (\operatorname{tg}\theta)'T, \tag{23}$$

avec x et y fixés aux deux appuis, et $M = 0$, donc $\theta' = 0$ en ces deux points. Poutre encastrée : x, y et θ fixés à une extrémité, $M = 0$ à l'extrémité libre.

Pour de petits déplacements (***élasticité infinitésimale***, ou ***linéaire***, ou ***linéarisée***), soit $\|w\| = \varepsilon$, et on s'intéresse à $u(x) = \lim y(s)/\varepsilon$ quand $\varepsilon \to 0$. On a $s \to x$, $\varepsilon^{-1} \operatorname{tg}\theta = y'/(\varepsilon x') \to u'$, et (23) devient

$$(EIu'')'' - Tu'' = w \tag{24}$$

(où w désigne maintenant la limite de w/ε).

Formulation variationnelle semi-faible de (24) si $u = 0$ en $x = a$ et $x = b$ (conditions ***essentielles***), et $u'' = 0$ en $x = a$ et $x = b$ (conditions ***naturelles***) :

$$\int_a^b EIu''v''\, dx + T\int_a^b u'v'\, dx = \int_a^b wv\, dx, \tag{25}$$

pour tout $v \in \mathscr{C}^2_I([a,b])$ et $v(a) = v(b) = 0$ (seules les conditions essentielles sont à considérer).

Pas de problème de coercivité pour la poutre en ***traction*** $T \geqslant 0$, puisque $u(x) = -\displaystyle\int_a^b G(x,t)u''(t)\, dt$ (vu en (15), p. 33), donc

$$(u(x))^2 \leqslant \int_a^b G^2(x,t)dt \quad \int_a^b (u''(t))^2 dt.$$

[8] Cf. L. Landau & E. Lifchitz, *Théorie de l'élasticité*, Mir, Moscou, 1967, pp. 105-123. C. Truesdell, The influence of elasticity on analysis : the classic heritage, *Bull. (New Ser.) Amer. Math. Soc.* **9** (1983) 293-310 [référence communiquée par J. Meinguet].

Et pour une poutre en ***compression*** $T < 0$? Evaluons le minimum de $a(u,u) = EI\int_a^b (u'')^2\,dx$
$+T\int_a^b (u')^2\,dx$ sous condition $\|u\|^2 = \int_a^b u^2\,dx = 1$, et u et u'' nuls en a et b. Lagrange
$\Rightarrow EI\int_a^b u''v''\,dx + T\int_a^b u'v'\,dx - \lambda\int_a^b uv\,dx = 0, \forall v \in U$, ou encore

$$EIu^{\mathrm{iv}} - Tu'' = \lambda u, \qquad u \in U.$$

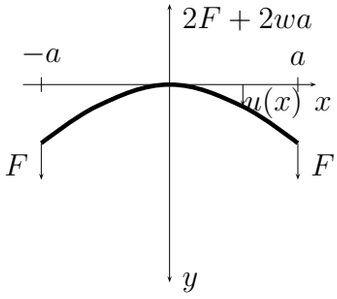
Solution : $u = \cos\left(\frac{\pi}{2}\,\frac{2x-a-b}{b-a}\right)$, donne λ (le minimum cherché) $= \left(\frac{\pi}{b-a}\right)^2\left[EI\left(\frac{\pi}{b-a}\right)^2 + T\right]$.
La forme a cesse d'être coercive, et même définie positive, dès que $T \leqslant -EI(\pi/(b-a))^2$, on est alors dans un cas de ***flambage*** de la poutre.

Même si la charge w est nulle, l'équation non linéaire (23) a alors au moins une solution autre que la solution nulle

$$x = a + \sqrt{\frac{EI}{-2T}}\int_{\theta_0}^{\theta}\frac{\cos\varphi\,d\varphi}{\sqrt{\cos\varphi-\cos\theta_0}}\,; y = \sqrt{\frac{EI}{-2T}}\int_{\theta_0}^{\theta}\frac{\sin\varphi\,d\varphi}{\sqrt{\cos\varphi-\cos\theta_0}} = 2\sqrt{\frac{EI}{-2T}}\,\sqrt{\cos\theta-\cos\theta_0},$$

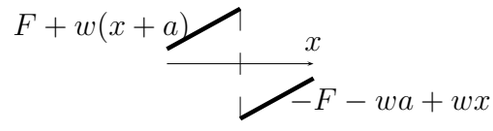
où $\theta_0 < 0$ est la pente (inconnue) de départ. Comme il faut une longueur $b-a$ quand y reprend une valeur nulle, donc en $\theta = -\theta_0$, on a l'équation pour θ_0 $b-a = \sqrt{\frac{EI}{-2T}}\,\mathcal{F}(\theta_0)$, où $\mathcal{F}(\theta_0) := \int_{\theta_0}^{-\theta_0}\frac{d\varphi}{\sqrt{\cos\varphi-\cos\theta_0}}$.
Cette équation est effectivement soluble si $T \leqslant -EI(\pi/(b-a))^2$, car $\mathcal{F}(\theta_0)$ est toujours supérieure à $\sqrt{2}\,\pi$ (à vérifier !, voir Landau & Lifchitz, *op. cit.*).

Exercice. Un exemple de résolution de problème d'élasticité finie (i.e., non infinitésimale) : corde ($E = 0$) uniformément chargée ($w =$ constante) : (23) $\Rightarrow (\ \mathrm{tg}\,\theta)' = -w/T \Rightarrow\ \mathrm{tg}\,\theta =\ \mathrm{tg}\,\theta_0 - sw/T \Rightarrow x' = \cos\theta = [1+(\ \mathrm{tg}\,\theta_0 - sw/T)^2]^{-1/2}, y' = \sin\theta = (\ \mathrm{tg}\,\theta_0 - sw/T)[1+(\ \mathrm{tg}\,\theta_0 - sw/T)^2]^{-1/2} \Rightarrow -\ \mathrm{tg}\,\theta_0 + sw/T = \sinh(w(x-x_0)/T)$ et $y =$ constante $-(T/w)\cosh(w(x-x_0)/T)$: équation d'une ***chaînette***.



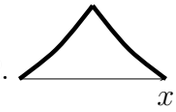
Un exemple plus simple : fléau de balance (en élasticité infinitésimale) soumise à une densité de poids w uniforme et à un poids F à chaque extrémité. La réaction au seul point d'appui $u(0) = 0$ est donc $2F + 2wa$.

$EIu'''(x) =$ résultante des forces entre $-a$ et x :

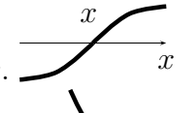


$$wx - (F + wa)\ \mathrm{sign}\ x.$$

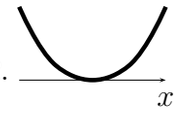
$EIu''(x) = wx^2/2 - (Fw + a)x\ \mathrm{sign}\ x + Fa + wa^2/2.$



$EIu'(x) = wx^3/6 - (Fw + a)x^2\ \mathrm{sign}\ x/2 + (Fa + wa^2/2)x.$



$EIu(x) = wx^4/24 - (Fw + a)x^3\ \mathrm{sign}\ x/6 + (Fa + wa^2/2)x^2/2.$



$J(u) = \dfrac{EI}{2}\displaystyle\int_{-a}^{a}(u''(x))^2dx - w\int_{-a}^{a}u(x)dx - Fu(-a) - Fu(a)$ à minimiser sur les fonctions de $\mathscr{C}^2_I(-a,a)$ nulles en $x=0$.

Solution (à vérifier) $EI\ J(u) = -Fwa^4/4 - w^2a^5/20 - F^2a^3/3$.

2. Plaque mince.

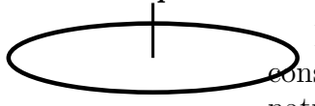
Sans autre commentaire, équation variationnelle : $a(u,v) = \varphi(v)$,

$$\begin{aligned} a(u,v) &= \int_\Omega \left\{ \Delta u\Delta v + (1-\sigma)\left[2\frac{\partial^2 u}{\partial x\partial y}\frac{\partial^2 v}{\partial x\partial y} - \frac{\partial^2 u}{\partial x^2}\frac{\partial^2 v}{\partial y^2} - \frac{\partial^2 u}{\partial y^2}\frac{\partial^2 v}{\partial x^2}\right]\right\} dxdy \\ &= \int_\Omega \left\{ \sigma\Delta u\Delta v + (1-\sigma)\left[2\frac{\partial^2 u}{\partial x\partial y}\frac{\partial^2 v}{\partial x\partial y} + \frac{\partial^2 u}{\partial x^2}\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\frac{\partial^2 v}{\partial y^2}\right]\right\} dxdy \qquad (26) \\ \varphi(v) &= \int_\Omega fv\,dxdy, \end{aligned}$$

sous conditions $u = v = 0$ et $\partial u/\partial n = \partial v/\partial n = 0$ sur $\Gamma = \partial\Omega$ (plaque encastrée).

Correspond à $\Delta^2 u = \Delta(\Delta u) = f$ [Ciarlet, pp. 29-34].

Exemple de solution $\notin \mathscr{C}^2_I(\Omega)$.



Plaque circulaire de densité uniforme suspendue en son centre. $\Delta\Delta u =$ constante $= f$ dans $x^2+y^2 < R^2$, avec $u(0,0) = 0$ et d'autres conditions naturelles (voir plus loin).

On suppose que u ne dépend que de $r = \sqrt{x^2+y^2}$:

$\left(\dfrac{d^2}{dr^2} + \dfrac{1}{r}\dfrac{d}{dr}\right)\left(\dfrac{d^2}{dr^2} + \dfrac{1}{r}\dfrac{d}{dr}\right)u(r) = f$, ou $r^{-1}[r[r^{-1}(ru')']']' = f$. $r[r^{-1}(ru')']' = r^2f/2 + A$. $r^{-1}(ru')' = r^2f/4 + A\ln r + B$. Donc, le laplacien de u contient un terme logarithmique si $A \neq 0$: $u \notin \mathscr{C}^2_I$. $ru' = r^4f/16 + A(r^2\ln r/2 - r^2/4) + Br^2/2 + C$. $u = r^4f/64 + A(r^2\ln r/4 - r^2/4) + Br^2/4 + C\ln r + D$. u continue et $u(0) = 0 : C = D = 0$.

Par formulation variationnelle :

$\displaystyle\int_0^R r^{-1}(ru')'r^{-1}(rv')'\,rdrd\theta = \int_0^R fv\,rdrd\theta$, $\forall v$ avec $v(0) = 0$.

$\lim\limits_{\varepsilon\to 0}[r^{-1}(ru')'rv']_\varepsilon^R - \displaystyle\int_0^R (r^{-1}(ru')')'rv'\,dr = \int_0^R fvr\,dr$.

$\lim\limits_{\varepsilon\to 0}[r^{-1}(ru')'rv']_\varepsilon^R - [(r^{-1}(ru')')'rv]_\varepsilon^R + \displaystyle\int_0^R (r(r^{-1}(ru')')')'v\,dr = \int_0^R fvr\,dr$. D'où les conditions naturelles $(ru')' = (r^{-1}(ru')')' = 0$ en $r = R$, $(ru')' \to 0$ et $r(r^{-1}(ru')')'$ borné quand $r \to 0$. On obtient $A = -R^2f/2$ et $B = -R^2f/4 + R^2f\ln R/2$.

2.2. Méthode de Ritz-Galerkin.

On considère maintenant un ***sous-espace de dimension finie*** $U_h \subset U$, et

$$?u_h \in U_h, \forall v \in U_h : a(u_h, v) = \varphi(v) \qquad (27)$$

où $U_h \subset U$ est un espace de dimension finie N.

En pratique, on considère une base $\{e_1, e_2, \ldots, e_N\}$ de U_h, on doit donc déterminer N scalaires (réels) $c_1, \ldots, c_N$, tels que

$$u_h = \sum_1^N c_j\,e_j,$$

vérifie (27), ce qui signifie

$$\sum_{j=1}^{N} c_j\, a(e_j, e_i) = \varphi(e_i), \qquad i = 1, 2, \ldots, N. \tag{28}$$

La matrice $\boldsymbol{A}_h := [a(e_j, e_i)]_{i,j=1}^{N}$ de (28) est appelée ***matrice de rigidité***.

On aura besoin de cet important rappel :

Définition. *Une matrice carrée réelle* $\boldsymbol{A}$ *est* **définie positive** *si, pour tout vecteur non nul* $\boldsymbol{x}$*, la forme quadratique* $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$.

Toute matrice symétrique réelle définie positive admet une factorisation triangulaire $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$, avec $\boldsymbol{L}$ et $\boldsymbol{U}$ non singulières et $\boldsymbol{U}$ = transposée $\boldsymbol{L}^T$ de $\boldsymbol{L}$ (factorisation de ***Cholesky***).

Les valeurs propres de $\boldsymbol{A}$ symétrique réelle sont toutes strictement positives $\iff$ $\boldsymbol{A}$ est définie positive.

On a le

Théorème. *Si* a *est bilinéaire symétrique définie positive,*

(1) *La matrice de* (28) *est symétrique définie positive.*

(2) *Le problème* (27) *a une et une seule solution dans* U_h.

(3) *La solution* u_h *de* (27) *minimise* $J(u) = a(u,u)/2 - \varphi(u)$ *dans* U_h.

(4) Si (21) *a une solution* u *dans* U, u_h *est l'élément de* U_h *le plus proche de* u*, au sens de la distance* $\sqrt{a(.,.)}$*, c'est-à-dire,* $a(u-u_h, u-u_h) \leqslant a(u-v, u-v)$, $\forall v \in U_h$ *; on a aussi* $a(u_h, u_h) \leqslant a(u,u)$.

En effet, la symétrie de la matrice se déduit immédiatement de la symétrie de a. Pour la définition positive, constatons que, $\forall \boldsymbol{\xi} \in \mathbb{R}^N, \boldsymbol{\xi} \neq 0$, $\boldsymbol{\xi}^T\boldsymbol{A}_h\boldsymbol{\xi} = \displaystyle\sum_{i=1}^{N} \xi_i \sum_{j=1}^{N} a(e_j, e_i)\xi_j =$

$$\sum_{i=1}^{N} \xi_i a\left(\sum_{j=1}^{N} \xi_j e_j, e_i\right) = a\left(\sum_{i=1}^{N} \xi_i e_i, \sum_{j=1}^{N} \xi_j e_j\right) > 0.$$

Une matrice symétrique définie positive est nécessairement non singulière : $\boldsymbol{A}_h\boldsymbol{\xi}$ ne peut être le vecteur nul quand $\boldsymbol{\xi} \neq 0$, sinon on aurait $\boldsymbol{\xi}^T\boldsymbol{A}_h\boldsymbol{\xi} = 0$.

Le problème (27) a donc nécessairement une solution unique u_h dans U_h. Remarquons que (on prend $v = u_h$),

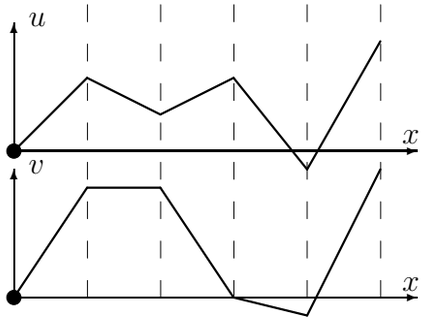
$$a(u_h, u_h) = \varphi(u_h) \;\Rightarrow J(u_h) = -\frac{a(u_h, u_h)}{2}. \tag{29}$$

L'équivalence entre (27) et la minimisation de $J(u_h)$ s'établit comme l'équivalence (21)-(22) établie p. 38.

Enfin, si (21) a une solution (forcément unique) dans U, on examine $a(u-v, u-v)$: soit $w = u_h - v \in U_h$,

$$\begin{aligned} a(u-v, u-v) = a(u-u_h+w, u-u_h+w) &= a(u-u_h, u-u_h) + 2a(u-u_h, w) + a(w,w) \\ &= a(u-u_h, u-u_h) + a(w,w) > a(u-u_h, u-u_h) \text{ si } v \neq u_h. \end{aligned}$$

Et, comme $J(u)$ minimise J sur un espace plus grand que U_h, $J(u) \leqslant J(u_h)$, donc, $a(u_h, u_h) \leqslant a(u,u)$, par (29). □



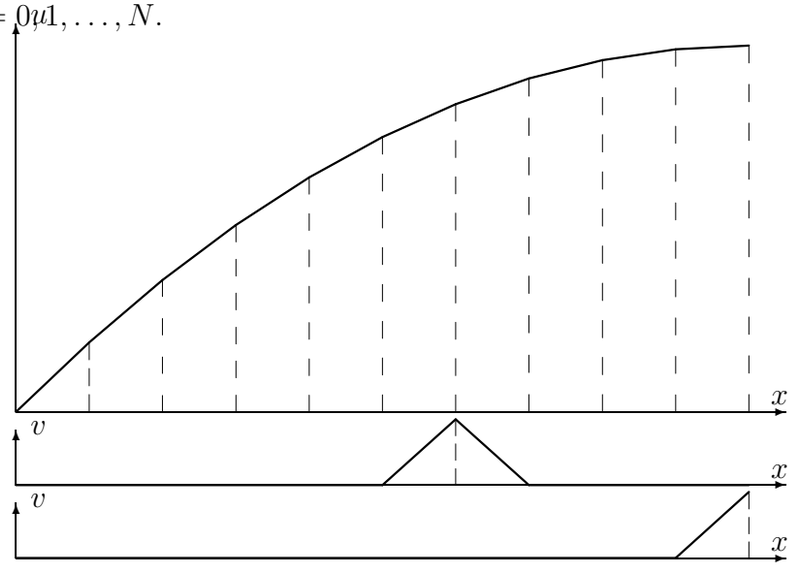
Par **exemple**, prenons $a(u,v) = \int_0^1 u'(x)v'(x)\,dx$, $\varphi(v) = \int_0^1 v(x)\,dx$, $h = 1/N$, $U_h = \{$fonctions continues linéaires par morceaux sur les intervalles $[jh, (j+1)h]$, $j = 0, \ldots, N-1$, avec $u(0) = 0\}$. Donc, (27) devient

$$a(u,v) = \sum_{j=0}^{N-1} \int_{jh}^{(j+1)h} u'(x)v'(x)\,dx =$$

$$= h^{-1} \sum_{j=0}^{N-1} (u_{j+1} - u_j)(v_{j+1} - v_j) = \varphi(v) = \frac{h}{2} \sum_{j=0}^{N-1} (v_j + v_{j+1}),$$

pour tout $(v_1, v_2, \ldots, v_N) \in \mathbb{R}^N$ ($u_0 = v_0 = 0$).

Avec la base canonique de $\mathbb{R}^N$, on obtient $h^{-2}(-u_{i-1} + 2u_i - u_{i+1}) = 1, i = 0, 1, \ldots, N-1$ (d'après les contributions de $v_1, \ldots, v_{N-1}$), et $u_N - u_{N-1} = h^2/2$ (d'après la contribution de v_N). On constate (tous les cas ne sont pas aussi simplement solubles) $u_i = u_h(ih) = ih - (ih)^2/2$, $i = 0, 1, \ldots, N$.



On voit que la condition *naturelle* $u'(1) = 0$ est de plus en plus près d'être réalisée quand $h \to 0$.

$$h^{-2} \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1/2 \end{bmatrix}$$

3. Formes bilinéaires et opérateurs dans des espaces de Hilbert.

On aborde maintenant un cadre assurant la solution de formulations variationnelles.

3.1. Espaces de Banach et de Hilbert.

D'après J.P. Bertrandias, *Analyse fonctionnelle*, Armand Colin, 1970.

(1) **Espaces vectoriels normés.**
E et F vectoriels normés, T application linéaire $E \to F$. Alors, T est continue $\Leftrightarrow T$ est bornée : $\|Tx\| \leqslant C\|x\|$. Alors $\|T\| = \sup_{x\neq 0}(\|Tx\|/\|x\|)$.
Une forme φ est continue sur E si et seulement si son noyau $\{x \in E : \varphi(x) = 0\}$ est fermé.
Une suite $\{x_n\}$ de E converge vers $x \in E$ si $\|x - x_n\| \to 0$ quand $n \to \infty$ (convergence forte) ;
Une suite $\{x_n\}$ de E converge ***faiblement*** vers $x \in E$ si, pour toute forme ℓ du dual E' de E, $\ell(x - x_n) \to 0$ quand $n \to \infty$;
Une suite $\{x_n\}$ de E est une ***suite de Cauchy*** de E si $\|x_n - x_m\|$ est aussi petit que l'on veut dès que n et m sont assez grands.

Une suite $\{T_n\}$ d'applications linéaires de E dans F converge vers T si $\forall x \in E$, $T_n x$ converge (fortement) vers Tx quand $n \to \infty$ (convergence forte des opérateurs) ;
Une suite $\{T_n\}$ d'applications linéaires de E dans F converge ***faiblement*** vers T si $\forall \ell \in F', \forall x \in E$, $\ell(T_n x)$ converge vers $\ell(Tx)$ quand $n \to \infty$ (convergence faible des opérateurs).

(2) **Espaces de Banach.**
Def. : un Banach est un normé *complet*, c'est-à-dire tel que toute suite de Cauchy de E a une limite dans E.
Un sous espace vectoriel *fermé* d'un Banach est encore un Banach.

(a) *Théorème de la borne uniforme.* Si la famille $\{T_\lambda\}_{\lambda\in\Lambda}$ d'applications linéaires continues du Banach E dans le normé F vérifie

$$\forall x \in E : \quad \|T_\lambda x\|_F \leqslant M(x), \quad \forall \lambda \in \Lambda,$$

avec $M(x)$ indépendant de λ, alors $\exists M : \|T_\lambda\| \leqslant M$.

(b) *Théorème de l'inverse continu.* Si T est linéaire continue bijective du Banach E sur le Banach F, T^{-1} est continue $F \to E$.

(c) *Théorème de l'application ouverte.* Si T est linéaire continue surjective $E \to F$, (E et F Banach), l'image par T de tout ouvert de E est un ouvert de F.

(d) *Théorème de Banach-Steinhaus.* Une suite $\{T_n\}_1^\infty$ d'applications linéaires continues d'un Banach E dans un Banach F converge (fortement) vers une application linéaire continue T si et seulement si

(i) $\|T_n\| \leqslant M < \infty\,, \qquad n = 1, 2, \ldots$

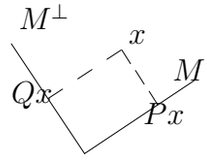
(ii) la suite $\{T_n x\}$ est Cauchy pour tout x dans une partie dense de E.

(3) **Espaces de Hilbert.**
Def. : Un espace *préhilbertien* est un espace vectoriel muni d'un produit scalaire

$(.,.)$ et normé[9] par $\|x\| = (x,x)^{1/2}$.
Inégalité de Cauchy-Schwarz : $(x,y) \leqslant \|x\|\,\|y\|$.
Identité du parallélogramme : $\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2)$.

Un *espace de Hilbert* est un préhilbertien complet.

(a) *Théorème de la projection orthogonale.* Soit M un sous espace vectoriel *fermé* de l'espace de Hilbert H. Alors, $\forall x \in H$ se décompose d'une seule manière en $x = Px + Qx$, avec $Px \in M$ et $Qx \perp M$. Px est l'élément de M le plus proche de x. P est le projecteur orthogonal sur M, $Q = I - P$ est le projecteur orthogonal sur $M^\perp$, complément orthogonal de M. H est donc la somme directe de M et $M^\perp$.



L'existence de Px en tant qu'élément de M le plus proche de x dépend essentiellement du caractère ***complet*** de M : soit $\{y_i\}, y_i \in M$ une suite minimisante de $\|x-y\|$, donc, $\forall \varepsilon > 0$, une infinité de $y_i \in M$ vérifient $\|x - y_i\| \leqslant \mu + \varepsilon$, où $\mu = \inf_{y\in M} \|x-y\|$.
Parallélogramme : $\|2x - y_i - y_j\|^2 + \|y_i - y_j\|^2 = 2(\|x-y_i\|^2 + \|x-y_j\|^2)$, donc,
$\|y_i - y_j\|^2 = 2(\|x-y_i\|^2 + \|x-y_j\|^2) - 4\|x - (y_i+y_j)/2\|^2 \leqslant 4[(\mu+\varepsilon)^2 - \mu^2] = 4\varepsilon(2\mu+\varepsilon)$.
$\{y_i\}$ est Cauchy, donc, comme M est complet, converge.

(b) Caractérisation des formes linéaires continues (*représentant de Riesz*). A toute forme (=application linéaire sur $\mathbb{R}$) φ continue de l'espace de Hilbert H, on peut faire correspondre exactement un élément de $\xi_\varphi \in H$ tel que $\varphi(x) = (x, \xi_\varphi)$, $\forall x \in H$.
On a $\|\varphi\| = \|\xi_\varphi\|$.
On peut donc identifier un espace de Hilbert et son dual. Un Hilbert est évidemment *réflexif* (identifiable à son bidual).

(c) **Familles orthonormales totales.**
Tout espace de Hilbert possède au moins une famille orthonormale totale (ou "base hilbertienne"), c'est-à-dire une famille $\{e_\lambda\}$, $e_\lambda \in H$, $\|e_\lambda\| = 1$, $(e_\lambda, e_\mu) = 0$ si $\lambda \neq \mu$, avec :

(i) $x \in H$ orthogonal à tous les $e_\lambda \Rightarrow x = 0$ (maximalité).

(ii) L'ensemble des combinaisons linéaires finies des e_λ est dense dans H (totalité[10]).

(iii) $\forall x \in H, \sum_\lambda (x, e_\lambda)^2 = \|x\|^2$ (Parseval).

(iv) $\forall x \in H, \left\| x - \sum_1^n (x, e_\lambda)\, e_\lambda \right\| \to 0$ quand $n \to \infty$ (séries de Fourier).

(v) $\forall x, y \in H, (x,y) = \sum_\lambda (x, e_\lambda)(y, e_\lambda)$ (Riesz-Fischer).

[9] N.B. Certains auteurs, dont L. Schwartz, Bourbaki, ou encore L. Chambadal et J.L. Ovaert dans leur article "Hilbert (espace de)" de l'*Encyclopedia Universalis* [1980], appellent préhilbertien un espace vectoriel muni d'une forme bilinéaire symétrique ***semi-***définie positive, et donc susceptible de n'être que semi-normé. Il faut alors préciser "préhilbertien séparé", ou "euclidien" (ou "hermitien" pour les espaces sur $\mathbb{C}$), pour retrouver la définition utilisée ici. (Remarque communiquée par J. Meinguet).

[10] Il n'est pas question ici de base *algébrique*, dont l'ensemble des combinaisons linéaires finies est exactement H.

Les propriétés qui précèdent sont toutes équivalentes.
Un espace de Hilbert est dit *séparable* s'il admet une base hilbertienne dénombrable.

(d) *Topologie faible.*
Une suite $\{u_k\}$ de H est faiblement Cauchy si $\{(u_k, v)\}$ est Cauchy (dans $\mathbb{R}$, donc convergente) pour tout $v \in H$.
Un espace de Hilbert est toujours faiblement complet : si $\{u_k\}$ est faiblement Cauchy dans H, $\exists u \in H : \forall v \in H, (u_k, v) \to (u, v)$ quand $k \to \infty$.
Dans un espace de Hilbert séparable, la boule unité fermée est faiblement compacte. Donc, si $\|u_k\| \leqslant C$, on peut extraire $\{u_{k_i}\}_i$ qui correspond à $u \in H$ telle que, $\forall v \in H, (u_{k_i}, v) \to (u, v)$ quand $i \to \infty$.
Théorème : si $\{u_k\}$ converge faiblement vers $u \in H$ (espace de Hilbert), *et si* $\|u_k\| \to \|u\|$ quand $k \to \infty$, alors $\{u_k\}$ converge fortement vers u, c'est-à-dire $\|u - u_k\| \to 0$.

Remarque. Concrétisation de la notion de représentant de Riesz : soit U Hilbert réel séparable de base hilbertienne $\{e_k\}_0^\infty$. Alors, si $\varphi(e_k) = \varphi_k$, le représentant de Riesz de φ est

$$f = \sum_0^\infty \varphi_k e_k,$$

et $\|f\| = \|\varphi\| = (\sum_0^\infty \varphi_k^2)^{1/2} = [\sum_0^\infty (\varphi(e_k))^2]^{1/2}$.

3.2. Le problème de l'existence de la solution. Théorème de Lax-Milgram.

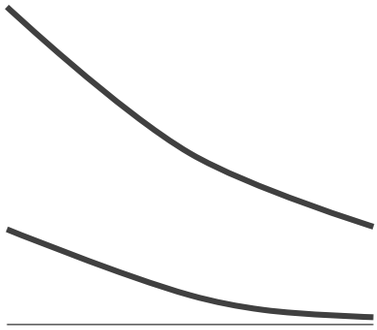
Soit le problème variationnel

$$?u \in U, \forall v \in U : a(u, v) = \varphi(v) \tag{30}$$

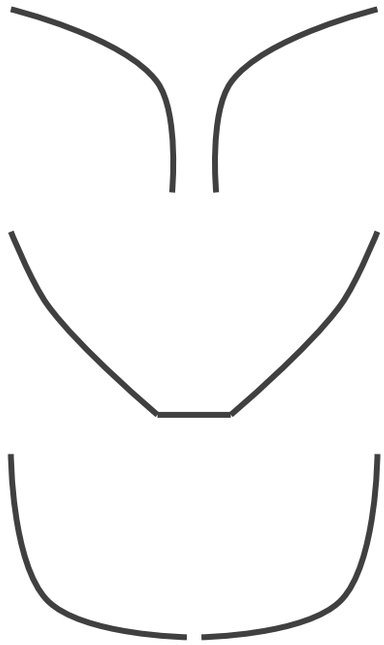
où φ est une forme linéaire continue sur U, et a bilinéaire symétrique continue sur U.

On a vu (p. 38) que, si a est définie positive, résoudre (30) revient à chercher le minimum en $u \in U$ de $J(u) = \dfrac{1}{2}\, a(u, u) - \varphi(u)$, et que ce minimum est unique, ***s'il existe*** dans U.

Cette présentation très séduisante cache quelques pièges :



Abscisses et ordonnées non bornées	Impossible si a coercive
Abscisses non bornées, ordonnée minimale $> -\infty$	Impossible si a coercive



Abscisses bornées, ordonnées → $-\infty$

Impossible si J convexe (J devrait être $-\infty$ partout)

non unicité

Impossible si a définie positive

non existence du minimum

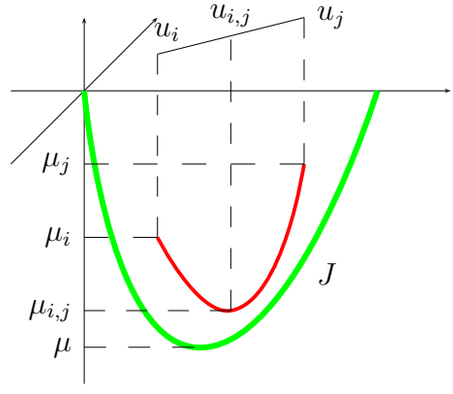
U doit être ***complet***

Et voici enfin le ***théorème d'existence*** :

Théorème de Lax-Milgram. *Si a est bilinéaire continue coercive sur l'espace de Hilbert U, le problème* (30) *a exactement une solution u pour tout $\varphi \in U^*$; de plus, $\|u\| \leqslant c^{-1}\|\varphi\|$* (***stabilité*** de la solution par rapport aux données).

N.B. : par *continuité* de a, on a $$\forall u, v \in U, \quad |a(u,v)| \leqslant d\|u\|\,\|v\|, \tag{31}$$

avec $0 \leqslant d < \infty$ (en effet, d est $\geqslant$ le sup de $|a|$ sur la sphère unité. Si ce sup était $+\infty$, il y aurait une suite $\{u_i, v_i\}$ avec u_i et v_i de norme unité, et $|a(u_i, v_i)| \to \infty$. Considérons alors $r_i = u_i/\sqrt{|a(u_i,v_i)|}$ et $s_i = v_i/\sqrt{|a(u_i,v_i)|}$: $|a(r_i, s_i)| = 1$, alors que r_i et $s_i \to 0 \Rightarrow a(r_i, s_i) \to 0$ par continuité de a).



Démonstration montrant l'existence dans U du minimum de J, si a est symétrique :
Par continuité de φ et coercivité de a, nous savons déjà que

$$\mu := \inf_u J(u) \geqslant -\frac{\|\varphi\|^2}{2c} > -\infty.$$

(Cf. Proposition p. 39).
Considérons une ***suite minimisante*** u_i, $i = 1, 2, \ldots$ telle que $\mu_i := J(u_i) \to \mu$ quand $i \to \infty$.

Il faut montrer que la suite des u_i converge vers un élément de U où J atteint effectivement son infimum μ. Comme l'espace U est complet, il suffit d'établir que la suite des u_i est une suite de Cauchy, donc que $\|u_i - u_j\|$ est aussi petit que l'on veut dès que i et j sont assez grands.

Estimons $\|u_i - u_j\|$ à partir de $\mu_i = J(u_i)$ et $\mu_j = J(u_j)$ proches de (et légèrement supérieurs à) μ. Effectuons encore un "line search" sur les combinaisons de u_i et u_j, soit $u = u_i + \lambda v$, où $v = \dfrac{u_j - u_i}{\|u_j - u_i\|}$:

$$J(u) = \frac{a(u,u)}{2} - \varphi(u) = \frac{a(u_i + \lambda v, u_i + \lambda v)}{2} - \varphi(u_i + \lambda v) =$$

$$\frac{a(u_i,u_i)}{2} - \varphi(u_i) + [a(u_i, v) - \varphi(v)]\lambda + \frac{a(v,v)}{2}\lambda^2$$

prend une valeur minimale $\mu_{i,j} := J(u_i) - \dfrac{[a(u_i,v) - \varphi(v)]^2}{2a(v,v)}$ en $u_{i,j} := u_i + \lambda_{i,j} v$, où $\lambda_{i,j}$ vaut $-\dfrac{a(u_i,v) - \varphi(v)}{a(v,v)}$.

Les valeurs de J autour de $u_{i,j}$ dans la direction v prennent une expression simple :

$$J(u_{i,j} + \lambda v) = \mu_{i,j} + \underbrace{[a(u_{i,j}, v) - \varphi(v)]}_{=0}\lambda + \frac{a(v,v)}{2}\lambda^2.$$

Donc, comme $u_i = u_{i,j} - \lambda_{i,j} v$, et $u_j = u_{i,j} - \lambda_{i,j} v + u_j - u_i = u_{i,j} + (\|u_j - u_i\| - \lambda_{i,j})v$,

$$\mu_i = \mu_{i,j} + \frac{a(v,v)}{2}\lambda_{i,j}^2,$$

$$\mu_j = \mu_{i,j} + \frac{a(v,v)}{2}(\|u_j - u_i\| - \lambda_{i,j})^2,$$

$$\|u_j - u_i\| = \lambda_{i,j} + (\|u_j - u_i\| - \lambda_{i,j}) = \sqrt{\frac{2(\mu_i - \mu_{i,j})}{a(v,v)}} + \sqrt{\frac{2(\mu_j - \mu_{i,j})}{a(v,v)}},$$

qui peut être rendu aussi petit que possible, puisque μ_i, μ_j et $\mu_{i,j}$ sont aussi proches que l'on veut (légèrement supérieurs à μ), et que $a(v,v) \geqslant c > 0$.

Enfin, comme a est continue, $J\left(\lim_{i\to\infty} u_i\right) = \lim_{i\to\infty} J(u_i) = \mu = \inf J$. □

Autres démonstrations :

(1) par compacité faible (a est symétrique) : Soit $\{u_k : \|u_k\| \leqslant \|\varphi\|/c\}$ une suite minimisante : $J(u_k) \to \mu$. Comme la boule $\|u\| \leqslant \|\varphi\|/c$ est faiblement compacte dans U, une suite extraite $\{u_{k_i}\}$ converge faiblement, disons vers $\tilde{u}$.

Montrons que la convergence est forte, et que $J(\tilde{u}) = \mu$:

$$J(u_{k_i}) = \frac{1}{2}a(\tilde{u} + u_{k_i} - \tilde{u}, \tilde{u} + u_{k_i} - \tilde{u}) - \varphi(u_{k_i})$$

$$= \frac{1}{2}a(\tilde{u},\tilde{u}) + a(\tilde{u}, u_{k_i} - \tilde{u}) + \frac{1}{2}a(u_{k_i} - \tilde{u}, u_{k_i} - \tilde{u}) - \varphi(u_{k_i}),$$

où $J(u_{k_i}) \to \mu$ par construction des u_k, $a(\tilde{u}, u_{k_i} - \tilde{u}) = (A\tilde{u}, u_{k_i} - \tilde{u}) \to 0$ par convergence faible, et $\varphi(u_{k_i}) \to \varphi(\tilde{u})$ pour la même raison, donc la suite *positive* $a(u_{k_i} - \tilde{u}, u_{k_i} - \tilde{u})$ a une limite qui est $2(\mu - J(\tilde{u})) \leqslant 0$ puisque μ est l'infimum de J, la limite est donc nulle, et J atteint son infimum = minimum en $\tilde{u}$. On montre ensuite aisément l'unicité de la limite. □

(2) Par l'opérateur linéaire associé à a. Cette démonstration est valable même si a n'est pas symétrique.

Si U est un *espace de Hilbert*, φ admet un *représentant de Riesz* $f \in U$ tel que $\varphi(v) = (f, v)$, et $a(u, .)$ le représentant de Riesz $\mathscr{A}_u$: $a(u, v) = (\mathscr{A}_u, v)$. Bien entendu, comme $a(u, v)$ est *aussi* linéaire en u, $\mathscr{A}_u = Au$, où A est linéaire $U \longrightarrow U$.

On retrouve donc la forme $Au = f$!

Remarquons que A est continu (borné) de norme $\leqslant d$: avec $v = Au$ dans (31),

$$\|Au\|^2 = (Au, Au) = a(u, Au) \leqslant d\|u\| \, \|Au\|,$$

d'où $\|Au\| \leqslant d\|u\|$, donc $\|A\| \leqslant d$.

En effet, a et φ étant continues, on a vu que le problème revient à résoudre $Au = f$.

A est injectif : si $Av = Au$, soit $w = u - v$, $Aw = 0 \Rightarrow c\|w\|^2 \leqslant (Aw, w) = a(w, w) = 0 \Rightarrow w = 0$.
De plus A^{-1} : $AU \mapsto U$ est un opérateur continu et $\|A^{-1}\| \leqslant c^{-1}$. En effet, si $v = Au \in AU$,

$$\|A^{-1}v\|^2 = (u, u) \leqslant c^{-1}a(u, u) = c^{-1}(Au, u) = c^{-1}(v, A^{-1}v) \leqslant c^{-1}\|v\| \, \|A^{-1}v\|.$$

AU est un sous-espace fermé de U : si $\{Au_k\}$ converge dans U, $\forall \varepsilon > 0, k$ et m assez grands $\Rightarrow \|Au_k - Au_m\| \leqslant \varepsilon$, alors
$\|u_k - u_m\|^2 \leqslant c^{-1}a(u_k - u_m, u_k - u_m) = c^{-1}(A(u_k - u_m), u_k - u_m) \leqslant c^{-1}\varepsilon\|u_k - u_m\|$, donc $\{u_k\}$ converge et $\lim(Au_k) = A \lim u_k \in U$ (par continuité de A [ou de a]).

Surjectivité, c'est-à-dire $AU = U$: supposons $f \notin AU$, on peut aussi supposer f orthogonal à AU (retirer de f sa projection $\perp$ sur AU), alors, avec $v = f$:

$$0 = (f, Af) = a(f, f) \geqslant c\|f\|^2 \Rightarrow f = 0.$$

Borne pour $\|u\|$: on a $Au = f$, prenons maintenant $v = u$,

$$c\|u\|^2 \leqslant a(u, u) = (Au, u) = (f, u) \leqslant \|f\| \, \|u\|,$$

$\|u\| \leqslant \|f\|/c$, d'où $\|A^{-1}\| \leqslant c^{-1}$. □

(D'ailleurs, la continuité de A^{-1} découle du théorème de l'inverse continu).

Continuité et stabilité. La borne $\|A^{-1}\| \leqslant c^{-1}$ représente bien un résultat de stabilité : à deux seconds membres f_1 et f_2 proches correspondent des solutions u_1 et u_2 proches : $\|u_2 - u_1\| = \|A^{-1}(f_2 - f_1)\| \leqslant c^{-1}\|f_2 - f_1\|$.

(3) par équivalence de produits scalaires

a bilinéaire symétrique définie positive peut servir de produit scalaire dans U, qui est encore un espace de Hilbert si a est coercive continue : si $\{u_n\}$ est une suite de Cauchy pour a : $\forall \varepsilon, m$ et n assez grands $\Rightarrow a(u_m - u_n, u_m - u_n) \leqslant \varepsilon$, donc $\|u_m - u_n\|^2 \leqslant c^{-1}\varepsilon$, donc $\{u_n\}$ est encore Cauchy pour le produit scalaire initial, donc converge vers $u \in U$, et $a(u - u_n, u - u_n) \leqslant d\|u - u_n\|^2 \to 0$, donc converge aussi selon $a(., .)$.

Alors u est tout simplement le représentant de Riesz de φ dans l'espace de Hilbert U muni du produit scalaire $a(., .)$!!

Bases hilbertiennes et méthodes spectrales. Si on disposait d'une suite orthonormale totale $\{g_k\}$ de U pour $a(.,.)$, c'est-à-dire $a(g_k, g_\ell) = \delta_{k,\ell}$, une bonne construction de u consiste à prendre les sommes partielles de la série

$$a(u,v) = \varphi(v) \Rightarrow u = \sum \varphi(g_k)\, g_k.$$

On disposera plus vraisemblablement d'une suite orthonormale totale de U pour le produit scalaire initial : $(e_k, e_\ell) = \delta_{k,\ell}$.

Peut-on faire d'une pierre deux coups ? Oui : si on veut $a(e_k, e_\ell) = (Ae_k, e_\ell) = 0$ pour tout $e_\ell, \ell \neq k$, sachant que $(e_k, e_\ell) = 0$, il suffit d'avoir $Ae_k = \lambda_k e_k$: valeurs et vecteurs propres de A.
On a $g_k = \lambda_k^{-1/2} e_k$.

Par **exemple**, pour la forme bilinéaire $a(u,v) = \displaystyle\int_0^1 u'(x)v'(x)\,dx$, $\{e_k(x) = \sqrt{2}\sin(k\pi x)\}$, $k = 1, 2, \ldots$ forme une suite orthonormale totale de $L^2(0,1)$;
$\{g_k(x) = \sqrt{2}(k\pi)^{-1}\sin(k\pi x)\}$, $k = 1, 2, \ldots$ vérifie bien $a(g_k, g_\ell) = \displaystyle\int_0^1 g_k'(x) g_\ell'(x)\,dx = \delta_{k,\ell}$.

Si on choisit de partir de L^2, A est un opérateur (non borné !) de valeurs propres $\lambda_k = k^2\pi^2$, $k = 1, 2, \ldots$ défini sur une partie de L^2 ; sur $\mathscr{C}^2 \bigcap \{u : u(0) = u(1) = 0\}$, on a bien $Au = -u''$.

Mais peut-on retrouver $-u'' = f$ à partir de $(u', v') = (f, v)$ avec le produit scalaire de L^2 ? Formellement, oui, mais $A : Au = -u''$ n'est pas un opérateur borné dans L^2 $\longrightarrow$ nécessité de considérer d'autres espaces de Hilbert.

Pour cette forme $a(u,v) = \displaystyle\int_0^1 u'(x)v'(x)\,dx$, un espace de Hilbert où le théorème de Lax-Milgram s'applique est l'espace de Sobolev (voir plus loin) $H_0^1((0,1))$, muni du produit scalaire

$$(u,v)_1 := \int_0^1 u(x)v(x)\,dx + \int_0^1 u'(x)v'(x)\,dx.$$

Concrétisons ici l'opérateur A, au moins dans le sous-espace $\mathscr{C}^2 \bigcap \{u : u(0) = u(1) = 0\}$ de H_0^1 :
la fonction $w = Au$ doit être telle que $(w,v)_1 = a(u,v)$, $\forall v$, ce qui fait

$$\int_0^1 w(x)v(x)\,dx + \int_0^1 w'(x)v'(x)\,dx = \int_0^1 u'(x)v'(x)\,dx \quad \forall v,$$

on retourne à une formulation classique par intégrations par parties, pour avoir

$$?w : \quad w - w'' = -u'' \qquad w(0) = w(1) = 0,$$

que l'on résout, un peu comme (15), en

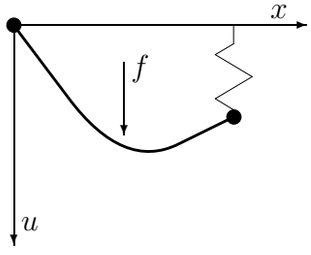
$$\left.\begin{aligned} w(x) = (Au)(x) &= -\int_0^1 \mathcal{G}(x,y)\, u''(y)\, dy \\ &= \int_0^1 \frac{\partial \mathcal{G}(x,y)}{\partial y}\, u'(y)\, dy \\ &= u(x) - \int_0^1 \mathcal{G}(x,y)\, u(y)\, dy\,, \end{aligned}\right\} \tag{32}$$

où $\mathcal{G}(x,y) = \min\left(\dfrac{\sinh(1-y)\sinh x}{\sinh 1}, \dfrac{\sinh(1-x)\sinh y}{\sinh 1}\right)$. Dans la troisième formule de (32), on a utilisé $\partial^2\mathcal{G}(x,y)/\partial y^2 = \mathcal{G}(x,y)$ dans l'intégration par parties ; le terme en $u(x)$ provient de la discontinuité de $\partial\mathcal{G}(x,y)/\partial y$ en $y = x$. En tout cas, on voit bien maintenant que A est un opérateur borné, les fonctions $\mathcal{G}$ et $\partial\mathcal{G}/\partial x$ étant bornées dans $[0,1]\times[0,1]$.

Et son inverse A^{-1} ? C'est plus simple ! Il faut résoudre $w - w'' = -u''$ en u, ce qui donne, d'après (15),

$$u(x) = (A^{-1}w)(x) = w(x) - \int_0^1 G(x,y)\, w(y)\, dy.$$

Encore un exemple, montrant le rôle des conditions aux limites essentielles et naturelles :



Une corde tendue est fixée à son extrémité de gauche à un point fixe, mais son extrémité de droite est attachée à un ressort vertical de constante de raideur $\alpha \geqslant 0$. L'énergie se constitue de

(1) Déplacement de la charge f : $\quad -\displaystyle\int_0^1 f u\, dx,$

(2) Déformation de la corde $\dfrac{1}{2}\displaystyle\int_0^1 u'^2\, dx, \qquad (T = 1),$

(3) Déformation du ressort $\dfrac{\alpha}{2}(u(1))^2,$

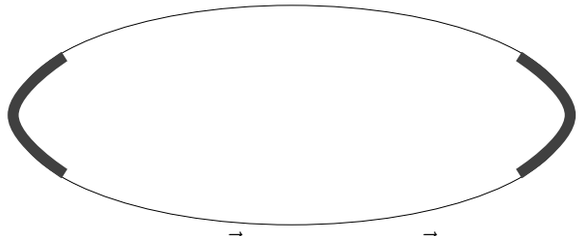
Ce qui correspond à $a(u,v) = \int_0^1 u'v'\, dx + \alpha u(1)v(1)$, à évaluer sur les fonctions s'annulant en 0. En intégrant par parties pour "chasser" v', on obtient $-u'' = f$ sur $(0,1)$, avec $u(0) = 0$ (condition essentielle), et $u'(1) = -\alpha u(1)$ (condition naturelle [condition de ***Robin***]).

Les conditions aux limites ***essentielles*** figurent donc dans les spécifications des espaces vectoriels de fonctions à considérer ; les conditions aux limites ***naturelles*** n'apparaissent pas dans ces espaces, mais peuvent apparaître dans la forme a.

Dans un problème à plusieurs dimensions, on aura typiquement u donnée (**condition de Dirichlet**) sur une partie de la frontière Γ, et la dérivée normale $\partial u/\partial n$ donnée (**condition**

de Neumann) sur la partie complémentaire de Γ. Enfin, si on fixe une combinaison de u et de sa dérivée normale, on parle de **condition de Robin**[11].

Exemple : résistance d'un conducteur[12]**.** Un volume Ω d'une substance conductrice isotrope, mais non nécessairement homogène, de conductivité $\sigma(x, y, z)$, est soumis à des potentiels V_1 et V_2 appliqués en deux électrodes formant une partie $\partial\Omega_D$ (D comme *Dirichlet*) de la frontière de Ω[13].



On a $\vec{I}(x, y, z) = \sigma(x, y, z)\vec{E}(x, y, z)$, $\mathbf{div}\,\vec{I} = 0$ ("le courant est incompressible") et $E = -\,\mathbf{grad}\,V$, d'où

$\mathbf{div}\,\sigma(x, y, z)\,\mathbf{grad}\,V(x, y, z) = 0$ dans Ω, avec V donné sur $\partial\Omega_D$ et la condition de *Neumann* $\partial V/\partial n = 0$ sur $\partial\Omega \setminus \partial\Omega_D$ (aucun courant ne passe par la partie de $\partial\Omega$ non couverte par les électrodes).

Constatons que la solution minimise la puissance dissipée $P = \int_\Omega \vec{I} \cdot \vec{E} dxdydz = \int_\Omega \sigma(x, y, z)\mathbf{grad}\,V \cdot \mathbf{grad}\,V\; dxdydz$ sous la seule contrainte $V = V_1$ ou V_2 sur $\partial\Omega_D$. En effet, on arrive à la forme variationnelle

$\int_\Omega \sigma(x, y, z)\mathbf{grad}\,V \cdot \mathbf{grad}\,W\; dxdydz = 0$, avec $W = 0$ sur $\partial\Omega_D$, ou, en appliquant (20) p. 35 avec $F = \sigma\ \mathbf{grad}\,V$ et $G = W$, $-\int_\Omega \mathbf{div}\ [\sigma\mathbf{grad}\,V]\ W\; dxdydz + \int_{\partial\Omega} \sigma W \frac{\partial V}{\partial n} dS = 0$, on voit bien que le terme aux frontières s'annule.

Enfin, la résistance est $(V_2 - V_1)/I_{\text{totale}}$, où I_{totale} est l'intégrale de $\vec{I}$ sur toute section équipotentielle de Ω (faire $W = 1$ plus haut pour voir que cette intégrale est la même sur tout section équipotentielle). Par exemple, pour un cylindre homogène de longueur L, $V(x, y, z) = V_1 + (V_2 - V_1)x/L$, $\vec{I} = \sigma(V_2 - V_1)[1, 0, 0]/L$, $I_{\text{totale}} = \sigma(V_2 - V_1)|S|/L$, $R = L/(\sigma|S|)$.

Répartition de courant (Landau & Lifshitz, *Electrodynamique des milieux continus*).

4. Relations entre méthodes de projection : Galerkin, etc.

Considérons encore le problème

$$?u \in U : \qquad a(u, v) = (f, v)\,, \qquad \forall v \in V, \tag{33}$$

[11] Sur la vie et l'oeuvre de Robin, lire K. Gustafson, T. Abe : "The third boundary condition- was it Robin's ?", *Math. Intelligencer* vol. **20** (1998), nr. 1, pp. 63-71 ; "(Victor) Gustave Robin : 1855-1897", *ibid.*, vol. **20**, nr. 2, pp. 47-53.

[12] Merci à J. Ruppert (MATH22, 1998-1999) pour cet exemple.

[13] Le symbole Ω n'est peut être pas très heureux ici ; par ailleurs, l'unité de conductance est évidemment le mho $\mho$.

posé dans les espaces de Hilbert (réels) U et V, où a est une forme bilinéaire, non nécessairement symétrique.

Comme vu précédemment, le problème revient à résoudre

$$Au = f,$$

où Au est le représentant de Riesz dans V de la forme $v \mapsto a(u, v)$.

Toutes les méthodes de projection de solution approchée de ce problème reviennent

(1) à choisir un espace $U_h \subset U$ de dimension finie N,

(2) à choisir un projecteur P_h sur un sous-espace $V_h \subset V$ de même dimension,

et à résoudre

$$?u_h \in U_h : \qquad P_h(Au_h - f) = 0,$$

ce qui donne lieu à un système de N équations linéaires à N inconnues, les coefficients c_j de u_h dans une base $\{\phi_1, \ldots, \phi_N\}$ de U_h :

$$\sum_{j=1}^{N} (P_h A \phi_j)_i \; c_j = (P_h f)_i \,, \qquad i = 1, \ldots, N, \tag{34}$$

où $(.)_i$ désigne le $i^{\text{ème}}$ coefficient du développement d'un élément de V_h dans une base de V_h.

On parle aussi de *méthode de résidu pondéré*. Le résidu est $Au_h - f$.

On n'a aucune assurance de non singularité de la matrice A_h de (34), même si on sait si A est inversible.

Les choix les plus courants sont :

4.1. Galerkin. $U = V$, $U_h = V_h$, $P_h =$ projecteur orthogonal sur U_h.

$P_h(Au_h - f) = 0$ signifie donc $Au_h - f$ (le résidu) est *orthogonal* à tout élément v de U_h : $a(u_h, v) = (Au_h, v) = (f, v)$, $\forall v \in U_h$, ce qui donne lieu au système d'équations linéaires à N inconnues

$$\sum_{j=1}^{N} a(\phi_j, \phi_i) \; c_j = (f, \phi_i) \,, \qquad i = 1, \ldots, N.$$

4.2. Ritz. Cas particulier du précédent, avec a symétrique.

C'est bien sûr le cas essentiellement traité dans ce cours.

4.3. Galerkin-Petrov. $P_h =$ projecteur orthogonal sur V_h.

V étant normalement différent de U (tous deux Hilbert), on a la généralisation suivante (dite "condition inf-sup") du théorème de Lax-Milgram :

Théorème (Nečas, cité dans [Quarteroni & Valli], § 5.1) : *si la forme a vérifie*

$$\left.\begin{aligned} |a(u, v)| &\leqslant d\|u\| \, \|v\|, \qquad \forall u \in U, v \in V, \\ \sup_{\substack{v \in V \\ v \neq 0}} \frac{a(u, v)}{\|v\|} &\geqslant c \, \|u\| \,, \qquad \forall u \in U, \\ \sup_{u \in U} a(u, v) &> 0 \qquad \forall v \in V, v \neq 0, \end{aligned}\right\} \tag{35}$$

le problème (33) *a une et une seule solution pour tout $f \in V$; cette solution est bornée par $\|f\|/c$.*

(La deuxième condition peut s'écrire sous forme "inf-sup" proprement dite $\inf\limits_{\substack{u\in U\\ u\neq 0}} \sup\limits_{\substack{v\in V\\ v\neq 0}} \dfrac{a(u,v)}{\|u\|\ \|v\|} \geqslant c > 0$, mais il faut remarquer que le sup est évalué avant l'inf, donc que v dépend de u. Dans le cas coercif, on prend $v = u$).

En effet, la première condition de (35) assure la continuité de a, donc l'existence du représentant de Riesz $a(u,v) = (Au, v)$. Par la deuxième condition, l'opérateur A est injectif, si $Au_1 = Au_2 = f$, $a(u_1 - u_2, v) = 0$ pour tout $v \in V$, d'où $\|u_1 - u_2\| = 0$: unicité de la solution.

On peut donc définir l'opérateur A^{-1} dans la partie AU de V, opérateur qui fait correspondre à tout $f \in AU$ la solution u de $Au = f$.

Cet opérateur A^{-1} est borné par $c^{-1} < \infty$ (stabilité) : toujours par la deuxième condition de (35),

$$\|f\|\ \|v\| \geqslant |(f,v)| = |a(u,v)| \geqslant c\|u\|\ \|v\| = c\|A^{-1}f\|\ \|v\|,$$

d'où $\|A^{-1}\| \leqslant c^{-1}$.

Enfin, A est surjectif : supposons $AU \subset V$ avec $AU \neq V$. Il y a donc un f hors de AU. L'espace AU est fermé : si la suite $\{Au_n\}$ converge dans V, il en est de même de la suite $\{u_n\}$, l'opérateur A^{-1} étant borné, donc la limite des Au_n vaut A fois la limite des u_n, qui est dans AU. L'espace AU étant fermé, donc complet, f y admet une projection orthogonale p. Enfin, avec $v = f - p$ dans la troisième condition de (35), on aurait $a(u,v) = (Au, v) = 0$ par orthogonalité de v avec tout l'espace AU, ce qui est impossible, par la troisième condition de (35), si $v = f - p \neq 0$. □

Les équations du problème approché sont

$$\sum_{j=1}^{N} a(\phi_j, \psi_i)\ c_j = (f, \psi_i)\,, \qquad i = 1, \ldots, N. \tag{36}$$

où $\{\psi_1, \ldots, \psi_N\}$ est une base de V_h.

La matrice du système est $A_h = [a(\phi_j, \psi_i)]_{i,j=1,\ldots,N}$. Si on multiplie A_h par un vecteur $[c_1, \ldots, c_N]^T$ de coefficients d'un élément $u = \sum_1^N c_j\phi_j$ de U_h, on obtient le vecteur des $a(u, \psi_i) = (Au, \psi_i)$ qui permet de reconstituer la projection de Au dans V_h : A_h est donc l'opérateur P_hA restreint à U_h.

Si les conditions inf-sup (35) restent valables quand on restreint a à $U_h \times V_h$, on établit l'existence de la solution u_h, et on a aussi

$$\|u - u_h\| \leqslant (1 + d/c) \inf_{w_h \in U_h} \|u - w_h\|.$$

(Babuška-Strang, cité dans [Quarteroni & Valli], § 5.3). Démonstration habituelle : on a, pour tout $w_h \in U_h$ et un v encore indéterminé dans V :

$$d\|u - w_h\|\ \|v\| \geqslant a(u - w_h, v) = a(u - u_h, v) + a(u_h - w_h, v)$$

et prenons $v \in V_h$ réalisant la condition inf-sup pour $u_h - w_h \in U_h$: on a donc $a(u_h - w_h, v) \geqslant c\|u_h - w_h\|\ \|v\|$ et $a(u - u_h, v) = 0$ (rappelons que $A(u - u_h)$ est orthogonal à V_h), donc $d\|u - w_h\| \geqslant c\|u_h - w_h\| \geqslant c[\|u - u_h\| - \|u - w_h\|]$ (tout côté est supérieur à la différence des deux autres).

Autre démonstration[14], aboutissant à d/c au lieu de $1 + d/c$

[14] Xu, Jinchao ; Zikatanov, Ludmil : Some observations on Babuška and Brezzi theories. *Numer. Math.* **94**, No.1, 195-202 (2003). Merci à B. Delsaute qui a signalé cette référence (mai 2005).

En plus de P_h, projecteur orthogonal dans V_h, on considère le projecteur orthogonal Π_h dans U_h. $\Pi_h u$ est donc l'élément de U_h le plus proche de u, tiens, tiens...

On partitionne $Au = f$ en

$$\left[\begin{array}{c|c} P_hA\Pi_h & P_hA(I-\Pi_h) \\ \hline (I-P_h)A\Pi_h & (I-P_h)A(I-\Pi_h) \end{array}\right] \left[\begin{array}{c} \Pi_h u \\ \hline (I-\Pi_h)u \end{array}\right] = \left[\begin{array}{c} P_hf \\ \hline (I-P_h)f \end{array}\right]$$

Le premier bloc $P_hA\Pi_h$ est identique à A_h (voir plus haut), on compare donc aisément u_h et $\Pi_h u$ en remplaçant simplement P_hf par A_hu_h. Alors,

$$A_h(\Pi_h u - u_h) = -P_hA(I-\Pi_h)u,$$

et, comme la condition inf-sup est supposée être encore valable pour A_h (en tant qu'opérateur $U_h \to V_h$), $\|A_h^{-1}\| \leqslant c^{-1}$, et comme $\|P_h\| = 1$ (projecteur orthogonal), $\|A\| \leqslant d$, on a

$$\|\Pi_h u - u_h\| \leqslant (d/c)\|u - \Pi_h u\|.$$

Passage à $u - u_h$ (pas évident ! !) : $u - u_h = \Pi_h u - u_h + u - \Pi_h u = (I - S_h)(u - \Pi_h u)$, où S_h est l'opérateur $A_h^{-1}P_hA$. On vient de voir que S_h a une norme bornée par d/c. Montrons que la même borne vaut pour $\|I - S_h\|$.

Lemme (Kato) *Si $\boldsymbol{P}$ est un projecteur (c.-à-d. $\boldsymbol{P}^2 = \boldsymbol{P}$) linéaire dans un espace préhilbertien E, avec $\boldsymbol{P} \neq 0$ et $\boldsymbol{P} \neq \boldsymbol{I}$, alors $\|\boldsymbol{I} - \boldsymbol{P}\| = \|\boldsymbol{P}\|$.*

Démonstrations.

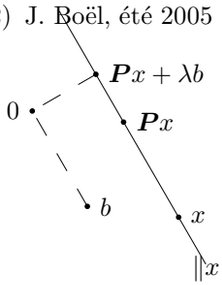
(1) Kato[15]). Montrons que $\|\boldsymbol{I} - \boldsymbol{P}\| \geqslant \|\boldsymbol{P}\|$ si $\boldsymbol{P} \neq \boldsymbol{I}$ et $\boldsymbol{P} \neq 0$. Ensuite, on permutera évidemment $\boldsymbol{P}$ et $\boldsymbol{I} - \boldsymbol{P}$.

En effet, $\boldsymbol{P}$ n'étant pas l'opérateur nul, on a déjà $\|\boldsymbol{P}\| \geqslant 1$ (car $0 \neq \boldsymbol{P} = \boldsymbol{P}^2$). Comme il en est de même de $\boldsymbol{I} - \boldsymbol{P}$, on a aussi $\|\boldsymbol{I} - \boldsymbol{P}\| \geqslant 1$. La relation cherchée $\|\boldsymbol{I} - \boldsymbol{P}\| \geqslant \|\boldsymbol{P}\|$ est donc vraie si $\|\boldsymbol{P}\| = 1$, supposons $\|\boldsymbol{P}\| > 1$. Alors, pour tout $a \in (1, \|\boldsymbol{P}\|)$, il existe $u \in E$ avec $\|\boldsymbol{P}u\| \geqslant a\|u\| > 0$. Soit v la combinaison linéaire de u et $\boldsymbol{P}u$ telle que v soit orthogonal à $\boldsymbol{P}u$: $v = u - \lambda\boldsymbol{P}u$, avec $\lambda = (u, \boldsymbol{P}u)/\|\boldsymbol{P}u\|^2$.

Montrons que $\dfrac{\|v - \boldsymbol{P}v\|}{\|v\|} \geqslant \dfrac{\|\boldsymbol{P}u\|}{\|u\|} \geqslant a$, pour tout $a < \|\boldsymbol{P}\|$, ce qui établira $\|\boldsymbol{I} - \boldsymbol{P}\| \geqslant \|\boldsymbol{P}\|$.

Ecrivons tout en fonction de u et $\boldsymbol{P}u$: $\boldsymbol{P}v = (1-\lambda)\boldsymbol{P}u$, donc $v - \boldsymbol{P}v = u - \boldsymbol{P}u$, $\|v - \boldsymbol{P}v\|^2 = (u - \boldsymbol{P}u, u - \boldsymbol{P}u) = \|u\|^2 - 2(u, \boldsymbol{P}u) + \|\boldsymbol{P}u\|^2$; $\|u\|^2 = \|v\|^2 + \lambda^2\|\boldsymbol{P}u\|^2$ par orthogonalité de v et $\boldsymbol{P}u$, donc, $\|v\|^2 = \|u\|^2 - (u, \boldsymbol{P}u)^2/\|\boldsymbol{P}u\|^2$. Il reste $\|u\|^2\{\|u\|^2 - 2(u, \boldsymbol{P}u) + \|\boldsymbol{P}u\|^2\} \geqslant \{\|u\|^2 - (u, \boldsymbol{P}u)^2/\|\boldsymbol{P}u\|^2\}\|\boldsymbol{P}u\|^2$, ou $\|u\|^4 - 2(u, \boldsymbol{P}u)\|u\|^2 + (u, \boldsymbol{P}u)^2 = \{\|u\|^2 - (u, \boldsymbol{P}u)\}^2 \geqslant 0$[16].

(2) J. Roël, été 2005 :



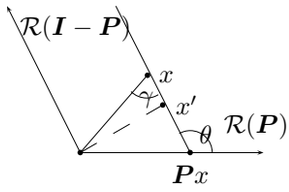
Montrons que pour tout $x \in E$ avec $x \neq 0$ et $b := x - \boldsymbol{P}x \neq 0$, il existe $y \neq 0$ tel que $\dfrac{\|x - \boldsymbol{P}x\|}{\|x\|} = \dfrac{\|b\|}{\|x\|} \leqslant \dfrac{\|\boldsymbol{P}y\|}{\|y\|}$. On aura alors effectivement $\|\boldsymbol{I} - \boldsymbol{P}\| \leqslant \|\boldsymbol{P}\|$. Ensuite, on permutera évidemment $\boldsymbol{P}$ et $\boldsymbol{I} - \boldsymbol{P}$. Comme $b \neq 0$, $\exists!\lambda \in \mathbb{R} : (\boldsymbol{P}x + \lambda b, b) = 0$, et prenons pour y ce $\boldsymbol{P}x + \lambda b$ orthogonal à b. Remarquons que $\boldsymbol{P}b = \boldsymbol{P}x - \boldsymbol{P}^2x = 0$, donc $\boldsymbol{P}y = \boldsymbol{P}x$. Normes : $\|\boldsymbol{P}y\|^2 = \|\boldsymbol{P}x\|^2 = \|y - \lambda b\|^2 = \|y\|^2 + \lambda^2\|b\|^2$,
$\|x\|^2 = \|y + (1-\lambda)b\|^2 = \|y\|^2 + (1-\lambda)^2\|b\|^2$.

Il faut montrer $\|b\|^2\|y\|^2 \leqslant \|\boldsymbol{P}y\|^2\|x\|^2$, ou $0 \leqslant \|y\|^4 + (\lambda^2 + (1-\lambda)^2 - 1)\|y\|^2\|b\|^2 + \lambda^2(1-\lambda)^2\|b\|^4$, ce qui vaut $[\|y\|^2 - \lambda(1-\lambda)\|b\|]^2$.

(3) J. Meinguet, été 2005 :

[15] Lemme 4 de l'appendice de Kato, Tosio : Estimation of iterated matrices, with application to the von Neumann condition. *Numer. Math.* **2** 1960 22–29.

[16] Si le champ des scalaires est $\mathbb{C}$, on a. $\|u\|^4 - 2\mathrm{Re}\,(u, \boldsymbol{P}u)\|u\|^2 + |(u, \boldsymbol{P}u)|^2 = \{\|u\|^2 - \mathrm{Re}\,(u, \boldsymbol{P}u)\}^2 + (\mathrm{Im}\,(u, \boldsymbol{P}u))^2 \geqslant 0$.



Soit θ l'angle entre $\boldsymbol{P}x$ et $x-\boldsymbol{P}x$ (nous prenons évidemment $\boldsymbol{P}$ et $\boldsymbol{I}-\boldsymbol{P}$ non triviaux, et x et $x-\boldsymbol{P}x$ non nuls). Loi des sinus dans le triangle $0, x, \boldsymbol{P}x$: $\dfrac{\|\boldsymbol{P}x\|}{\|x\|} = \dfrac{\sin\gamma}{\sin\theta} \leqslant \dfrac{1}{\sin\theta}$. On peut d'ailleurs toujours atteindre la borne $1/\sin\theta$, il suffit de remplacer x par x' = la projection orthogonale de 0 sur le côté $(x, \boldsymbol{P}x)$: $\dfrac{\|\boldsymbol{P}x = \boldsymbol{P}x'\|}{\|x'\|} = \dfrac{1}{\sin\theta}$. On a donc $\|\boldsymbol{P}\| = 1/\sin\theta_{\inf}$, où $\theta_{\inf}$ est l'infimum sur tous les $x \in E$ (avec $x \neq 0$ et $x - \boldsymbol{P}x \neq 0$) de l'angle entre $\boldsymbol{P}x$ et $x-\boldsymbol{P}x$. Cette quantité est évidemment la même pour $\boldsymbol{P}$ et $\boldsymbol{I}-\boldsymbol{P}$[17].

□

Nous avons donc

$$\|u-u_h\| = \|(I-S_h)(u-\Pi_h u)\| \leqslant \|S_h\|\,\|(u-\Pi_h u)\| \leqslant (d/c)\|u-\Pi_h u\|.$$

4.4. Moindres carrés. On cherche u_h dans U_h qui minimise $\|Au_h - f\|$ dans V_h : $\|A(u_h+\varepsilon w_h)-f\|^2 - \|Au_h-f\|^2 = (A(u_h+\varepsilon w_h)-f, A(u_h+\varepsilon w_h)-f) - (Au_h-f, Au_h-f) = 2(Au_h - f, Aw_h)\varepsilon + (Aw_h, Aw_h)\varepsilon^2$ positif pour $\forall\varepsilon$ réel seulement si $(Au_h - f, Aw_h) = 0$, $\forall w_h \in U_h$: cela revient à prendre $V_h = AU_h$.

4.5. Collocation. P_h = interpolant dans V_h : l'équation est résolue exactement en un nombre fini de points imposés.

[17] $\theta_{\inf}$ est l'angle entre les espaces d'arrivée $\mathcal{R}(\boldsymbol{P})$, $\mathcal{R}(\boldsymbol{I}-\boldsymbol{P})$ de $\boldsymbol{P}$ et $\boldsymbol{I}-\boldsymbol{P}$. En effet, l'angle entre deux espaces F et $G \subset E$ est le plus petit angle entre $u \in F$ et $v \in G$ (G. Golub, C. Van Loan, *Matrix Computations*, North Oxford Academic, 1983 (au moins deux autres éditions)). Ici, $u = \boldsymbol{P}y$, $v = (\boldsymbol{I}-\boldsymbol{P})z$, avec y et z quelconques dans E, et indépendants semble-t-il. Mais, avec $x = u+v$, on a bien $\boldsymbol{P}x = \boldsymbol{P}u + \boldsymbol{P}v = u$ et $(\boldsymbol{I}-\boldsymbol{P})x = (\boldsymbol{I}-\boldsymbol{P})u + (\boldsymbol{I}-\boldsymbol{P})v = v$.

Chapitre 2

Méthode des éléments finis.

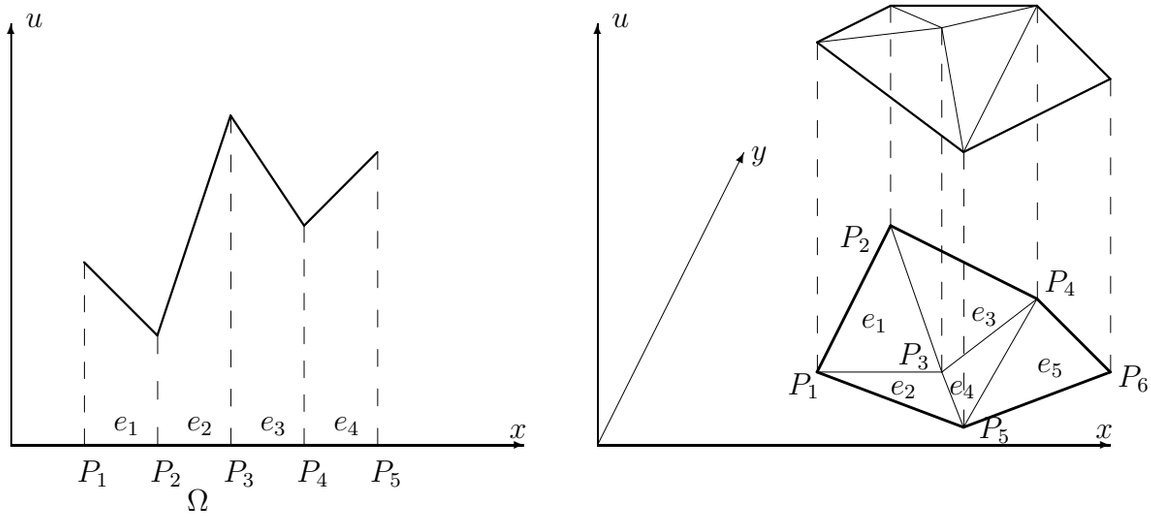
1. Introduction et définition.

Il est question de résoudre le problème variationnel maintenant familer

$$?u_h \in U_h : \qquad a(u_h, v) = \varphi(v), \qquad \forall v \in U_h,$$

(il s'agit donc bien de l'approximation de Ritz)

dans des espaces de dimension finie U_h dont les échantillons les plus simples sont, respectivement pour $\Omega \subset \mathbb{R}$ et $\Omega \subset \mathbb{R}^2$:



On note que

(1) Ω est divisé en sous-ensembles, les ***éléments***, suffisamment simples et petits. Evidemment ($n > 1$), on ne peut déjà plus prendre n'importe quel domaine Ω si on se limite à des éléments polygonaux ou polyédriques. On acceptera souvent de modifier Ω.

(2) La restriction de toute fonction de U_h à un e_i est une fonction simple, souvent un polynôme de degré faible (ici, degré 1 : ligne brisée à 1 dimension, espace de Courant à deux dimensions).

(3) La valeur d'une fonction de U_h dans un élément e_k est entièrement déterminée par ses valeurs aux points $P_i \in e_k$ (***interpolation***).

La méthode des éléments finis est ce dialogue (trialogue, en fait) entre géométrie des éléments, espace de fonctions associées à un élément et formes d'interpolation.

Plus précisément :

Définition. *Un espace U_h d'éléments finis est un espace de fonctions définies par leurs restrictions sur des parties e_k, $k = 1, 2, \ldots, E < \infty$ de $\overline{\Omega}$, et ayant les propriétés suivants :*

(1) (a) *Chaque e_k est un fermé régulier pour le théorème de la divergence (cf. p. 31),*

(b) *la mesure de $e_k \cap e_\ell$ est nulle si $k \neq \ell$($\iff$ $\overset{\circ}{e_k} \cap \overset{\circ}{e_\ell} = \varnothing$ si $k \neq \ell$: intérieurs disjoints),*

(c) $\overline{\Omega} = \bigcup_{k=1}^{E} e_k$.

(2) *La restriction d'une fonction de U_h à e_k est dans un espace donné (de dimension finie) V_k.*

(3) *On se donne des formes F_i associées à des points P_i (non nécessairement distincts) $\in \overline{\Omega}$.* ***La restriction d'une fonction u de U_h à e_k doit pouvoir être entièrement déterminée par les valeurs $F_i(u)$ avec $P_i \in e_k$*** *(interpolation).*
$F_i(v)$ est le plus souvent une combinaison linéaire de $v(P_i)$ et de dérivées de v au point P_i.
*On dit que l'ensemble $\{F_i\}_{i\in Q_k}$, où Q_k est l'ensemble des indices des noeuds appartenant à e_k, doit être V_k-*unisolvant.
En particulier, la dimension de V_k doit être égale à la cardinalité de Q_k.

$u_{\mathbf{c}}$ est alors la fonction dont la restriction à e_k est la solution du problème

$$(P_k) \qquad\qquad F_i(v) = c_i \ , i \in Q_k.$$

On obtient ainsi un grand ensemble $\{F_1, \ldots, F_N\}$ U_h-unisolvant à partir d'ensembles beaucoup plus simples. En particulier, quand h (diamètre moyen des e_k) tend vers zéro, N tend vers l'infini, mais on demandera que les dimensions des V_k restent bornées.

Des auteurs (Ciarlet [*The Finite Element Method for Elliptic Problems*, p. 78], , Raviart) ont donné cette *définition* formelle d'un élément fini : un élément fini est un ensemble de triples $(e_k, V_k, \{F_i\}_{i\in Q_k})$ associés aux e_k, avec $\{F_i\}_{i\in Q_k}\}V_k-$unisolvant **et** $F_i(v)$ prenant la même valeur sur tous les e_k tels que $i \in Q_k$: cette condition remplace la mention d'un point P_i et autorise des éléments finis (partiellement) non nodaux.
Exemple : $e_k = [x_k, x_{k+1}]$, $F_{2k}(v) = v(x_k)$, $F_{2k+1}(v) = \int_{x_k}^{x_{k+1}} v(t)dt$; $V_k = \mathscr{P}_2$. $Q_k = \{2k, 2k+1, 2k+2\}$.

2. Ensembles unisolvants, interpolation.

Soient $F_1, \ldots, F_m$ des formes définies sur un espace vectoriel (réel) W de dimension m.
On dit que $\{F_1, \ldots, F_m\}$ **forme un ensemble $W-$unisolvant** si le problème

$$F_i(v) = c_i \ , i = 1, \ldots, m \tag{37}$$

a exactement une seule solution $v \in W$ pour tout $\mathbf{c} = [c_1, \ldots, c_m]^T \in \mathbb{R}^m$.

Les F_i doivent évidemment être indépendantes : si une combinaison $\sum_1^m \lambda_i F_i$ était la forme nulle, avec au moins un $\lambda_i \neq 0$, on ne pourrait résoudre (P) avec des c_i vérifiant $\sum_1^m \lambda_i c_i \neq 0 \ldots$

Soit $w_1, \ldots, w_m$ une base de W. La condition d'unisolvance est alors :

$$\det \mathbf{M} = \det \begin{bmatrix} F_1(w_1) & \ldots & F_1(w_m) \\ \ldots & \ldots & \ldots \\ F_m(w_1) & \ldots & F_m(w_m) \end{bmatrix} \neq 0$$

puisque chercher à résoudre (37) par la représentation $w = \sum_1^m \mu_k w_k$ revient précisément à résoudre le système d'équations linéaires $\mathbf{M}\boldsymbol{\mu} = \mathbf{c}$.

Important : on évite souvent le recours au calcul explicite d'un déterminant en exploitant un des énoncés équivalents suivants :

Théorème. *Soient $F_1, \ldots, F_m$ des formes définies sur un espace vectoriel (réel) W de dimension m. Alors*

(1) *$\{F_1, \ldots, F_m\}$ est $W-$unisolvant.*

(2) *$\det[F_i(w_j)]_1^m \neq 0$ pour toute base $\{w_i\}$ de W.*

(3) *Seul l'élément nul de W annule toutes les valeurs $F_i(w)$, $i = 1, \ldots, m$.*

(4) *On peut trouver m éléments $L_1, \ldots, L_m$ de W vérifiant $F_i(L_j) = \delta_{i,j}$, $i, j = 1, \ldots, m$.*

sont équivalents.

En effet, on a déjà 1) $\iff$ 2).
2) $\iff$ 3) : le déterminant d'une matrice carrée est non nul si et seulement si le système d'équations linéaires homogènes construit avec cette matrice n'admet que la solution nulle [1].
1) $\Rightarrow$ 4) : si $\{F_1, \ldots, F_m\}$ est unisolvant, on peut trouver $v \in W$ tel que $F_i(v) = \delta_{i,j}$, $i = 1, \ldots, m$. Appelons L_j cette solution v.
4) $\Rightarrow$ 1) : constatons que, $\forall \mathbf{c} \in \mathbb{R}^m, v = \sum_{j=1}^{m} c_j L_j$ est un élément de W qui vérifie $F_i(v) = c_i$, $i = 1, \ldots, m$. On peut donc résoudre $\mathbf{M}\mathbf{v} = \mathbf{c}$ pour tout $\mathbf{c} \in \mathbb{R}^m$, donc, $\det \mathbf{M} \neq 0$. □

Un étudiant MAP22 a très justement remarqué (en septembre 2000) que
3) signifie que la matrice $\mathbf{M}$ représente une application ***injective*** dans $\mathbb{R}^m$, et que

4) signifie que la matrice $\mathbf{M}$ représente une application ***surjective*** !

On appelle $\{L_1, \ldots, L_m\}$ la ***base de Lagrange*** de W associée à $\{F_1, \ldots, F_m\}$.

Changement de base : chaque élément $\tilde{w}_\ell$ d'une nouvelle base de W s'exprime en fonction des w_k :

$$[\tilde{w}_1, \ldots, \tilde{w}_m] = [w_1, \ldots, w_m]\mathbf{Q},$$

avec $\det \mathbf{Q} \neq 0$, d'où

$$\mathbf{M} = \begin{bmatrix} F_1 \\ \ldots \\ F_m \end{bmatrix} [w_1, \ldots, w_m] = \begin{bmatrix} F_1 \\ \ldots \\ F_m \end{bmatrix} [\tilde{w}_1, \ldots, \tilde{w}_m]\mathbf{Q}^{-1} = \widetilde{\mathbf{M}}\mathbf{Q}^{-1}.$$

La base de Lagrange de W associée à $\{F_1, \ldots, F_m\}$ est la base **biorthogonale** à $\{F_1, \ldots, F_m\}$, c'est-à-dire $\mathbf{M} = \mathbf{I}$:

$$F_i(L_j) = \delta_{i,j}\ , i, j = 1, \ldots, m.$$

[1] cas élémentaire de l'alternative de Fredholm...

La solution de (37) dans la base de Lagrange de W est donc

$$v = \sum_1^m c_i\, L_i \quad = \sum_1^i F_i(v)\, L_i. \tag{38}$$

L'interpolation d'une fonction f dans W selon $\{F_1, \ldots, F_m\}$ consiste à trouver $v \in W$ tel que $F_i(v) = F_i(f)$, $i = 1, \ldots, m$. L'interpolant de f est donc

$$v = \sum_1^i F_i(f)\, L_i. \tag{39}$$

Les L_i sont aussi appelées ***fonctions de forme*** [Zienkiewicz].

La base de Lagrange de U_h pour $\{F_1, \ldots, F_N\}$ est simplement $\{u_1, \ldots, u_N\}$ avec $u_j = u_{[0,\ldots,0,1,0\ldots,0]}$, c'est-à-dire la fonction dont la restriction à e_k est l'élément correspondant de la base de Lagrange de V_k. Il s'ensuit que la restriction de u_j à e_k sera la fonction nulle dès que $j \notin Q_k$, c'est-à-dire dès que $P_j \notin e_k$: *le support de u_j n'est constitué que des éléments e_k contenant P_j.*

Les bon choix d'éléments finis conduisent à des bases de Lagrange de petits supports. La matrice de rigidité $[a(u_i, u_j)]$ aura alors de très nombreux éléments nuls (matrice creuse).

Remarque : si $k \neq \ell$, les intérieurs de e_k et e_ℓ sont disjoints mais les (fermés) e_k et e_ℓ peuvent avoir une intersection non vide. Il importe alors que les définitions de $u \in U_h$ par ses restrictions sur e_k et e_ℓ coincident sur $e_k \cap e_\ell$ (éléments finis *conformes*) !

Dans les exemples qui suivent, on décrira

(1) l'élément e_k,

(2) l'espace V_k,

(3) les formes F_i,

et on *vérifiera*

(1) l'unisolvance des F_i par rapport à V_k, par une ou plusieurs des trois méthodes vues plus haut : déterminant non nul ou/et unicité de la solution pour tous les $F_i = 0$ ou/et construction de base de Lagrange (méthode la plus utile),

(2) le degré de *continuité* d'un élément de U_h, c'est-à-dire le plus grand p tel que $U_h \subseteq \mathscr{C}_I^p$.

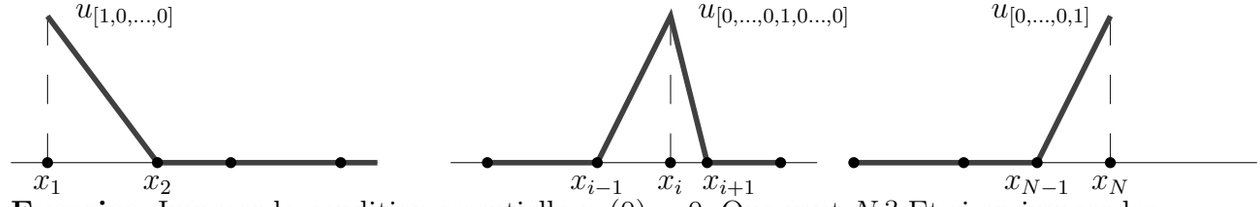
3. Eléments unidimensionnels.

3.1. Interpolation linéaire par morceaux.

On prend $x_1 < \cdots < x_N$, $e_k = [x_k, x_{k+1}], k = 1, \ldots, E = N-1$; $P_i = x_i, F_i(v) = v(x_i), i = 1, \ldots, N$; $V_k = \mathscr{P}_1$, $k = 1, \ldots, N-1$.
On trouve $Q_k = \{k, k+1\}, k = 1, \ldots, N-1$. $\{F_k, F_{k+1}\}$ est évidemment V_k-unisolvant.
$u_{\mathbf{c}}$ est la fonction linéaire par morceaux prenant la valeur c_i en x_i.
Notez que toutes les fonctions de U_h sont continues sur $\overline{\Omega} = [x_1, x_N] : U_h \subseteq \mathscr{C}_I^1$.

Base de Lagrange : $u_{[0,0,\ldots,0,1,0,\ldots,0]}$ est nulle en dehors de son *support* $[x_{i-1}, x_{i+1}]$. Elle vaut $(x - x_{i-1})/(x_i - x_{i-1})$ sur $[x_{i-1}, x_i]$; $(x_{i+1} - x)/(x_{i+1} - x_i)$ sur $[x_i, x_{i+1}]$. On vérifie

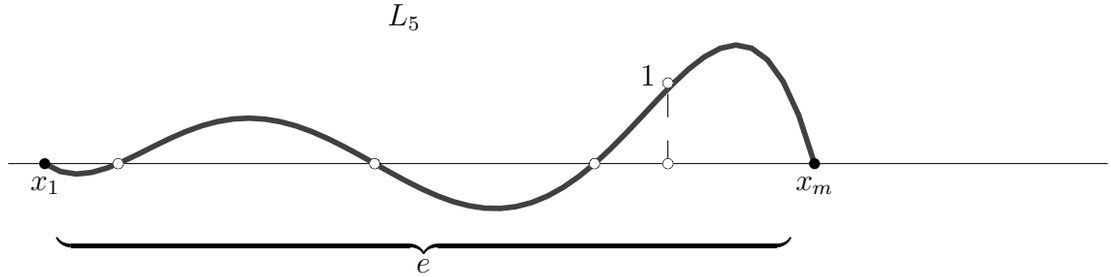
bien le raccord en x_{i-1}, x_i et x_{i+1}. N.B. Si $i = 1$ ou $i = N$, le support se réduit à un seul intervalle.



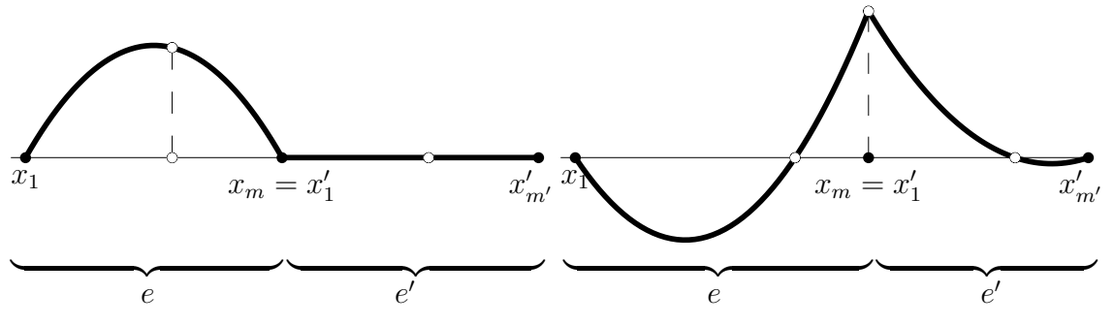
Exercice. Imposer la condition essentielle $u_{\mathbf{c}}(0) = 0$. Que vaut N ? Et si on impose les deux conditions essentielles $u_{\mathbf{c}}(0) = u_{\mathbf{c}}(1) = 0$?

3.2. Interpolation polynomiale de Lagrange.

On peut atteindre une approximation plus précise (= une division de $\overline{\Omega}$ en des éléments plus grands suffira à atteindre une précision demandée) en se donnant $x_1, \ldots, x_m$ distincts *dans un même élément* e ; $F_i(v) = v(x_i), i = 1, \ldots, m$, $V = \mathscr{P}_{m-1}$.
Avec la base $w_j(x) = x^{j-1}, j = 1, \ldots, m$, $\mathbf{M} = [F_i(w_j) = x_i^{j-1}]$, matrice de Vandermonde...
Par le système homogène : tous les $F_i(v) = 0 \Rightarrow v$ doit s'annuler en chaque $x_i \Rightarrow$
$v(x) = K(x - x_1)\ldots(x - x_m) \Rightarrow v$ serait dans $\mathscr{P}_m \setminus V$ si $K \neq 0 \Rightarrow v = 0$.
Base de Lagrange : $L_j(x) = \prod_{\substack{\ell=1 \\ \ell \neq j}}^{m} \dfrac{x - x_\ell}{x_j - x_\ell}$ vaut bien 1 en x_j, 0 aux autres points.



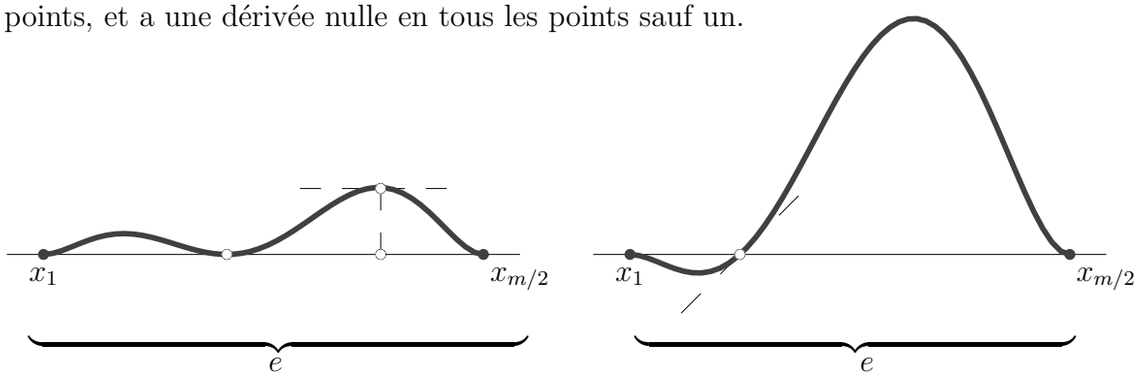
Raccord avec les éléments voisins : si x_1 et x_m sont les extrémités de e, $L_2, \ldots, L_{m-1}$ ont un support réduit au seul élément e : ces fonctions se prolongent par la fonction nulle en dehors de e. Seules L_1 et L_m doivent être prolongées par des fonctions non nulles sur un élément voisin de e (L_1 est prolongée à gauche par le L_m de l'élément précédant e ; L_m est prolongée à droite par le L_1 de l'élément suivant e). On ne dépasse pas $\mathscr{C}^1_I$.



3.3. Interpolation polynomiale d'Hermite.

On donne m pair, $x_1, \dots, x_{m/2}$ distincts dans l'intervalle e, $F_{2i-1}(v) = v(x_i), F_{2i}(v) = v'(x_i), i = 1, \dots, m/2$, $V = \mathscr{P}_{m-1}$.
Par le système homogène : tous les $F_i(v) = 0 \Rightarrow v$ doit avoir un zéro *double* en chaque x_i
$\Rightarrow$
$v(x) = K(x - x_1)^2 \dots (x - x_{m/2})^2 \Rightarrow v$ serait dans $\mathscr{P}_m \setminus V$ si $K \neq 0 \Rightarrow v = 0$.

Base de Lagrange[2] : une moitié des éléments de la base de Lagrange s'annule en tous les points sauf un, et a une dérivée nulle en tous les points ; l'autre moitié s'annule en tous les points, et a une dérivée nulle en tous les points sauf un.



En général, on prend $\{F_i(v)\}$ = réunion de suites *consécutives* de dérivées de v en des points de e, en partant de la valeur de v. L'espace V est $\mathscr{P}_{m-1}$, m étant le nombre total de formes F_i retenues.

Si les dérivées vont jusque l'ordre p aux deux extrémités de l'intervalle e, toutes les fonctions de U_h se raccordent jusqu'à l'ordre $p \Rightarrow U_h \subset \mathscr{C}_I^{p+1}$.

Un exemple plus simple : $m = 3$, $F_1(v) = v(a)$, $F_2(v) = v(b)$, $F_3(v) = v'(b)$ $(a \neq b)$; $W = \mathscr{P}_2$. Vérifiez que $\det \mathbf{M} \neq 0$ avec la base de votre choix.

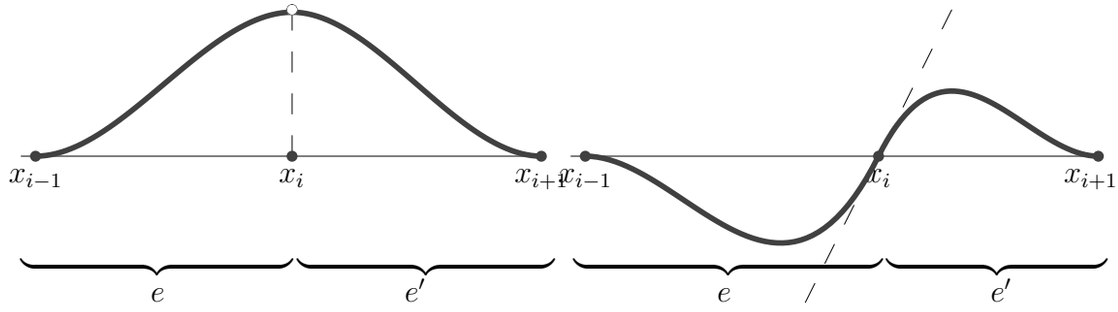
Un exemple d'ensemble non unisolvant : $m = 5$, $F_1(v) = v(a)$, $F_2(v) = v'(a)$, $F_3(v) = v(b)$, $F_4(v) = v'(b)$, $F_5(v) = v'((a+b)/2)$, $\quad (a \neq b)$; $W = \mathscr{P}_4$.
Montrez que $\det \mathbf{M} = 0$ ou qu'il existe des $v \in \mathscr{P}_4$, $v \neq 0$, tels que $F_i(v) = 0, i = 1, \dots, 5$.

3.4. Interpolation cubique d'Hermite par morceaux.

On prend N pair, des points $x_1 < \dots < x_{N/2+1}$, $e_k = [x_k, x_{k+1}], k = 1, \dots, E = N/2$;
$P_1 = x_1$, $F_1(v) = v(x_1)$,
$P_{2i} = x_{i+1}$, $F_{2i}(v) = v(x_{i+1}), i = 1, \dots, N/2 - 1$,
$P_{2i+1} = P_{2i} = x_{i+1}$, $F_{2i+1}(v) = v'(x_{i+1}), i = 1, \dots, N/2 - 1$,
$P_N = x_{N/2+1}$, $F_N(v) = v(x_{N/2+1})$;
$V_1 = V_{N/2} = \mathscr{P}_2$, $V_k = \mathscr{P}_3, k = 2, \dots, N/2 - 1$.
On a $Q_1 = \{1, 2, 3\}$, $Q_k = \{2k-2, 2k-1, 2k, 2k+1\}, k = 2, \dots, N/2 - 1$, $Q_{N/2} = \{N-2, N-1, N\}$.
$u_{\mathbf{c}}$ est la fonction de degré $\leqslant 2$ sur $[x_1, x_2]$ et $[x_{N/2}, x_{N/2+1}]$, de degré $\leqslant 3$ sur $[x_k, x_{k+1}], k =$

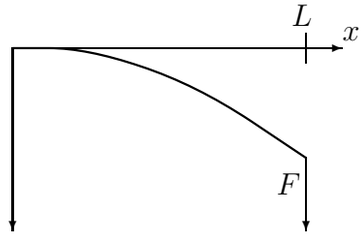
[2] Cf. le cours d'analyse numérique 1a MATH2171.

$2, \ldots, N/2-1$, qui prend les valeurs $c_1, c_2, c_4, \ldots, c_{N-2}, c_N$ en $x_1, x_2, x_3, \ldots, x_{N/2}, x_{N/2+1}$, et dont la dérivée prend les valeurs $c_3, c_5, \ldots, c_{N-1}$ en $x_2, x_3, \ldots, x_{N/2}$. Cette fonction est continûment dérivable sur $\overline{\Omega} = [x_1, x_{N/2+1}]$.



Et le spline cubique d'interpolation ?
Le spline cubique d'interpolation est encore de degré $\leqslant 3$ par intervalle, prend des valeurs arbitraires aux extrémités de ces intervalles, mais des dérivées telles que cette fonction soit deux fois continûment dérivable. Ce n'est pas un élément fini. D'ailleurs, cette fonction n'est déterminée qu'en considérant globalement tous les e_k.

Un exemple (unidimensionnel) de problème d'ordre 4 :



Poutre encastrée fléchie par une force F appliquée à son extrémité :

$$\int_0^L u''(x)v''(x)\,dx = \frac{F}{EI}v(L)\,,$$
$$u(0) = u'(0) = v(0) = v'(0) = 0,$$
$$\iff \min J(u) = \frac{1}{2}\int_0^L (u''(x))^2\,dx - \frac{F}{EI}u(L).$$

Premier essai : $U_h = \{\alpha x^2,\ 0 \leqslant x \leqslant L\}$, donc de dimension 1. Ritz avec $v(x) = x^2$: $4\alpha L = FL^2/(EI)$, donc $\alpha = FL/(4EI)$, $\qquad J = -F^2L^3/(8E^2I^2)$.

Éléments finis avec $e_1 = [0, L/2]$, $e_2 = [L/2, L]$, $V_1 = \{\alpha x^2 + \beta x^3\}$, $V_2 = \mathscr{P}_2$ (donc, de dimensions 2 et 3), $P_1 = P_2 = L/2, P_3 = L$, $F_1(f) = f(L/2), F_2(f) = f'(L/2), F_3(f) = f(L)$.

On vérifie que $\{F_1, F_2\}$ est bien V_1-unisolvant, que $\{F_1, F_2, F_3\}$ est bien V_2-unisolvant. D'ailleurs, on arrive bien à construire les bases de Lagrange $\{\ell_1, \ell_2\}$ de V_1 :

$\ell_1(x) = \dfrac{-3Lx^2/4 + x^3}{-L^3/16}$ vérifie bien $F_1(\ell_1) = \ell_1(L/2) = 1$ et $F_2(\ell_1) = \ell_1'(L/2) = 0$;

$\ell_2(x) = \dfrac{-Lx^2/2 + x^3}{L^2/4}$ vérifie bien $F_1(\ell_2) = \ell_2(L/2) = 0$ et $F_2(\ell_2) = \ell_2'(L/2) = 1$,

et $\{m_1, m_2, m_3\}$ de V_2 :

$m_1(x) = \dfrac{x(L-x)}{L^2/4}$ vérifie bien $F_1(m_1) = m_1(L/2) = 1$, $F_2(m_1) = m_1'(L/2) = 0$ et $F_3(m_1) = m_1(L) = 0$;

$m_2(x) = \dfrac{(x-L/2)(L-x)}{L/2}$ vérifie bien $F_1(m_2) = m_2(L/2) = 0$, $F_2(m_2) = m_2'(L/2) = 1$ et $F_3(m_2) = m_2(L) = 0$;
$m_2(x) = \dfrac{(x-L/2)^2}{L^2/4}$ vérifie bien $F_1(m_3) = m_3(L/2) = 0$, $F_2(m_3) = m_3'(L/2) = 0$ et $F_3(m_3) = m_3(L) = 1$,

d'où on tire la base de Lagrange $\{u_{[1,0,0]}, u_{[0,1,0]}, u_{[0,0,1]}\}$ de U_h (voir figure).

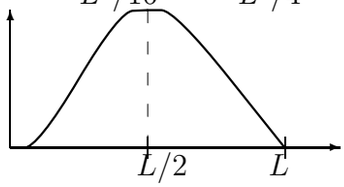
Ritz pour ce U_h : on cherche $u_h = Au_{[1,0,0]} + Bu_{[0,1,0]} + Cu_{[0,0,1]}$ telle que $a(u_h, v) = \varphi(v)$, pour $v =$ les 3 éléments de la base de Lagrange de U_h, d'où les 3 équations

$$\begin{bmatrix} 128/L^3 & -8/L^2 & -32/L^3 \\ -8/L^2 & 16/L & -16/L^2 \\ -32/L^3 & -16/L^2 & 32/L^3 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ F/EI \end{bmatrix},$$

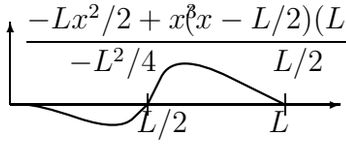
ce qui donne

$$C = u_h(L) = \frac{31}{96}\frac{FL^3}{EI} \ , J(u_h) = -\varphi(u_h)/2 = -\frac{31}{192}\frac{F^2L^3}{E^2I^2} \ .$$

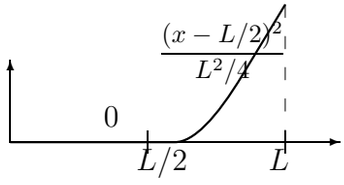
$\dfrac{-3Lx^2/4 + x^3}{-L^3/16}$ $\dfrac{x(L-x)}{L^2/4}$



$\dfrac{-Lx^2/2 + x^3}{-L^2/4}$ $\dfrac{(x-L/2)(L-x)}{L/2}$



$\dfrac{(x-L/2)^2}{L^2/4}$




```
diary
% testfem2.m : test de int de 0 a L de u'' v'' = F v(L)
%
clear;n=input('nombre de subdivisions '),
xi=symmul('L',symdiv((1 :n),n)); xxi=symsub('x',xi);
%
%base de Lagrange :
% le (2i-1)eme element est relatif a la VALEUR au ieme point,
% le (2i)eme element est relatif a la DERIVEE au ieme point.
%
% Lg = restrictions au sous-intervalle de gauche
% Ld = restrictions au sous-intervalle de droite
Lg=zeros(1,2*n-1);Lg=sym(Lg); Ld=zeros(1,2*n-1);Ld=sym(Ld);
% Lg(2i-1)=(1+n(x-xi)/L)^2(1-2n(x-xi)/L) ,

   % Ld(2i-1)=(1-n(x-xi)/L)^2(1+2n(x-xi)/L) ,
% Lg(2i)=(1+n(x-xi)/L)^2 (x-xi) , Ld(2*i)=(1-n(x-xi)/L)^2 (x-xi)
.
for j=1 :n-1,
aux=sym(xxi,1,j);
aux1=symop(1,'+',n,'*',aux,'/','L');aux1=sympow(aux1,2);
aux2=symmul(aux1,aux); Lg=sym(Lg,1,2*j,aux2);
aux2=symop(1,'-',2,'*',n,'*',aux,'/','L');aux1=symmul(aux1,aux2);

 Lg=sym(Lg,1,2*j-1,aux1);
aux1=symop(1,'-',n,'*',aux,'/','L');aux1=sympow(aux1,2);
aux2=symmul(aux1,aux);   Ld=sym(Ld,1,2*j,aux2);
aux2=symop(1,'+',2,'*',n,'*',aux,'/','L');aux1=symmul(aux1,aux2);
 Ld=sym(Ld,1,2*j-1,aux1);
end;
% cas du dernier point
aux=sym(xxi,1,n);
aux1=symop(1,'+',n,'*',aux,'/','L');aux1=sympow(aux1,2);
                          Lg=sym(Lg,1,2*n-1,aux1);
```

```
% a droite de l'avant dernier point:
    aux1=sym(xxi,1,n-1);
    aux2=symop(1,'-',n*n,'*',aux1,'*',aux1,'/','L^2');
                              Ld=sym(Ld,1,2*n-3,aux2);
    aux2=symop(0,'-',n,'*',aux,'*',aux1,'/','L');
                              Ld=sym(Ld,1,2*n-2,aux2);
  Lg=simplify(Lg),       Ld=simplify(Ld),
%
%  matrice de rigidite
%
A=zeros(2*n-1,2*n-1);A=sym(A);
%
for k=1:2*n-1,
    ii=floor((k+1)/2);   % point associe a Lk
     xii0=sym(xi,1,ii);
     if ii>1,xim1=sym(xi,1,ii-1);else,xim1=0;end;
     if ii<n,xip1=sym(xi,1,ii+1);else,xip1=sym(xi,1,n);end;
    auxgk=sym(Lg,1,k);   auxdk=sym(Ld,1,k);
    auxgk2=diff(auxgk,'x',2); auxdk2=diff(auxdk,'x',2);
 for p=1:2*n-1;   % recherche des Lp appropries
   %  ne pas faire comme ca si n est grand!!!
 matel=0;
  % A(k,p)=int de xi(ii-1) a xi(ii+1)  de Lk''  Lp''
    pp=floor((p+1)/2);   % point associe a Lp
     xpp0=sym(xi,1,pp);
     if pp>1,xpm1=sym(xi,1,pp-1);else,xpm1=0;end;
     if pp<n,xpp1=sym(xi,1,pp+1);else,xpp1=sym(xi,1,n);end;
    auxgp=sym(Lg,1,p);   auxdp=sym(Ld,1,p);
    auxgp2=diff(auxgp,'x',2); auxdp2=diff(auxdp,'x',2);
     if xpp1==xii0,
      matel=symadd(matel,int( symmul(auxgk2,auxdp2),'x',xim1,xii0) );
     end;
     if xpp0==xii0,
      matel=symadd(matel,int( symmul(auxgk2,auxgp2),'x',xim1,xii0) );
      matel=symadd(matel,int( symmul(auxdk2,auxdp2),'x',xii0,xip1) );
     end;
     if xpm1==xii0,
      matel=symadd(matel,int( symmul(auxdk2,auxgp2),'x',xii0,xip1) );
     end;
   A=sym(A,k,p,matel);
 end;
end;
A,
% second membre
phi=zeros(1,2*n-1);phi=sym(phi);phi=sym(phi,1,2*n-1,'F/(EI)');
% solution
u=symdiv(phi,A),
J=symop('-1/2','*',sym(u,1,2*n-1)),
```

```
testfem2

nombre de subdivisions 2
n =     2
```

```
Lg = [-4*x^2*(-3*L+4*x)/L^3, 2*x^2*(2*x-L)/L^2, (2*x-L)^2/L^2]

Ld = [-4*x*(x-L)/L^2,        -(x-L)*(2*x-L)/L , 0]


   [128/L^3, -8/L^2,-32/L^3]
A= [ -8/L^2,   16/L,-16/L^2]
   [-32/L^3,-16/L^2, 32/L^3]

u = [5/48*F/EI*L^3, 3/8*F/EI*L^2, 31/96*F/EI*L^3]

J = -31/192*F/EI*L^3

testfem2

nombre de subdivisions 3
n =       3


Lg = [-27*x^2*(2*x-L)/L^3, 3*x^2*(-L+3*x)/L^2, -(-L+3*x)^2*(-5*L+6*x)/L^3,
                           1/3*(-L+3*x)^2*(3*x-2*L)/L^2, (3*x-2*L)^2/L^2]

Ld = [(3*x-2*L)^2*(-L+6*x)/L^3, 1/3*(3*x-2*L)^2*(-L+3*x)/L^2,
                          -3*(L^2+3*x^2-4*x*L)/L^2, -(x-L)*(3*x-2*L)/L, 0]

  [ 648/L^3,      0,-324/L^3, 54/L^2,       0]
  [       0,   24/L, -54/L^2,    6/L,       0]
A=[-324/L^3,-54/L^2, 432/L^3,-18/L^2,-108/L^3]
  [  54/L^2,    6/L, -18/L^2,   24/L, -36/L^2]
  [       0,      0,-108/L^3,-36/L^2, 108/L^3]


u =  [4/81*F/EI*L^3, 5/18*F/EI*L^2, 14/81*F/EI*L^3, 4/9*F/EI*L^2, 107/324*F/EI*L^3]


J =  -107/648*F/EI*L^3



testfem2

nombre de subdivisions 5
n =       5
...

  [ 3000/L^3,       0,-1500/L^3, 150/L^2,        0,       0,        0,       0,       0]
  [       0,    40/L, -150/L^2,    10/L,        0,       0,        0,       0,       0]
  [-1500/L^3,-150/L^2, 3000/L^3,       0,-1500/L^3, 150/L^2,        0,       0,       0]
  [  150/L^2,    10/L,        0,    40/L, -150/L^2,    10/L,        0,       0,       0]
A=[       0,       0,-1500/L^3,-150/L^2, 3000/L^3,       0,-1500/L^3, 150/L^2,       0]
  [       0,       0,  150/L^2,    10/L,        0,    40/L, -150/L^2,    10/L,       0]
  [       0,       0,        0,       0,-1500/L^3,-150/L^2, 2000/L^3, -50/L^2,-500/L^3]
  [       0,       0,        0,       0,  150/L^2,    10/L,  -50/L^2,    40/L,-100/L^2]
  [       0,       0,        0,       0,        0,       0, -500/L^3,-100/L^2, 500/L^3]
```

```
u =  [7/375*F/EI*L^3, 9/50*F/EI*L^2, ...    , 499/1500*F/EI*L^3]


J =  -499/3000*F/EI*L^3

inverse(A)

ans =

[ 1/375*L^3, 1/50*L^2,  1/150*L^3, 1/50*L^2,  4/375*L^3,  1/50*L^2, 11/750*L^3,  1/50*L^2,    7/375*L^3]
[  1/50*L^2,    1/5*L,   3/50*L^2,    1/5*L,   1/10*L^2,     1/5*L,   7/50*L^2,     1/5*L,     9/50*L^2]
[ 1/150*L^3, 3/50*L^2,  8/375*L^3, 2/25*L^2, 14/375*L^3,  2/25*L^2,   4/75*L^3,  2/25*L^2,   26/375*L^3]
[  1/50*L^2,    1/5*L,   2/25*L^2,    2/5*L,   4/25*L^2,     2/5*L,   6/25*L^2,     2/5*L,     8/25*L^2]
[ 4/375*L^3, 1/10*L^2, 14/375*L^3, 4/25*L^2,  9/125*L^3,  9/50*L^2, 27/250*L^3,  9/50*L^2,   18/125*L^3]
[  1/50*L^2,    1/5*L,   2/25*L^2,    2/5*L,   9/50*L^2,     3/5*L,   3/10*L^2,     3/5*L,    21/50*L^2]
[11/750*L^3, 7/50*L^2,   4/75*L^3, 6/25*L^2, 27/250*L^3,  3/10*L^2, 64/375*L^3,  8/25*L^2,   88/375*L^3]
[  1/50*L^2,    1/5*L,   2/25*L^2,    2/5*L,   9/50*L^2,     3/5*L,   8/25*L^2,     4/5*L,    12/25*L^2]
[ 7/375*L^3, 9/50*L^2, 26/375*L^3, 8/25*L^2, 18/125*L^3, 21/50*L^2, 88/375*L^3, 12/25*L^2, 499/1500*L^3]


testfem2

nombre de subdivisions 8
n =        8

...

A =

[12288/L^3,       0,-6144/L^3, 384/L^2,        0,       0,        0,       0,        0,       0,        0,       0,
[        0,    64/L, -384/L^2,    16/L,        0,       0,        0,       0,        0,       0,        0,       0,
[-6144/L^3,-384/L^2,12288/L^3,       0,-6144/L^3, 384/L^2,        0,       0,        0,       0,        0,       0,
[  384/L^2,    16/L,        0,    64/L, -384/L^2,    16/L,        0,       0,        0,       0,        0,       0,
[        0,       0,-6144/L^3,-384/L^2,12288/L^3,       0,-6144/L^3, 384/L^2,        0,       0,        0,       0,
[        0,       0,  384/L^2,    16/L,        0,    64/L, -384/L^2,    16/L,        0,       0,        0,       0,
[        0,       0,        0,       0,-6144/L^3,-384/L^2,12288/L^3,       0,-6144/L^3, 384/L^2,        0,       0,
[        0,       0,        0,       0,  384/L^2,    16/L,        0,    64/L, -384/L^2,    16/L,        0,       0,
[        0,       0,        0,       0,        0,       0,-6144/L^3,-384/L^2,12288/L^3,       0,-6144/L^3, 384/L^2,
[        0,       0,        0,       0,        0,       0,  384/L^2,    16/L,        0,    64/L, -384/L^2,    16/L,
[        0,       0,        0,       0,        0,       0,        0,       0,-6144/L^3,-384/L^2,12288/L^3,       0,-6
[        0,       0,        0,       0,        0,       0,        0,       0,  384/L^2,    16/L,        0,    64/L, -
[        0,       0,        0,       0,        0,       0,        0,       0,        0,       0,-6144/L^3,-384/L^2, 8
[        0,       0,        0,       0,        0,       0,        0,       0,        0,       0,  384/L^2,    16/L, -
[        0,       0,        0,       0,        0,       0,        0,       0,        0,       0,        0,       0,-2


u = [23/3072*F/EI*L^3, 15/128*F/EI*L^2, ... ,     2047/6144*F/EI*L^3]

J =  -2047/12288*F/EI*L^3

exit

 1149495 flops.
```

Par ailleurs, la vraie solution est $u(x) = \dfrac{F}{EI}\left(L\dfrac{x^2}{2} - \dfrac{x^3}{6}\right)$, comme on le voit en passant à

$$[u''v']_0^L - \int_0^L u'''(x)v'(x)\,dx = \frac{F}{EI}v(L) = \frac{F}{EI}\int_0^L v'(x)\,dx \Rightarrow -u''' = \frac{F}{EI},$$

avec $u''(L) = 0$ (condition naturelle). On voit que

$$u_(L) = \frac{1}{3}\frac{FL^3}{EI} \ , J(u) = -\varphi(u)/2 = -\frac{1}{6}\frac{F^2L^3}{E^2I^2}\ .$$

Un exemple plus élaboré de traitement d'un problème de 4ème ordre, d'après `beam.m` tiré de `femlab-1d.tar.Z` :

`beam1.m`

```
% beam1.m: probleme  (au'')'' + (bu')' +cu = f      0 <= x  <= 1
% d'apres  beam.m de      femlab-1d.tar.Z         a(x)>0  (ou a=0 et b<0)
%  with boundary conditions
%
%          u = u0,  u' = u1   (clamped)
%                                seul le cas encastre est au point, hum
%  or
%          u'' = 0,  (au'')' = 0  (free)
%
%  at  x = 0  and  x = 1,  respectively.
%
%  using  C1  piecewice qubic elements and mesh size  h.
%
%
%echo off
clear;
fnom=input('nom du fichier donnees (entre apostrophes)  ');
%  formules de a, b,c et f  dans un fichier
% exemple:   beamd1.d
%1
%2.10/(4*14)
%14*(x^3-x)
%14*(3*x^2-1)
%c
%-1
%0
%c
%1
%0
fid=fopen(fnom);
acopy=fscanf(fid,'%s',1);bcopy=fscanf(fid,'%s',1);
ccopy=fscanf(fid,'%s',1);fcopy=fscanf(fid,'%s',1);
bctype=[1 1];
bounda=fscanf(fid,'%s',1);      % c=clamped , f=free
if bounda=='c'; bctype(1)=0; u0(1)=eval(fscanf(fid,'%s',1));
                             u1(1)=eval(fscanf(fid,'%s',1));
elseif bounda=='f'; bctype(1)=1;
end
bounda=fscanf(fid,'%s',1);
if bounda=='c'; bctype(2)=0; u0(2)=eval(fscanf(fid,'%s',1));
                             u1(2)=eval(fscanf(fid,'%s',1));
elseif bounda=='f'; bctype(2)=1;
end
refine=1;       %Boolean used for refinement loop
```

```
first=1;        %Boolean used to indicate if turn in first loop
while refine
 %Clear old data structures (otherwise, selecting next h > previous h
 % will cause an error)
 clear h hh ni nn ndf rhs f fi a b c u ucomp Ducomp DDucomp  xplot;
 disp('Give desired mesh size h')
 h=input('h = ');
 ni=round(1/h);  h = 1/ni; %Modify h according to ni
 disp(['Number of sub-intervals = ',int2str(ni)])
 nn=(ni+1);
 disp(['Number of nodes = ',int2str(nn)])
 ndf=2*nn;
 disp(['Number of degrees of freedom = ',int2str(ndf-2*sum(bctype))])
 %
 %disp('Before you use this progam, enter formulas for the four basis functions and their second derivatives on the u
     b1copy='1-3*x^2+2*x^3';
     b2copy='x-2*x^2+x^3';
     b3copy='1-3*(1-x)^2+2*(1-x)^3';
     b4copy='-(1-x)+2*(1-x)^2-(1-x)^3';
     Db1copy='-6*x+6*x^2';
     Db2copy='1-4*x+3*x^2';
     Db3copy='6*(1-x)-6*(1-x)^2';
     Db4copy='1-4*(1-x)+3*(1-x)^2';
     DDb1copy='-6+12*x';
     DDb2copy='-4+6*x';
     DDb3copy='-6+12*(1-x)';
     DDb4copy='4-6*(1-x)';
 %Set up six-diagonal stiffnes matrix
 %
 gauss=sqrt(3/5);
 g(1)=1/2-gauss/2;
 g(2)=1/2;
 g(3)=1/2+gauss/2;
 w(1)=5/18;
 w(2)=8/18;
 w(3)=5/18;
 %
 %
  diag=zeros(size(1:ndf));
  subdiag=zeros(size(1:ndf-1));
  sub2diag=zeros(size(1:ndf-2));
  sub3diag=zeros(size(1:ndf-3));
  rhs=zeros(size(1:ndf));
  integJ=0;
   %
   %
   %
  for gp=1:3;
  %
  x=g(gp);
  %
  b1=eval(b1copy);
  b2=eval(b2copy);
  b3=eval(b3copy);
  b4=eval(b4copy);
  %
  Db1=eval(Db1copy)/h;
  Db2=eval(Db2copy)/h;
  Db3=eval(Db3copy)/h;
```

```
  Db4=eval(Db4copy)/h;
  %
  DDb1=eval(DDb1copy)/h^2;
  DDb2=eval(DDb2copy)/h^2;
  DDb3=eval(DDb3copy)/h^2;
  DDb4=eval(DDb4copy)/h^2;
  %
  for i=1:ni;
   x=(i-1)*h+g(gp)*h;
   a(i)=eval(acopy);
   b(i)=eval(bcopy);
   c(i)=eval(ccopy);
   f(i)=eval(fcopy);
  end;
  %
  hh=h*w(gp);
  %
  diag(3:2:ndf-1)=diag(3:2:ndf-1)+(a.*DDb3.*DDb3-b.*Db3.*Db3+c.*b3.*b3).*hh;
  diag(1:2:ndf-3)=diag(1:2:ndf-3)+(a.*DDb1.*DDb1-b.*Db1.*Db1+c.*b1.*b1).*hh;
  diag(4:2:ndf)=diag(4:2:ndf)+(a.*DDb4.*DDb4-b.*Db4.*Db4+c.*b4.*b4).*hh;
  diag(2:2:ndf-2)=diag(2:2:ndf-2)+(a.*DDb2.*DDb2-b.*Db2.*Db2+c.*b2.*b2).*hh;
  %
  subdiag(3:2:ndf-1)=subdiag(3:2:ndf-1)+(a.*DDb3.*DDb4-b.*Db3.*Db4+c.*b3.*b4).*hh;
  subdiag(1:2:ndf-3)=subdiag(1:2:ndf-3)+(a.*DDb1.*DDb2-b.*Db1.*Db2+c.*b3.*b4).*hh;
  subdiag(2:2:ndf-2)=subdiag(2:2:ndf-2)+(a.*DDb2.*DDb3-b.*Db2.*Db3+c.*b3.*b4).*hh;
  %
  sub2diag(1:2:ndf-3)=sub2diag(1:2:ndf-3)+(a.*DDb1.*DDb3-b.*Db1.*Db3+c.*b1.*b3).*hh;
  sub2diag(2:2:ndf-2)=sub2diag(2:2:ndf-2)+(a.*DDb2.*DDb4-b.*Db2.*Db4+c.*b2.*b4).*hh;
  %
  sub3diag(1:2:ndf-3)=sub3diag(1:2:ndf-3)+(a.*DDb1.*DDb4-b.*Db1.*Db4+c.*b1.*b4).*hh;
  %
   rhs(3:2:ndf-1)=rhs(3:2:ndf-1)+f.*b3.*hh;
   rhs(1:2:ndf-3)=rhs(1:2:ndf-3)+f.*b1.*hh;
   rhs(4:2:ndf)=rhs(4:2:ndf)+f.*b4.*hh;
   rhs(2:2:ndf-2)=rhs(2:2:ndf-2)+f.*b2.*hh;
  %
  end %gp
  %
 disp('Start of stiffness matrix:');
 [ diag(1)    subdiag(1) sub2diag(1) sub3diag(1) 0 ; ...
   subdiag(1)   diag(2)  subdiag(2)  sub2diag(2) sub3diag(2) ; ...
  sub2diag(1) subdiag(2)   diag(3)   subdiag(3)  sub2diag(3) ; ...
  sub3diag(1) sub2diag(2) subdiag(3)  diag(4)     subdiag(4) ; ...
       0      sub3diag(2) sub2diag(3) subdiag(4)   diag(5)
 ],
  %
  ndff=ndf;
  if bctype(2)==0;
% rhs=rhs-stiff(:,ndf-1)*u0(2)-stiff(:,ndf)*u1(2)*h;
   rhs(ndf-2)=rhs(ndf-2)-subdiag(ndf-2)*u0(2)-sub2diag(ndf-2)*u1(2)*h;
   rhs(ndf-3)=rhs(ndf-3)-sub2diag(ndf-3)*u0(2)-sub3diag(ndf-3)*u1(2)*h;
   rhs(ndf-4)=rhs(ndf-4)-sub3diag(ndf-4)*u0(2);
   diag=diag(1:ndf-2);
   subdiag=subdiag(1:ndf-3);sub2diag=sub2diag(1:ndf-4);sub3diag=sub3diag(1:ndf-5);
   ndff=ndf-2;
  end
  if bctype(1)==0;
% rhs=rhs-stiff(:,1)*u0(1)-stiff(:,2)*u1(1)*h;
   rhs(3)=rhs(3)-sub2diag(1)*u0(1)-subdiag(2)*u1(1)*h;
```

```
   rhs(4)=rhs(4)-sub3diag(1)*u0(1)-sub2diag(2)*u1(1)*h;
   rhs(5)=rhs(5)-sub3diag(2)*u1(1)*h;
  rhs=rhs(3:ndff);
  diag=diag(3:ndff);
  subdiag=subdiag(3:ndff-1);sub2diag=sub2diag(3:ndff-2);sub3diag=sub3diag(3:ndff-3);
  ndff=ndff-2;
  end
  %
 disp('Stiffness matrix after boundary cond.:');
 [ diag(1)    subdiag(1) sub2diag(1) sub3diag(1) 0 ; ...
   subdiag(1)   diag(2)  subdiag(2)  sub2diag(2) sub3diag(2) ; ...
  sub2diag(1) subdiag(2)   diag(3)   subdiag(3)  sub2diag(3) ; ...
  sub3diag(1) sub2diag(2) subdiag(3)  diag(4)     subdiag(4) ; ...
       0      sub3diag(2) sub2diag(3) subdiag(4)   diag(5)
 ],
  %
  %
  %
% Cholesky
%
%  [ d1  s1  ss1  sss1       ] [c1               ][c1 cs2 css3 csss4     ]
%  [ s1  d2   s2   ss2 sss2  ] [cs2  c2          ][   c2  cs3  css4 csss5]
%  [ss1  s2   d3    s3  ss3  ]=[css3 cs3 c3      ][       c3   cs4  css5 ]
%  ...
  endchol=0;
  rhs=[0 0 0 rhs];
  subdiag=[subdiag 0];sub2diag=[sub2diag 0 0];sub3diag=[sub3diag 0 0 0];
  sub1chol=zeros(size(1:ndff+1));  sub2chol=zeros(size(1:ndff+2));
  sub3chol=zeros(size(1:ndff+3));
  for j=1:ndff,
   chol(j)=diag(j)-sub1chol(j)^2-sub2chol(j)^2-sub3chol(j)^2;
   if chol(j)<0, endchol=1;break; end;
   chol(j)=sqrt(chol(j));
   sub1chol(j+1)=(subdiag(j)-sub1chol(j)*sub2chol(j+1)-sub2chol(j)*sub3chol(j+1) )/chol(j);
   sub2chol(j+2)=(sub2diag(j)-sub1chol(j)*sub3chol(j+2) )/chol(j);
   sub3chol(j+3)= sub3diag(j) /chol(j);
   rhs(j+3)=(rhs(j+3)-sub1chol(j)*rhs(j+2)-sub2chol(j)*rhs(j+1)  ...
                               -sub3chol(j)*rhs(j) )/chol(j);
  end
  %
 disp('Start of Cholesky upper factor:');
 [ chol(1)    sub1chol(2) sub2chol(3) sub3chol(4) 0 ; ...
     0          chol(2)  sub1chol(3)  sub2chol(4) sub3chol(5) ; ...
     0           0         chol(3)   sub1chol(4)  sub2chol(5) ; ...
     0           0           0        chol(4)     sub1chol(5) ; ...
     0           0           0           0         chol(5)
 ],
  %
% u=stiff\rhs;
  rhs=rhs(4:ndff+3);
  u(ndff)=rhs(ndff)/chol(ndff);
  u(ndff-1)=(rhs(ndff-1)-sub1chol(ndff)*u(ndff))/chol(ndff-1);
  u(ndff-2)=(rhs(ndff-2)-sub1chol(ndff-1)*u(ndff-1)-sub2chol(ndff)*u(ndff))/chol(ndff-2);
  for j=ndff-3:-1:1,
   u(j)=(rhs(j)-sub1chol(j+1)*u(j+1)-sub2chol(j+2)*u(j+2)-sub3chol(j+3)*u(j+3))/chol(j);
  end
if bctype(2)==0;   u=[u u0(2) u1(2)*h];  end;
if bctype(1)==0;   u=[u0(1) u1(1)*h u ];  end;
  %
```

```
  %  plot
  %
  % Graphics preparations
  %
  for gp=1:3;
  x=g(gp);
  b1=eval(b1copy);
  b2=eval(b2copy);
  b3=eval(b3copy);
  b4=eval(b4copy);
  Db1=eval(Db1copy)/h;
  Db2=eval(Db2copy)/h;
  Db3=eval(Db3copy)/h;
  Db4=eval(Db4copy)/h;
  DDb1=eval(DDb1copy)/h^2;
  DDb2=eval(DDb2copy)/h^2;
  DDb3=eval(DDb3copy)/h^2;
  DDb4=eval(DDb4copy)/h^2;
  ucomp(gp:3:3*ni-3+gp)=u(1:2:2*ni-1)*b1+u(2:2:2*ni)*b2+u(3:2:2*ni+1)*b3+u(4:2:2*ni+2)*b4;
  Ducomp(gp:3:3*ni-3+gp)=u(1:2:2*ni-1)*Db1+u(2:2:2*ni)*Db2+u(3:2:2*ni+1)*Db3+u(4:2:2*ni+2)*Db4;
  DDucomp(gp:3:3*ni-3+gp)=u(1:2:2*ni-1)*DDb1+u(2:2:2*ni)*DDb2+u(3:2:2*ni+1)*DDb3+u(4:2:2*ni+2)*DDb4;
%
   integJ=integJ+sum(a.*DDucomp(gp:3:3*ni-3+gp).^2 ...
            -b.*Ducomp(gp:3:3*ni-3+gp).^2+c.*ucomp(gp:3:3*ni-3+gp))*hh;
  end
ucomp=[u(1) ucomp u(2*ni+1)];
 disp('Valeurs de u en 0.25, 0.5 et 0.75:');
 for j=1:3,
  xj=0.25*j;ixj=floor(xj/h);x=xj/h-ixj;
  b1=eval(b1copy);b2=eval(b2copy);b3=eval(b3copy);b4=eval(b4copy);
  uc(j)=u(2*ixj+1)*b1+u(2*ixj+2)*b2+u(2*ixj+3)*b3+u(2*ixj+4)*b4;
 end;
 format long;uc, format short;
  rn=1:ni;
  rn=rn*h;
  xplot(3:3:3*ni)=rn-ones(size(rn))*g(1)*h;
  xplot(2:3:3*ni-1)=rn-ones(size(rn))*g(2)*h;
  xplot(1:3:3*ni-2)=rn-ones(size(rn))*g(3)*h;
  xplot=[0 xplot 1];
    %Erase old plots?
  figure(gcf)
    if first
      clgwin=1;     %If first time clear window and plot uexact, uh
      first=~first;
    else
      disp('Do you want to clear the graph window?');
      clgwin = (input('answer y or n ','s')=='y');
    end
    if clgwin
  %   if uexactflag
  %     hold off;  plot(xxplot,uplot,':',xplot,ucomp,'-');  hold on;
  %     title(['u(x) = ' uexact ' dotted;     uh solid']);
  %   else
        hold off;  plot(xplot,ucomp,'-');  hold on;
        title('uh');
  %   end
    else
      plot(xplot,ucomp,'-');
    end
```

```
 disp(' Cholesky  min max :');
  format long;  [ min(chol(1:ndff)), max(chol(1:ndff))], format short ;
 disp(' integrale:');
  format long;   integJ, format short ;
 disp('Do you want to refine the mesh?');
 refine=(input('Answer y or n ','s')=='y');
 end; %of refinement while loop
hold  off
disp('If  you want a hardcopy of the plot, enter the command print');
disp('to close all: fclose(''all'')');

-------------------------------------------------------------------

diary beam1.l

beam1

nom du fichier donnees (entre apostrophes)  'beamd2.d' Give desired mesh size h

u''' + u'' +u = 3 exp(x)

h = 0.1 Number of sub-intervals = 10
Number of nodes = 11
Number of degrees of freedom = 22
Start of stiffness matrix:

  1.0e+004 *
    1.1988    0.5999   -1.1988    0.5999         0
    0.5999    0.3999   -0.5999    0.2000         0
   -1.1988   -0.5999    2.3976    0.0000   -1.1988
    0.5999    0.2000    0.0000    0.7997   -0.5999
         0         0   -1.1988   -0.5999    2.3976

Stiffness matrix after boundary cond.:

  1.0e+004 *
    2.3976    0.0000   -1.1988    0.5999         0
    0.0000    0.7997   -0.5999    0.2000         0
   -1.1988   -0.5999    2.3976    0.0000   -1.1988
    0.5999    0.2000    0.0000    0.7997   -0.5999
         0         0   -1.1988   -0.5999    2.3976

Start of Cholesky upper factor:

  154.8421   -0.0001  -77.4207   38.7427         0
         0   89.4278  -67.0821   22.3682         0
         0         0  116.1124   38.7554 -103.2447
         0         0         0   67.0375  -29.8000
         0         0         0         0  111.4835

Valeurs de u en 0.25, 0.5 et 0.75:

uc =    1.28402086518283   1.64871737300429   2.11700053953865

 Cholesky  min max : 1.0e+002 *  [ 0.38814353891161   1.54842093760062]

 integrale:     integJ =       1.43157790352493

Do you want to refine the mesh?
```

```
Answer y or n n If  you want a hardcopy of the plot, enter the command print
to close all: fclose('all')

format long;uc-exp(0.25*(1:3)),format short

ans = 1.0e-005 * [-0.45515049156286  -0.38976958367520   0.05229259727457 ]

beam1

h = 0.01 Number of sub-intervals = 100    Number of nodes = 101      202
Start of stiffness matrix:

  1.0e+007 *
    1.2000    0.6000   -1.2000    0.6000         0
    0.6000    0.4000   -0.6000    0.2000         0
   -1.2000   -0.6000    2.4000    0.0000   -1.2000
    0.6000    0.2000    0.0000    0.8000   -0.6000
         0         0   -1.2000   -0.6000    2.4000
...

Valeurs de u en 0.25, 0.5 et 0.75:
uc =     1.28402537719733   1.64872120069143   2.11699997858784

 Cholesky  min max :  1.0e+003 *  [1.01967588887132   4.89895499136295 ]
 integrale:    integJ =      1.43189805256995

Do you want to refine the mesh?
n
format long;uc-exp(0.25*(1:3)),format short

ans = 1.0e-007 * [ -0.39490408276777  -0.70008693731083  -0.38024830395500]


beam1

nom du fichier donnees (entre apostrophes)  'beamd2.d' Give desired mesh size h
h = 0.005 Number of sub-intervals = 200    Number of nodes = 201       402
Start of stiffness matrix:

  1.0e+008 *
    0.9600    0.4800   -0.9600    0.4800         0
    0.4800    0.3200   -0.4800    0.1600         0
   -0.9600   -0.4800    1.9200    0.0000   -0.9600
    0.4800    0.1600    0.0000    0.6400   -0.4800
         0         0   -0.9600   -0.4800    1.9200
...

Valeurs de u en 0.25, 0.5 et 0.75:
uc =     1.28402539432393   1.64872122936657   2.11699999257843

 Cholesky  min max :   1.0e+004 *  [  0.28560380015633   1.38563891401656]
 integrale:   integJ =      1.43190057189301

Do you want to refine the mesh?
n
format long;uc-exp(0.25*(1:3)),format short

  1.0e-007 *[-0.22363812357540  -0.41333557021517  -0.24034246148830 ]
```

```
beam1

nom du fichier donnees (entre apostrophes)  'beamd2.d' Give desired mesh size h
h = 0.001 Number of sub-intervals = 1000    Number of nodes = 1001   2002
Start of stiffness matrix:

  1.0e+010 *
    1.2000    0.6000   -1.2000    0.6000         0
    0.6000    0.4000   -0.6000    0.2000         0
   -1.2000   -0.6000    2.4000    0.0000   -1.2000
    0.6000    0.2000    0.0000    0.8000   -0.6000
         0         0   -1.2000   -0.6000    2.4000
...
Start of Cholesky upper factor:

  1.0e+005 *
    1.5492    0.0000   -0.7746    0.3873         0
         0    0.8944   -0.6708    0.2236         0
         0         0    1.1619    0.3873   -1.0328
         0         0         0    0.6708   -0.2981
         0         0         0         0    1.1155

Valeurs de u en 0.25, 0.5 et 0.75:
uc =    1.28402714845518   1.64872476751512   2.11700228847090

 Cholesky  min max :   1.0e+005 * [ 0.31684135694133   1.54919326102332 ]
 integrale:  integJ =   1.43191380744285

Do you want to refine the mesh?
n
format long;uc-exp(0.25*(1:3)),format short

1.0e-005 *  [ 0.17317674421147   0.34968149900116   0.22718582233239  ]

exit             1320773 flops.

------------------------------------------------------------

u''' + u'' = 2 exp(x)

beam1
format long;uc-exp(0.25*(1:3)),format short;

h=0.1    1.0e-006 *[-0.33410285626978   0.00059470495195  -0.55106918317449]
h=0.01   1.0e-009 *[-0.08527201167396  -0.14673950943234  -0.06117817363815]
h=0.005  1.0e-009 *[-0.28261637474714  -0.28264346418894   0.09384271137947]
h=0.001  1.0e-005 *[-0.10207032616893  -0.24702783443242  -0.18865660327272]

exit   715102 flops.
```

4. Éléments bidimensionels.

Exemple : fonctions linéaires par morceaux triangulaires (Espace de Courant).

$\overline{\Omega}$ est un polygone du plan, les e_k forment une *triangulation* de $\overline{\Omega}$: en plus des conditions *supra*, si $k \neq \ell$, les triangles e_k et e_ℓ ont en commun, soit le vide, soit un sommet, soit tout un côté. Les P_i sont les sommets des triangles e_k, $F_i(v) = v(P_i)$; $V_k = \mathscr{P}_1 = \{\alpha + \beta x + \gamma y\}$. Q_k est formé des indices des trois sommets de e_k. $u_{\mathbf{c}}$ est la fonction linéaire sur chaque e_k prenant la valeur c_i au sommet P_i. Cette fonction est bien définie sur les côtés des triangles (ce n'est pas difficile à montrer, mais ce n'est pas évident !) et continue sur $\overline{\Omega}$.

Exemple de triangulation, cf. `http ://www.math.ucl.ac.be/~magnus/num2/triangle.html`

Programme réalisé par J.R. Shewchuk, Carnegie Mellon University,
`http ://www.cs.cmu.edu/~quake/triangle.html`

From : Jorge More' ¡more@mcs.anl.gov¿
Date : Sat, 26 Jul 2003 11 :13 :39 -0500

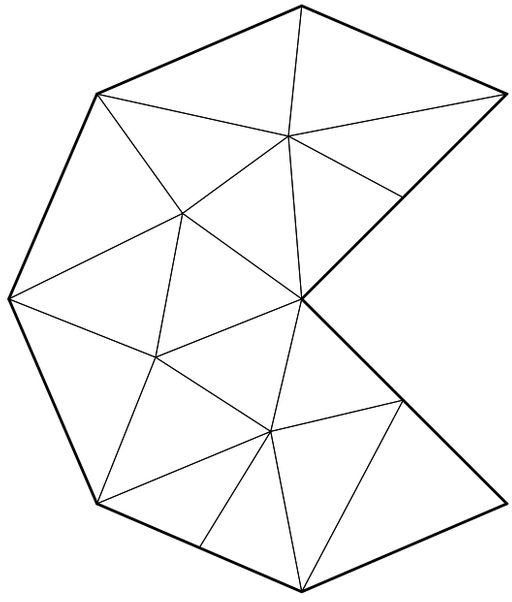
The 2003 Wilkinson Prize for Numerical Software has been awarded to Jonathan Shewchuk for Triangle : A Two-dimensional Mesh Generator and Delaunay Triangulator. The presentation took place July 10, at the 5th International Congress on Industrial and Applied Mathematics (ICIAM 2003) in Sydney, Australia.

Triangle generates high-quality unstructured triangular meshes. Triangle also generates two-dimensional Delaunay triangulations, constrained Delaunay triangulations, Voronoi diagrams, and convex hulls. The speed and accuracy of this code is a result of novel algorithms for extended precision floating-point arithmetic and the use of adaptive computation controlled by forward error analysis.

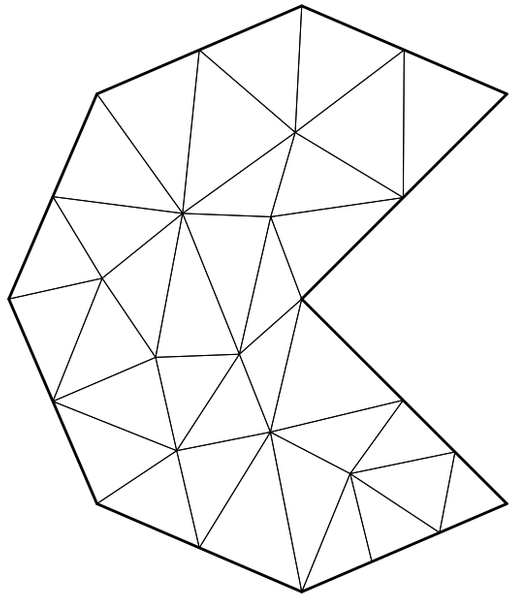
The algorithms and software in Triangle are unquestionably innovative, in both scientific and engineering senses. Triangle includes many significant new algorithmic ideas, including Shewchuk's robust geometric primitives. In an engineering sense, Shewchuk has done a magnificent job of building a flexible piece of software that combines the best of the existing algorithms with his own.

Triangle has thousands of users, and is downloaded more than 30 times per day. Triangle has been licensed for inclusion in eleven commercial software packages.

The Wilkinson Prize for Numerical Software is awarded in honor of the outstanding contributions of James Hardy Wilkinson to the field of numerical software by Argonne National Laboratory, the National Physical Laboratory, and the Numerical Algorithms Group.



/u18/grpanma/magnus @ ux12 [3] %triangle -pqa0.2 trifig1
Opening trifig1.poly. Constructing Delaunay triangulation by divide-and-conquer method. Delaunay milliseconds : 11 Inserting segments into Delaunay triangulation. Segment milliseconds : 1 Removing unwanted triangles. Hole milliseconds : 7 Adding Steiner points to enforce quality. Quality milliseconds : 17
Writing trifig1.1.node. Writing trifig1.1.ele. Writing trifig1.1.poly.
Output milliseconds : 91 Total running milliseconds : 144
Statistics :
Input points : 8 Input segments : 8 Input holes : 0
Mesh points : 15 Mesh triangles : 17 Mesh edges : 31 Mesh boundary edges : 11 Mesh segments : 11



/u18/grpanma/magnus @ ux12 [7] %triangle -pqa0.1 trifig1
Opening trifig1.poly. Constructing Delaunay triangulation by divide-and-conquer method. Delaunay milliseconds : 4 Inserting segments into Delaunay triangulation. Segment milliseconds : 1 Removing unwanted triangles. Hole milliseconds : 1 Adding Steiner points to enforce quality. Quality milliseconds : 8
Writing trifig1.1.node. Writing trifig1.1.ele. Writing trifig1.1.poly.
Output milliseconds : 61 Total running milliseconds : 87
Statistics :
Input points : 8 Input segments : 8 Input holes : 0
Mesh points : 27 Mesh triangles : 34 Mesh edges : 60 Mesh boundary edges : 18 Mesh segments : 18

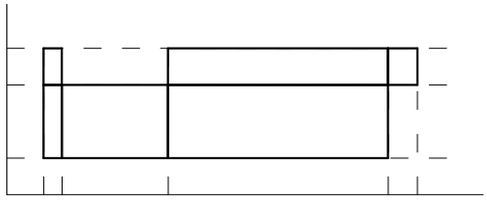
4.1. Éléments rectangulaires.

4.1.1. *Éléments produits.* Une première famille d'éléments bidimensionnels se déduit immédiatement de familles unidimensionnelles par produit :
Proposition. *Si $\{e_k, V_k, F_i\}$ et $\{f_\ell, W_\ell, G_j\}$ sont deux éléments finis, le produit $\{e_k \times f_\ell, V_k \times W_\ell, F_i G_j\}$ est encore un élément fini. Les $F_i G_j$ sont donc $V_k \times W_\ell-$unisolvants. D'ailleurs, si $\{\ell_i\}$ et $\{m_j\}$ sont les bases de Lagrange des deux éléments donnés, $\{p_{i,j}(x,y) = \ell_i(x) m_j(y)\}$ est une base du produit.*

Par $F_i G_j$, il faut comprendre une forme définie sur des fonctions de deux variables par $F_i G_j u(x,y) = F_i$ appliqué en x seul de G_j appliqué en y seul. Ainsi, si $F_i(v) = v(P_i)$ et $G_j(w) = w(Q_j)$, $F_i G_j u(x,y) = u(P_i, Q_j)$; si $F_i(v) = v''(P_i)$ et $G_j(w) = w'''(Q_j)$, $F_i G_j u(x,y) = \dfrac{\partial^5}{\partial x^2 \partial y^3} u(P_i, Q_j)$.

Preuve de la proposition : $(F_i G_j)(p_{r,s}(x,y)) = (F_i G_j)(\ell_r(x) m_s(y)) = F_i\{\ell_r(x)(G_j m_s)\} = (F_i \ell_r)(G_j m_s) = \delta_{i,r}\delta_{j,s} = \delta_{(i,j),(r,s)}$. □

Notez qu'on n'est pas obligé de prendre *tous* les couples possibles,

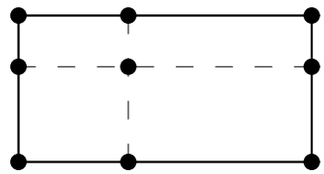


mais que l'on veille à ce que deux éléments aient en commun 1) le vide, 2) un sommet, ou 3) tout un côté.

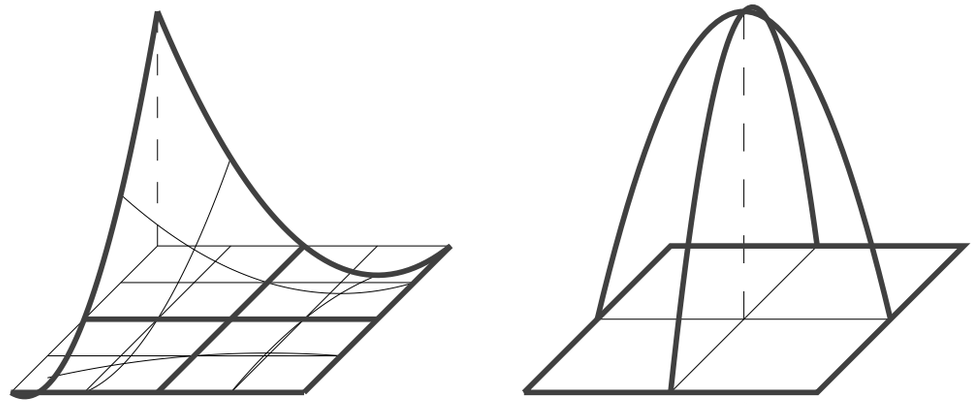
Élément rectangulaire le plus simple (***bilinéaire***) : $F_{i,j}$ = valeurs aux 4 sommets, $V \times W = \mathscr{P}_1 \times \mathscr{P}_1$, de base $\{1, x, y, xy\}$.

Important : toute fonction de U_h est continue dans tout $\overline{\Omega}$, car sa restriction à un côté commun à deux éléments est une *même* fonction du premier degré (en x ou en y) déterminée par les deux *mêmes* valeurs aux extrémités du côté.

Élément rectangulaire produit de deux interpolations de degré 2 en 3 points distincts (***biquadratique***) : $F_{i,j}$ = valeurs aux sommets, et en chaque fois un point d'abscisse et ordonnée intermédiaire (pas nécessairement les milieux des côtés, mais mêmes abscisses et mêmes ordonnées pour ces point intemédiaires) ; $V \times W = \mathscr{P}_2 \times \mathscr{P}_2$, de base $\{1, x, y, x^2, xy, y^2, x^2 y, xy^2, x^2 y^2\}$.
Le principe s'étend aisément (en théorie) aux interpolants de degré quelconque.

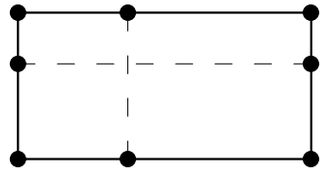


On garde $U_h \subset \mathscr{C}^1_I(\overline{\Omega})$ (sans plus), la restriction d'une fonction de U_h à un côté commun à deux éléments est une *même* fonction du deuxième degré (en x ou en y) déterminée par les trois *mêmes* valeurs sur ce côté.



Deux éléments de la base de Lagrange correspondant 1) à un sommet du rectangle, 2) à un point intérieur.

4.1.2. *L'élément "**serendipity**".* Et si on supprime le point intérieur ? Il reste 8 formes,



il faut imaginer un espace de dimension 8. On choisit l'espace de base $1, x, y, x^2, xy, y^2, x^2y$ et xy^2 (élément ***serendipity***[3]. Ce n'est *pas* un élément produit ! Il faut donc vérifier l'unisolvance des 8 formes d'interpolation selon l'espace proposé.

On peut

(1) Établir que seule la fonction nulle de l'espace annule les 8 formes : si $f(x,y) = \alpha + \beta x + \gamma y + \delta x^2 + \epsilon xy + \phi y^2 + \xi x^2 y + \eta x y^2$ est nulle en $(x,y) = (a_i, b_j)$ pour $i = 1,2,3$, $j = 1,2,3$ SAUF $(i,j) = (2,2)$, on a,
en $y = b_1$: $\alpha + \gamma b_1 + \phi b_1^2 + (\beta + \epsilon b_1 + \eta b_1^2)x + (\delta + \xi b_1)x^2 = 0$ en 3 valeurs de x, d'où
$\alpha + \gamma b_1 + \phi b_1^2 = 0, \beta + \epsilon b_1 + \eta b_1^2 = 0, \delta + \xi b_1 = 0.$
De même, en $y = b_3$:
$\alpha + \gamma b_3 + \phi b_3^2 = 0, \beta + \epsilon b_3 + \eta b_3^2 = 0, \delta + \xi b_3 = 0,$
d'où on tire déjà $\delta = \xi = 0$. On examine maintenant f comme un polynôme de degré 2 en y, en x fixé en a_1 et a_3 :
$\alpha + \beta a_1 + \delta a_1^2 = 0, \gamma + \epsilon a_1 + \xi a_1^2 = 0, \phi + \eta a_1 = 0,$ et
$\alpha + \beta a_3 + \delta a_3^2 = 0, \gamma + \epsilon a_3 + \xi a_3^2 = 0, \phi + \eta a_3 = 0,$
d'où $\phi = \eta = 0$, et, comme nous avons déjà $\xi = 0$: $\gamma = \epsilon = 0$, et enfin, comme on sait que $\delta = 0$, $\alpha = \beta = 0$. □

---

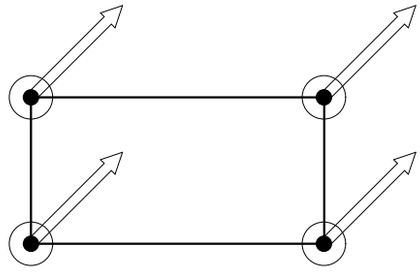
[3] **SERENDIPITY** N. a seeming gift for finding good things accidentally.
Taken by Horace Walpole from *The Three Princes of Serendip*, a 1754 Persian tale.
(The Literary Dictionary (27KB, indexed Oct 19 1997) Thanks to Holly Koslow,
`http ://www.geocities.com/SoHo/Lofts/2297/literary_word.html`)
`http ://www.citadelle.org/fr/forums/lecture.cfm?M_ID=835&Niveau=1`
`http ://sos-net.eu.org/red&s/communic/idl/serendip.htm`

(2) Construire la base de Lagrange de l'élément investigué à partir de la base de Lagrange d'un élément voisin. Ici, partons de $\{\ell_i(x)m_j(y)\}$, i et $j = 1, 2, 3$, base de Lagrange de l'élément produit à 9 points. Retenons les $\ell_i m_j$ pour i et $j = 1, 2, 3$ SAUF $(i, j) = (2, 2)$. On a bien $\ell_i(a_r)m_j(b_s) = \delta_{(i,j),(r,s)}$. Le problème est que les $\ell_i(x)m_j(y)$ ont normalement un coefficient non nul en x^2y^2. Prenons $\ell_i(x)m_j(y) - K_{i,j}\ell_2(x)m_2(y)$ avec $K_{i,j}$ tel que la différence n'ait plus de coefficient en x^2y^2. Il suffit donc de vérifier que ℓ_2 et m_2 sont bien de degré exact 2. Bien sûr : une fonction du premier degré ne saurait prendre les valeurs $0, 1, 0$ en trois points distincts. □

On peut construire des éléments de type serendipity avec plus d'un point intérieur par côté ([Strang & Fix], p. 90).

4.1.3. *Hermite bicubique.* (ou ***rectangle de Bogner-Fox-Schmit***).

Retour à un élément produit : les formes unidimensionnelles sont $F_i(f) = f(a)$ et $f'(a)$ en deux points, il y a donc 4 formes par dimension. Les 16 formes de l'élément produit



sont donc

$$f(a,b), \frac{\partial f}{\partial x}(a,b), \frac{\partial f}{\partial y}(a,b), \text{ et } \frac{\partial^2 f}{\partial x \partial y}(a,b)$$

aux 4 sommets du rectangle. On aura deviné le code graphique[4] : point pour valeur, cercle pour toutes les dérivées premières, d'autres cercles concentriques si on donne toutes les dérivées secondes, etc. Pour les dérivées isolées : flèche orientée simple, double, etc.

Espace : $\mathscr{P}^3 \times \mathscr{P}^3$, soit toutes les combinaisons de $1, x, y, x^2, xy, \ldots, x^3y^3$.

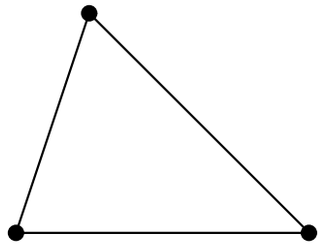
Intérêt principal : $U_h \subset \mathscr{C}^2_I$. Il faut évidemment se concentrer sur ce qui se passe sur les côtés communs à deux éléments. Pour fixer les idées, soit un côté *horizontal* $a \leqslant x \leqslant b, y = c$ commun à e_k et e_ℓ. La restriction de $u_h \in U_h$ à e_k est un certain polynôme de degré 3 en chaque variable, disons $f_k(x, y)$. Sur le côté, soit $g_k(x) := f_k(x, c)$. g_k est un polynôme (d'une variable) de degré $\leqslant 3$, entièrement déterminé par ses valeurs *et les valeurs de sa dérivée première* en $x = a$ et b. En effet, $g'_k = \partial f_k/\partial x$. Comme ces *mêmes* valeurs déterminent également g_ℓ et g'_ℓ, $g_k(x) \equiv g_\ell(x)$: continuité $\mathscr{C}^0$.

Ensuite, il faut montrer que le gradient de u_h est continu, donc que $\mathbf{grad}\, f_k = \mathbf{grad}\, f_\ell$ sur le côté commun. Pas de problème pour la dérivée en x : elle vaut $g'_k = g'_\ell$ sur ce côté. Pour la dérivée en y, soit $h_k(x) := \partial f_k(x, y)/\partial y$ en $y = c$. h_k est un nouveau polynôme de degré 3, fixé aux deux extrémités (valeurs de $\partial f_k/\partial y$), et de dérivées premières fixées en ces deux extrémités (valeurs de $\partial^2 f_k/\partial x \partial y = h'_k(x)$). Les polynômes h_k et h_ℓ sont déterminés par les mêmes données et sont donc identiques : $\partial f_k/\partial y = \partial f_\ell/\partial y$ sur $e_k \cap e_\ell$.

[4] introduit par P. Ciarlet.

4.2. Éléments triangulaires.

4.2.1. *Élément linéaire (Courant).* Nous avons déjà rencontré ***l'élément de Courant***[5]



avec $V_k = \mathscr{P}_1 = \{a + bx + cy\}$ qui ne présente guère difficulté quant à l'unisolvance. Nous pouvons même donner une interprétation du déterminant qui apparaît dans la vérification : $\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}$ n'est autre que deux fois *l'aire* (orientée) du triangle dont les (x_i, y_i) sont les sommets. Preuve supplémentaire qu'il faut et il suffit que ces trois sommets soient non colinéaires.

Cet élément est tellement utilisé qu'il n'est pas inutile de développer quelque peu le sujet :

On appelle *coordonnées barycentriques* d'un point relativement à un triangle les distances aux côtés divisées par les hauteurs correspondantes. Les coordonnées barycentriques d'un point intérieur au triangle sont trois nombres positifs de somme unité. Si $d_i(x, y) = 0$ est l'équation du côté opposé au sommet (x_i, y_i), $i = 1, 2, 3$, la coordonnée barycentrique associée au côté d_i est

$$\ell_i(x, y) = \frac{d_i(x, y)}{d_i(x_i, y_i)},$$

et $\{\ell_1, \ell_2, \ell_3\}$ n'est autre que la base de Lagrange de V !

---

[5] **Richard Courant** Born : 8 Jan 1888 in Lublinitz, Prussia (now Lubliniec, Poland), Died : 27 Jan 1972 in New Rochelle, New York, USA.
Courant obtained his doctorate from Göttingen in 1910 under Hilbert's supervision. He taught mathematics at Göttingen, where he was Klein's successor, until the start of World War I. A few years after the war, Courant returned to Göttingen, where he founded the university's Mathematics Institute. From 1920 until 1933 he was director of the Mathematics Institute.
His most important work was in mathematical physics. In 1924 he published, jointly with Hilbert, an important text *Methoden der mathematischen Physik*. He left for England in 1933, going to New York University the following year. He built up an applied mathematics research centre in New York based on the Göttingen style, making many new appointments such as Friedrichs.
References (6 books/articles)
1. Biography in *Encyclopaedia Britannica*.
2. C Reid, *Courant in Gottingen and New York : the story of an improbable mathematician* (New York, 1976). 3. C Reid, *Hilbert-Courant* (New York, 1986).
4. Professor Richard Courant : A biographical note, *J. Mathematical and Physical Sci.* **7** (1973), i-iv.
5. Richard Courant (1888-1972) (Bulgarian), *Fiz.-Mat. Spis. Bulgar. Akad. Nauk.* **15** (48) (1972), 171.
6. F Williamson Jr., Richard Courant and the finite element method : a further look, *Historia Mathematica* **7** (4) (1980), 369-378.
(`http ://www-history.mcs.st-and.ac.uk/history/Mathematicians/Courant.html`),
voir aussi Courant Institute's Picture Gallery (I-III), `http ://www.cims.nyu.edu/gallery/`

Exercice. Quel sont les lieux de points ayant deux coordonnées barycentriques égales ? Montrez qu'une droite passant par un sommet est caractérisée par un rapport constant de deux coordonnées barycentriques. Et que peut on dire d'une droite parallèle à un des côtés ?

Simplexes dans $\mathbb{R}^n$: cf. [Ciarlet], pp. 45-46.

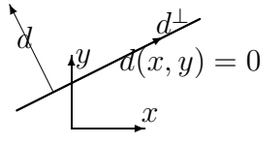
Pour des éléments plus compliqués, les remarques suivantes seront utiles :

(1) L'espace $\mathscr{P}_m$ désigne maintenant les polynômes de degré *total* $\leqslant m$, c'est-à-dire les combinaisons de $x^i y^j$ avec $i + j \leqslant m$.Les bases de monômes et les dimensions sont donc

m	base	dim. $= (m+1)(m+2)/2$
0	1	1
1	1, x, y	3
2	1, x, y, x^2, xy, y^2	6
3	1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3	10
4	1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3, x^4, x^3y, x^2y^2, xy^3, y^4	15
5	1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2, y^3, x^4, x^3y, x^2y^2, xy^3, y^4, x^5, x^4y, x^3y^2, x^2y^3, xy^4, y^5	21

(2) Factorisation : on cherchera encore à établir des propriétés d'unisolvance par $F_i(f) = 0, f \in V \Rightarrow f = 0$. Au passage, il est très utile de pouvoir factoriser f. En fonctions de plusieurs variables, $f = 0$ en un point n'implique pas de factorisation : ce point, comme une infinité d'autres points, fait partie de la courbe d'équation $f(x, y) = 0$, voila tout.
Mais si la courbe $f = 0$ contient des lieux remarquables, on retrouve des factorisations :

(a) **Proposition.** *Si f est un polynôme de deux variables, et si le lieu $f(x, y) = 0$ contient entièrement une droite d'équation $d = 0$, on a $f = dg$, avec g polynomial.*
De plus, si on a également **grad** $f = 0$ *sur la droite $d = 0$, on a $f = d^2h$, avec h polynomial.*



En effet, si $d(x, y) = \alpha x + \beta y + \gamma$, soit $d^\perp(x, y) = -\beta x + \alpha y$. On peut alors exprimer x et y en fonction de d et $d^\perp$ (le déterminant est $\alpha^2 + \beta^2 \neq 0$; on s'est tout simplement placé dans un système d'axes $(d, d^\perp)$ [ce système est d'ailleurs orthogonal]).
On écrit donc $f(x, y) = F(d, d^\perp)$, toujours un polynôme.
Séparons les puissances de d : $f(x, y) = F(d, d^\perp) = F_0(d^\perp) + dF_1(d^\perp) + d^2F_2(d^\perp) + \cdots$
Dès lors, si $f = F = 0$ quand $d = 0$, $F_0(d^\perp) \equiv 0$ et d peut se mettre en évidence ; si, de plus, le gradient de $f = F$ s'annule quand $d = 0$, $\partial F/\partial d = 0$, en $d = 0 \Rightarrow F_1(d^\perp) \equiv 0$ et d^2 apparaît. □

(On a utilisé $\partial F/\partial d = \lim_{\delta\to 0} \dfrac{F(d+\delta, d^\perp) - F(d, d^\perp)}{\delta} =$
$\lim \dfrac{f(x+\alpha\delta/(\alpha^2+\beta^2), y+\beta\delta/(\alpha^2+\beta^2)) - f(x,y)}{\delta} = (\alpha^2+\beta^2)^{-1}[\alpha,\beta]\cdot \mathbf{grad}\, f$).

(b) **Proposition.** *Si f est un polynôme de degré $\leqslant m$ de deux variables, et si la droite d'équation $d(x,y)=0$ contient $m+1$ points, comptés avec leurs multiplicités, où f et d'éventuelles dérivées tangentielles de f s'annulent, on a $f = dg$, avec $g \in \mathscr{P}_{m-1}$.*
Si, de plus, la dérivée normale de f s'annule en m points (comptés avec leurs multiplicités) de la même droite, $f = d^2h$, avec $h \in \mathscr{P}_{m-2}$.
En effet, on passe aux variables $d, d^\perp$ comme au point précédent. Sur $d=0$, $f = F(0, d^\perp)$ est un polynôme d'une variable ayant au moins $m+1$ zéros (comptés avec leurs multiplicité) $\Rightarrow F(0,d^\perp) \equiv 0$ et $f = F = dg$.
On a donc $f = F = dF_1(d^\perp) + d^2F_2(d^\perp) + \cdots$ La dérivée normale $\partial F/\partial d = F_1 + 2dF_2 + \cdots$ est donc le polynôme $F_1 \in \mathscr{P}_{m-1}$ sur $d=0$ et est le polynôme nul s'il s'annule en au moins m points. □

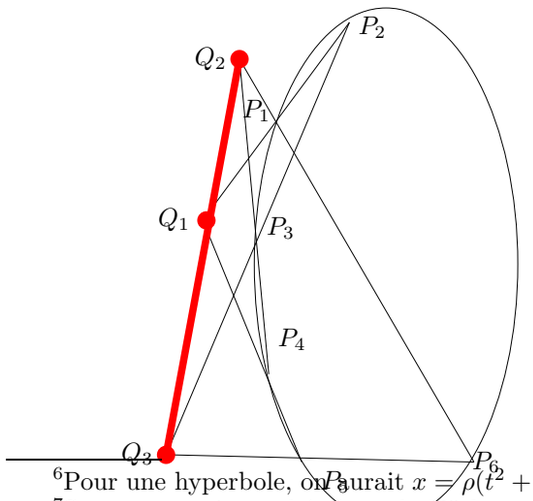
(3) ***Problème.*** Et si la courbe $f=0$ contient des points situés sur une même ***conique*** ?
Si la courbe $f(x,y)=0$, $f \in \mathscr{P}_m$, $m \geqslant 2$, contient $2m+1$ points distincts appartenant à une conique non dégénérée [= non réduite à deux droites] d'équation $C(x,y)=0$, alors $f = Cg$, avec $g \in \mathscr{P}_{m-2}$.
Démonstration (partielle) : supposons pouvoir amener C sur le cercle unité par transformation affine. Alors, on passe à des coordonnées (ρ, t) par

$$x = \rho\frac{t^2-1}{t^2+1}, y = \rho\frac{2t}{t^2+1},$$

(remarquez que $\rho = \sqrt{x^2+y^2}$, il s'agit en fait des coordonnées polaires[6], la conique est le lieu d'équation $\rho^2 - 1 = 0$), que l'on porte dans $f(x,y)$:

$$(t^2+1)^m f(x,y) := F(\rho,t) = F_0(t) + (\rho-1)F_1(t) + \cdots$$

F_0 est un polynôme de degré $\leqslant 2m$, puisque $(t^2+1)^m x^i y^j = (t^2+1)^{m-i-j}(t^2-1)^i(2t)^j \in \mathscr{P}_{2m}$. Ce polynôme devant s'annuler en $2m+1$ points doit donc être le polynôme nul, donc f contient $\rho - 1$ en facteur. On reprend avec les puissances de $\rho+1$ pour obtenir le facteur $\rho+1$, donc $\rho^2 - 1 = (x^2+y^2)^{1/2} - 1 = C(x,y)$[7].



Application : ***Théorème de Pascal***, *les côtés opposés d'un hexagone inscrit à une conique se coupent en trois points colinéaires.*
En effet, soient $d_{1,2}=0$, $d_{2,3}=0,\ldots$, $d_{5,6}=0$ et $d_{6,1}=0$ les équations des côtés de l'hexagone. On considère le polynôme du ***troisième*** degré

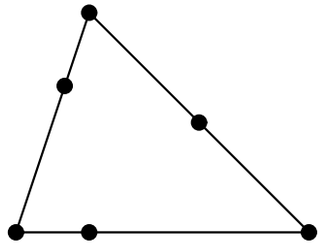
$$f = d_{1,2}d_{3,4}d_{5,6} - \lambda d_{2,3}d_{4,5}d_{6,1}.$$

Ce polynôme s'annule en P_1 (où $d_{1,2}$ et $d_{6,1}=0$), en $P_2,\ldots$, en P_6, ET en Q_1 (où $d_{1,2}$ et $d_{4,5}=0$), Q_2 (où $d_{3,4}$ et $d_{6,1}=0$), et Q_3 (où $d_{5,6}$ et $d_{2,3}=0$). Fixons λ pour que la courbe $f=0$ contienne un septième point de la conique $C=0 \Rightarrow f = Cd$, où d est du premier degré $\Rightarrow Q_1, Q_2$ et Q_3 sont sur $d=0$, donc colinéaires.

[6]Pour une hyperbole, on aurait $x = \rho(t^2+1)/(t^2-1)$.

[7]Sur ces questions, théorème de Nöther, etc., voir, par exemple, E.L. Wachpress, *A Rational Finite Element Basis*, Academic Press, New York, 1975, et des ouvrages sur les courbes planes algébriques.

4.2.2. *Triangle à six points.* On interpole aux trois sommets et en un point intérieur de chaque côté (souvent le point milieu).



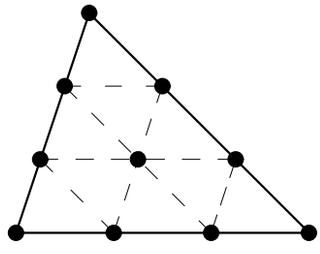
$V = \mathscr{P}_2$ est bien de dimension 6. Unisolvance : si $f \in \mathscr{P}_2$ s'annule aux 6 points, f s'annule aux 3 points du côté $d_1 = 0 \Rightarrow f = d_1 g$. De même pour les deux autres côtés : $f = d_1 d_2 d_3 h$, impossible avec $f \in \mathscr{P}_2$, sauf si $h = 0$.

Autre argument : les 6 points sont sur la conique $f = 0$ et forment donc un hexagone inscrit. Théorème de Pascal : Q_1, Q_2 et Q_3 sont colinéaires, impossible, ce sont les trois sommets !

Base de Lagrange : $\ell_i(x, y)$ vaut 1 au point P_i, vaut 0 aux 5 autres points. Parmi ces 5 points, il y a toujours un triplet colinéaire (sur un côté) $\Rightarrow \ell_i = d_i g_i$, où $d_i = 0$ est l'équation de ce côté, et $g_i = 0$ est l'équation de la droite passant par les deux derniers points.

Continuité : étant bien entendu que le point intérieur est le *même* sur le côté commun de deux triangles jointifs, la restriction d'une fonction de U_h à ce côté est un polynôme du second degré d'une variable (la distance à un sommet) et est donc entièrement déterminée par les valeurs aux trois points situés sur ce côté : $U_h \subset \mathscr{C}^1_I$.

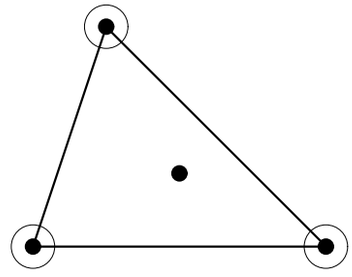
4.2.3. *Autres éléments triangulaires d'interpolation.* On prend les sommets, $m-1$ points intérieurs aux côtés, et $(m-2) + (m-3) + \cdots = (m-2)(m-1)/2$ points intérieurs au triangle, ce qui fait $(m+1)(m+2)/2$ points. On s'arrange pour pour avoir des alignements de m, $m-1$,etc. points, par exemple sur des parallèles aux côtés (voir figure pour $m = 3$).



On prend $V = \mathscr{P}_m$. Unisolvance : directement par base de Lagrange : un élément $\ell_i(x, y)$ de la base de Lagrange doit être un polynôme de $\mathscr{P}_m$ s'annulant en tous les points sauf un. Comme il y a au moins un côté contenant $m + 1$ points où ℓ_i doit s'annuler, $\ell_i = d_i g_i$, où $d_i = 0$ est l'équation de ce côté et $g_i \in \mathscr{P}_{m-1}$. On recommence avec un alignement de m points, etc.

Continuité : $U_h \subset \mathscr{C}^1_I$, puisque u_h sur un côté est un polynôme de degré $\leqslant m$ en une variable, entièrement déterminé par ses valeurs en $m + 1$ points.

4.2.4. *Triangle d'Hermite.* On fixe les valeurs et les deux composantes du gradient aux 3 sommets, et la valeur en un point intérieur.



$V = \mathscr{P}_3$ (dimension 10).

Unisolvance : si $f \in \mathscr{P}_3$ annule toutes les formes, f a deux zéros doubles le long de chaque côté $\Rightarrow f = d_1 d_2 d_3 c$, où c ne peut plus être qu'une constante, et il est alors impossible d'avoir $f = 0$ au point intérieur sauf si $c = 0$. □

Ou par base de Lagrange ([Nicaise], exercice 8.32) : soient $d_1 = 0, d_2 = 0, d_3 = 0$ les équations des 3 côtés, telles que $d_i = 1$ au sommet opposé au côté d'équation $d_i = 0$. (Les d_i forment la base de Lagrange du triangle de Courant).

(1) Le polynôme de $\mathscr{P}_3$ qui vaut 1 au sommet opposé au côté d'équation $d_i = 0$, nul aux deux autres sommets et au point intérieur, et de gradient nul aux trois sommets est $d_i^2(3 - 2d_i) -$ const. $d_1 d_2 d_3$, avec la constante telle que le résultat soit nul au point intérieur.

(2) Les deux polynômes indépendants de $\mathscr{P}_3$ nuls aux trois sommets et au point intérieur, et de gradient nul aux deux sommets situés sur le côté d'équation $d_i = 0$ sont
$\alpha d_i d_j (A d_i + d_j - 1) + \beta d_i d_k (B d_i + d_k - 1)$,
où A et B sont tels que $A d_i + d_j - 1 = B d_i + d_k - 1 = 0$ au point intérieur. En effet, au sommet $d_i = 0, d_j = 1$, (et donc, $d_k = 0$: $d_i d_k$ a déjà son gradient nul), la dérivée $\partial/\partial d_i$ vaut $\alpha d_j (d_j - 1) = 0$; la dérivée $\partial/\partial d_j$ vaut $\alpha[d_i(A d_i + d_j - 1) + d_i d_j = 0$.

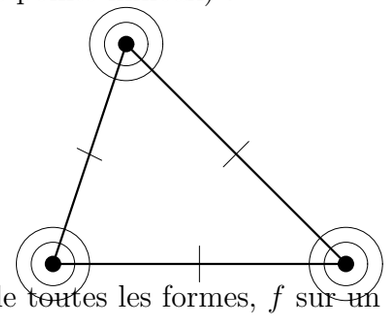
(3) Enfin, le polynôme de $\mathscr{P}_3$ nul avec son gradient aux trois sommets est const. $d_1 d_2 d_3$, avec la constante telle que le résultat $= 1$ au point intérieur.

Continuité : examinons u_h le long d'un côté commun à deux éléments. Effectuons une translation et une rotation pour amener ce côté sur $a \leqslant x \leqslant b, y = 0$. u_h est un polynôme de degré $\leqslant 3$ en x déterminé par ses valeurs et les valeurs de sa dérivée première en a et b (interpolation d'Hermite), donc, continuité $\mathscr{C}^0$: $u_h \in \mathscr{C}^1_I$.

Et la dérivée normale ? C'est-à-dire $\partial u_h/\partial y$, qui est un polynôme du second degré en x sur $[a, b]$, dont on ne fixe que les valeurs (deuxième composante du gradient de u_h) en a et b, ce qui est insuffisant pour le déterminer ! La dérivée normale subira donc le plus souvent une discontinuité lors du passage d'un élément à un élément voisin (aux points intérieurs d'un côté).

Il existe une modification de type serendipity du triangle d'Hermite ne nécessitant pas de point intérieur (triangle de Zienkiewicz).

4.2.5. *Triangle d'Argyris.* On atteint la continuité $\mathscr{C}^1$ en imposant les valeurs, celles des deux dérivées premières ET celles des trois dérivées secondes en chaque sommet[8], ce qui fait 18 conditions. Il n'y a pas de $\mathscr{P}_m$ de dimension 18 ! Le triangle d'Argyris utilise $\mathscr{P}_5$ (dim. 21) et impose 3 formes supplémentaires, les dérivées normales en des points intérieurs des côtés (presque toujours les points milieux) :



Unisolvance : si $f \in \mathscr{P}_5$ annule toutes les formes, f sur un côté est un polynôme de degré $\leqslant 5$ ayant 2 zéros *triples* aux deux sommets $\Rightarrow f = 0$ sur tout le côté ; la dérivée normale de f est un polynôme de degré $\leqslant 4$ le long d'un côté, ayant deux zéros doubles aux sommets ET un zéro en un point intérieur $\Rightarrow f$ et sa dérivée normale sont nulles sur tout un côté $\Rightarrow f = d^2 g$, où $d = 0$ est l'équation de ce côté, et $g \in \mathscr{P}_3$. Pour les trois côtés : $f = d_1^2 d_2^2 d_3^2 h$, impossible sauf si $h = 0$. □

Continuité : portons encore un côté sur $x \in [a, b]$, $y = 0$. Sur $[a, b]$, u_h est de degré $\leqslant 5$, entièrement déterminée par ses valeurs et celles de ses dérivées première et seconde en a et b ; la dérivée normale ($\partial u_h/\partial y$ en $y = 0$) est de degré $\leqslant 4$ (en x), entièrement déterminée par ses valeurs et dérivées premières ($\partial^2 u_h/\partial x \partial y$) en a et b, ET par une valeur en un point intérieur au côté.

Donc, $U_h \subset \mathscr{C}_I^2$. □

En s'arrangeant pour que les éléments de l'espace V aient des dérivées normales de degré $\leqslant 3$ le long des côtés, on arrive à un élément fini à 18 degrés de liberté (*triangle de Bell*).

Il existe bien sûr des éléments tridimensionnels : briques, tétraèdres, etc.

V_k peut lui-même être un espace de polynômes par morceaux (*macro-éléments* : triangles de Hsieh-Clough-Tocher, quadrilatère de Fraeijs de Veubeke-Sander,. . .[Ciarlet], chap. 6).

5. Résumé de ce chapitre.

La résolution de $a(u_h, v) = \varphi(v)$ pour tout v dans un espace U_h d'éléments finis consiste à prendre pour u_h et v des fonctions

(1) dont la restriction à des éléments e_k appartient à un espace prédéfini V_k de dimension faible (souvent un espace de polynômes). Les e_k sont d'intérieurs disjoints et leur réunion est $\overline{\Omega}$ (éventuellement modifié pour permettre un choix simple des éléments e_k),

(2) entièrement déterminées sur tout $\overline{\Omega}$ par les valeurs prises par de nombreuses formes imposées F_i, $i = 1, \ldots, N$. Une fonction u_h de U_h est donc représentée par le vecteur $[F_1(u_h), \ldots, F_N(U_h)]$ (et U_h est de dimension N).
Cependant, les valeurs d'un petit nombre de formes, généralement associées à des

[8]Sur la nécessité de ces conditions, cf. A. Ženíšek, Interpolation polynomials on the triangle, *Numer. Math.* **15** (1970) 283-296 ; A general theorem on triangular finite $\mathscr{C}^{(m)}-$elements, *R.A.I.R.O. Analyse Numérique* **R-2** (1974) 119-127.

points de e_k, doivent suffire à déterminer la restriction de u_h à e_k (unisolvance de $\{F_i\}, i \in Q_k$ par rapport à V_k),

(3) qui sont toutes dans $\mathscr{C}^m_I(\overline{\Omega})$ si $a(u,v)$ fait appel à des dérivées d'ordre $\leqslant m$ de u et v (éléments finis ***conformes***, il existe toute une théorie [en fait, d'abord une pratique, puis une théorie] des éléments finis non conformes).

La *base de Lagrange* de U_h relativement aux formes F_j est formée des N fonctions $\ell_i \in U_h$ telles que $F_j(\ell_i) = \delta_{i,j}$, $i,j = 1,\ldots,N$. Comme la restriction de ℓ_i à chaque e_k est entièrement déterminée par les quelques formes associées à des points de e_k, cette restriction sera la fonction nulle pour la plupart des éléments e_k : les fonctions ℓ_i ont de petits supports (lieux où $\ell_i \neq 0$).

Dans le problème $a(u_h,v) = \varphi(v)$, les inconnues sont les réels α_j tels que $u_h = \sum_{j=1}^{N} \alpha_j \ell_j$, et on applique $v = \ell_i$, $i = 1,\ldots,N$: $\boldsymbol{A}_h\boldsymbol{\alpha} = \boldsymbol{\varphi}$, avec $\boldsymbol{A}_h = [a(\ell_i,\ell_j)]_{i,j=1}^N$ (*matrice de rigidité*), $\boldsymbol{\alpha} = [\alpha_1,\ldots,\alpha_N]^T$, $\boldsymbol{\varphi} = [\varphi(\ell_1),\ldots,\varphi(\ell_N)]^T$.

La matrice $\boldsymbol{A}_h$ est creuse, ce qui permet de traiter des problèmes de très grande taille.

Exemple d'élément fini tridimensionnel non conforme : *brique de Wilson* : e = un pavé de $\mathbb{R}^3$, V de base $\{1,\ x_1, x_2, x_3,\ x_1^2, x_1x_2, x_2^2, x_1x_3, x_2x_3, x_3^2,\ x_1x_2x_3\}$, donc de dimension 11, et $\{F_i\}$ = les valeurs aux 8 sommets et les 3 intégrales $\displaystyle\int\int\int_e \frac{\partial^2 f}{\partial x_i^2}\, dx_1dx_2dx_3$ ([Ciarlet], ch. 4, § 4.2.1).

Crimes variationnels : voir chap. 3, § 5

Chapitre 3

Espaces de Sobolev, convergence.

1. Introduction et définitions.

Nous nous préoccupons de résoudre $a(u,v) = \varphi(v)$ dans un espace

(1) suffisamment grand pour qu'il y ait existence d'une solution,

(2) pas trop grand pour qu'il y ait unicité.

Ainsi, on a vu un exemple p. 37 où, sur un secteur d'ouverture $\alpha\pi$, on devait admettre des fonctions se comportant comme $r^{1/\alpha}$ près de l'origine, et rejeter des fonctions se comportant comme $r^{-1/\alpha}$.

L'existence et l'unicité sont assurées (moyennant vérifications de conditions sur la forme a) dans un *espace de Hilbert* (p. 48). Commençons donc par établir un produit scalaire sur $\mathscr{C}^m_I(\overline{\Omega})$:

1.1. Produit scalaire de Sobolev.

Si Ω est un ouvert borné de $\mathbb{R}^n$,

$$(f,g)_m := \sum_{|\alpha|\leqslant m} \int_\Omega D^\alpha f \, D^\alpha g \, d\boldsymbol{x} \tag{40}$$

est un produit scalaire sur $\mathscr{C}^m_I(\overline{\Omega})$.

On rappelle que α est un vecteur de n entiers $\geqslant 0$, que $|\alpha|$ est la somme des composantes de α, et que $D^\alpha f$ est la dérivée partielle $\dfrac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}$.

Les différents termes de (40) sont donc les intégrales de fg, de $\mathbf{grad}\, f \cdot \mathbf{grad}\, g$, de $\displaystyle\sum_{i=1}^n \frac{\partial^2 f}{\partial x_i^2}$
$+ \displaystyle\sum_{i=1}^n \sum_{j<i} \frac{\partial^2 f}{\partial x_i \partial x_j}$, etc.

Les intégrales sur Ω s'entendent comme une somme d'intégrales sur les parties de Ω (les morceaux) où les fonctions sont m fois continûment dérivables, on ne doit donc intégrer que des fonctions continues sur des bornés.

(40) est bien un produit scalaire sur $\mathscr{C}^m_I(\overline{\Omega})$:
– $(f,g)_m$ est définie pour tout $f, g \in \mathscr{C}^m_I(\overline{\Omega})$,
– $(.,.)_m$ est bien une forme bilinéaire [les scalaires sont les réels], et
– définie positive : $(f,f)_m \geqslant 0$; $(f,f)_m = 0 \Rightarrow \displaystyle\int_\Omega f^2 \, d\boldsymbol{x} = 0 \Rightarrow f = 0$ sur chaque partie de Ω où f est continue $\Rightarrow f = 0$.

Exemples.
$(|x|,|x|)_1$ sur $(-a,b)$ vaut $(a^3+b^3)/3+a+b>0$ si a et $b\geqslant 0$, $a+b>0$.
$(|x|^\alpha,|x|^\alpha)_1$ sur $(-R,R)$ vaut $2R^{2\alpha+1}/(2\alpha+1)+2\alpha^2R^{2\alpha-1}/(2\alpha-1)$ a encore un sens dès que $\alpha>1/2$, pourtant $|x|^\alpha$ n'est plus dans $\mathscr{C}^1$ quand $\alpha<1$.
$(r^\alpha,r^\alpha)_1$ sur le disque $r<R$ vaut $2\pi R^{2\alpha+2}/(2\alpha+2)+2\pi\alpha^2R^{2\alpha}/(2\alpha)$ a encore un sens dès que $\alpha>0$, pourtant r^α n'est plus dans $\mathscr{C}^1$ quand $\alpha<1$.

Si Ω est non borné, on peut encore définir $(f,g)_m$ si f et g sont de support borné (support d'une fonction continue f = adhérence du lieu où $f\neq 0$).

Avec la norme

$$\|f\|_m := \sqrt{(f,f)_m} = \left[\sum_{|\alpha|\leqslant m}\int_\Omega (D^\alpha f)^2\ d\boldsymbol{x}\right]^{1/2} \tag{41}$$

on munit donc $\mathscr{C}^m_I(\overline{\Omega})$ d'une structure d'espace *préhilbertien*.

Remarque. Si f est la constante C sur tout Ω borné, $\|f\|_m=\|f\|_0=|C|\sqrt{\mu(\Omega)}$, où $\mu(\Omega)$ est la mesure (de Lebesgue) de Ω.

1.2. Espaces de Sobolev.

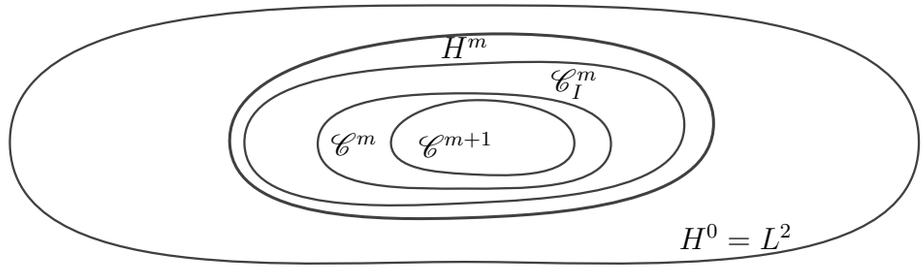
1.2.1. ***Définition***. *L'espace de Sobolev*[1] $H^m(\Omega)$ *est le complété de l'espace préhilbertien* $\mathscr{C}^m_I(\overline{\Omega})$ *selon la norme* (41).
On entend par là qu'un élément de $H^m(\Omega)$ doit toujours pouvoir être associé à une suite de Cauchy de $\mathscr{C}^m_I(\overline{\Omega})$, c'est-à-dire une suite $\{f_N\}$ telle que $\forall\varepsilon>0, \|f_N-f_M\|_m\leqslant\varepsilon$ si N et M sont assez grands.

1.2.2. *Identification à un espace de fonctions.* L'espace $H^m(\Omega)$ est donc un espace d'éléments associés à des suites de Cauchy selon la norme $\|\ \|_m$. Comme $\|f\|_m\geqslant\|f\|_0$, $H_m(\Omega)$ est contenu dans un espace plus familier, le complété de $\mathscr{C}$ selon l'intégrale du carré de fonctions. On sait que ce dernier espace est l'espace $L^2(\Omega)$ des classes d'équivalence de fonctions de carré intégrable sur Ω, deux fonctions de $L^2(\Omega)$ étant considérées comme identiques si elles coïncident *presque partout*[2]. Avec cette condition en tête, nous notons donc

$$H^m(\Omega)\subseteq H^0(\Omega)=L^2(\Omega).$$

Bien entendu, on suivra l'usage en parlant de L^2, et donc de H^m, comme d'un *espace de fonctions*.

[1] S.L. Sobolev (1908-1989).
[2] J. Mawhin, *Analyse. Fondements, techniques, évolution*, De Boeck, 2ème édition, 1997, p. 684.



Exemple. Constater que $|x|^{2/3}$ est bien dans $H^1((-R,R))$: on construit $f_n(x) = n^{1/3}|x|$ sur $(-n^{-1}, n^{-1})$; $f_n(x) = |x|^{2/3}$ sur $|x| > 1/n$. Cette fonction f_n est bien dans $\mathscr{C}^1_I((-R,R))$. Ensuite, si $n > m$, $f_n(x) - f_m(x)$ vaut $(n^{1/3} - m^{1/3})|x|$ sur $(-n^{-1}, n^{-1})$; $|x|2/3 - m^{1/3}|x|$ sur $(n^{-1} < |x| < m^{-1})$; 0 sur $m^{-1} < |x| < R$. Enfin,

$$\|f_n - f_m\|_1^2 = \frac{2}{3}(n^{2/3}-m^{2/3})^2 n^{-3} + 2(n^{2/3}-m^{2/3})^2 n^{-1} + \frac{6}{7}(m^{-7/3}-n^{-7/3}) - \frac{3}{2}m^{1/3}(m^{-8/3}-n^{-8/3}) + \frac{2}{3}m^{2/3}(m^{-3}-n^{-3})$$
$$+ \frac{8}{3}(m^{-1/3} - n^{-1/3}) - 4m^{1/3}(m^{-2/3} - n^{-2/3}) - 2m^{2/3}(m^{-1} - n^{-1}) < \text{const. } m^{-1/3}$$

aussi petit que l'on veut si m est assez grand.

On n'exhibe jamais de suite de Cauchy pour vérifier si $f \in H^m$! On vérifiera si une fonction de $L^2(\Omega)$ admet des *dérivées distributionnelles* (au sens (43), voir plus loin, p. 95) de carré intégrable jusqu'à l'ordre m.

On définit aussi l'espace $H^{m,p}(\Omega)$, $1 \leqslant p \leqslant \infty$, comme complété de $\mathscr{C}^m_I(\overline{\Omega})$ selon la norme

$$\|u\|_{m,p} = \left\{ \sum_{0 \leqslant |\alpha| \leqslant m} \|D^\alpha u\|_p^p \right\}^{1/p}.$$

2. Propriétés des espaces de Sobolev.

2.1. Formes définies sur $H^m(\Omega)$.

Une certaine familiarité avec L^2 nous habitue à quelques précautions dans l'examen d'opérations usuelles de l'analyse. Ainsi, on veillera à ce qui suit :

Définition. *Une forme F définie sur $\mathscr{C}^m_I(\overline{\Omega})$ se* ***prolonge continûment*** *à $H^m(\Omega)$ si $F(f_N)$ a la même limite pour toute suite de Cauchy associée à un même élément de $H^m(\Omega)$.*

Pour cela, il faut et il suffit que F soit bornée dans $\mathscr{C}^m_I(\overline{\Omega})$:

$$\|F\| = \sup_{f \neq 0} \frac{F(f)}{\|f\|_m} < \infty.$$

En effet, si $\|F\| < \infty$ et si $\{f_N\}$ est Cauchy dans $\mathscr{C}^m_I(\overline{\Omega})$, $\{F(f_N)\}$ est Cauchy dans $\mathbb{R}$, puisque

$$|F(f_N) - F(f_M)| = |F(f_N - f_M)| \leqslant \|F\| \, \|f_N - f_M\|_m,$$

donc converge vers un réel que l'on appelle $F(f)$. On vérifie que deux suites de Cauchy définissant un même élément f de $H^m(\Omega)$ donnent bien lieu à une même limite $F(f)$.

C'est certainement le cas pour des intégrales de dérivées de f sur Ω ou une partie de Ω :

Proposition. *Si $|a| \leqslant m$, $E \subseteq \Omega$ et la fonction $w \in L^2(E)$, alors*

$$\left| \int_E w(x)\, D^a f(x)\, dx \right| \leqslant \|w\|_{L^2(E)}\, \|f\|_m. \tag{42}$$

En effet, on applique Cauchy-Schwarz sur $L^2(E)$, et la norme de $D^a f$ sur $L^2(E)$ est évidemment inférieure à la norme sur $L^2(\Omega)$. □

Des valeurs ponctuelles peuvent encore avoir un sens ou non :

(1) $a \leqslant c \leqslant b$, $F(f) = f(c)$ n'a pas de sens dans $H^0(a,b)$: on peut construire f continue avec $f(c) = 1$ et $\|f\|_0 = \left[\int_a^b f^2(x)\, dx\right]^{1/2}$ aussi petit que l'on veut.

(2) $a \leqslant c \leqslant b, a < b$, $F(f) = f(c)$ a un sens dans $H^1(a,b) : \forall x \in [a,b], f(x) = f(c) + \int_c^x f'(t)\, dt$, donc $h = f - g$, où h est la fonction constante de valeur $f(c)$ et où $g(x) = \int_c^x f'(t)\, dt$. Passons aux normes dans $H^0(a,b)$: $\|h\|_0 = |f(c)|\sqrt{b-a} \leqslant \|f\|_0 + \|g\|_0$. Estimons $\|g\|_0$: par Cauchy-Schwarz, $|g(x)| \leqslant \|f'\|_0 \sqrt{b-a}$, donc $\|g\|_0^2 \leqslant \|f'\|_0^2 (b-a)^2$, et

$$|f(c)| \leqslant \frac{1}{\sqrt{b-a}} \|f\|_0 + \sqrt{b-a}\, \|f'\|_0.$$

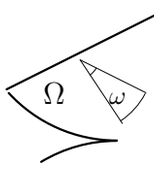
Enfin, de $\|f\|_1^2 = \|f\|_0^2 + \|f'\|_0^2$, il vient

$$|f(c)| \leqslant \sqrt{\frac{1}{b-a} + b - a} \qquad \|f\|_1.$$

Remarquons que la majorante de la norme de F ne dépend pas de c.

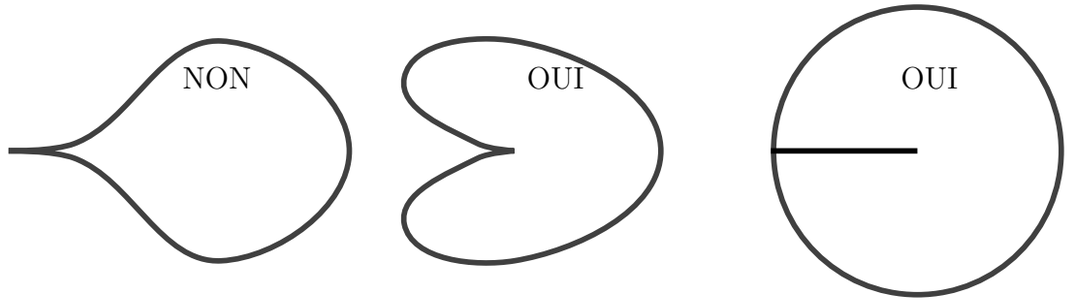
(3) $\Omega \subset \mathbb{R}^2$, $c \in \Omega$, $F(f) = f(c)$ n'a pas de sens dans $H^1(\Omega)$: on peut construire f continue avec $f(c) = 1$ et $\|f\|_1^2 = \int_\Omega f^2(\boldsymbol{x})\, d\boldsymbol{x} + \int_\Omega \|\operatorname{grad} f(\boldsymbol{x})\|^2\, d\boldsymbol{x}$ aussi petit que l'on veut. Prendre par exemple $f(\boldsymbol{x}) = 1 - \|\boldsymbol{x} - c\|^\alpha / \rho^\alpha$ sur $r := \|\boldsymbol{x} - c\| \leqslant \rho$ et $f = 0$ ailleurs, avec ρ petit et $\alpha > 0$. L'intégrale de f^2 est bornée par une constante fois ρ^2 et peut donc être rendue $< \varepsilon$; le gradient de f est dirigé radialement (à partir de c), sa norme vaut $|\partial f / \partial r| = \alpha r^{\alpha-1} / \rho^\alpha$. L'intégrale du carré de cette norme sur $r < \rho$ est donc $\int_0^\rho \int_0^{2\pi} \alpha^2 r^{2\alpha-2} \rho^{-2\alpha}\, r\, dr\, d\theta = \pi\alpha$, donc aussi petit que l'on veut, puisque α est positif arbitraire.
Autre fonction : $f(\boldsymbol{x}) = 1$ si $r := \|\boldsymbol{x} - c\| \leqslant \rho \exp(-1/\rho)$, $f(\boldsymbol{x}) = -\rho \log(r/\rho)$ si $\rho \exp(-1/\rho) \leqslant r \leqslant \rho$, $f(\boldsymbol{x}) = 0$ si $r \geqslant \rho$.

(4) Quand une valeur ponctuelle $f(c)$ a-t-elle encore un sens dans $H^m(\Omega)$? La condition est essentiellement $m > n/2$:



Définition. *On dit que l'ouvert borné $\Omega \subset \mathbb{R}^n$ possède la* **propriété du cône** *si tout point de $\overline{\Omega}$ peut être le sommet d'un cône fixe, de volume > 0, entièrement contenu dans $\overline{\Omega}$.* On prendra en fait un secteur sphérique de rayon ρ et d'angle solide $\omega > 0$.
Ω ne peut donc présenter de structure infiniment effilée, d'épine

infiniment pointue, etc. Par contre, un ouvert possédant la propriété de cône peut admettre une épine rentrante ou même entourer partout une partie de sa frontière.



Lemme de Sobolev : *si Ω possède la propriété de cône, si $m > n/2$, on a pour tout $f \in \mathscr{C}^m_I(\overline{\Omega})$,*

$$\forall c \in \overline{\Omega}, \qquad |f(c)| \leqslant C\|f\|_m,$$

où C ne dépend que de Ω (donc, ni de f ni de c).

En effet, il faut exprimer $f(c)$ selon une intégrale de volume de f et de ses dérivées jusqu'à l'ordre m. Soit τ une fonction $\mathscr{C}^m$ [d'une variable] telle que $\tau(0) = 1$, $\tau'(0) = \cdots = \tau^{(m)}(0) = 0$ et $\tau(\rho) = \tau'(\rho) = \cdots = \tau^{(m)}(\rho) = 0$. Sur tout segment de longueur ρ et d'extrémités c et d, $f(c) = \tau(0)f(c) - \tau(\rho)f(d) = -\displaystyle\int_0^\rho \frac{\partial[\tau(r)f(\boldsymbol{x})]}{\partial r}\,dr$, où r est la distance de c au point courant $\boldsymbol{x}$ sur le segment. Intégrons $m-1$ fois par parties, il n'y a jamais de termes aux limites, puisque $r^k\partial^k(\tau f)/\partial r^k$ est nul en $r = 0$ et $r = \rho$, $k = 1, 2, \ldots, m-1$:

$$f(c) = (-1)^m \int_0^\rho \frac{\partial^m[\tau(r)f(\boldsymbol{x})]}{\partial r^m}\frac{r^{m-1}}{(m-1)!}\,dr.$$

Intégrons selon les variables angulaires sur un angle solide ω, on obtient enfin une intégrale de volume sur le cône :

$$f(c) = \frac{(-1)^m}{\omega(m-1)!}\int_{\text{cône}} \frac{\partial^m[\tau(r)f(\boldsymbol{x})]}{\partial r^m} r^{m-n}\,d\boldsymbol{x},$$

où on a utilisé $d\boldsymbol{x} = r^{n-1}\,dr\,d\omega$. Appliquons (42) sur $L^2(\text{cône})$:

$$\left|\int_c F(\boldsymbol{x})G(\boldsymbol{x})\,d\boldsymbol{x}\right| \leqslant \sqrt{\int_c F^2(\boldsymbol{x})\,d\boldsymbol{x}}\sqrt{\int_c G^2(\boldsymbol{x})\,d\boldsymbol{x}},$$

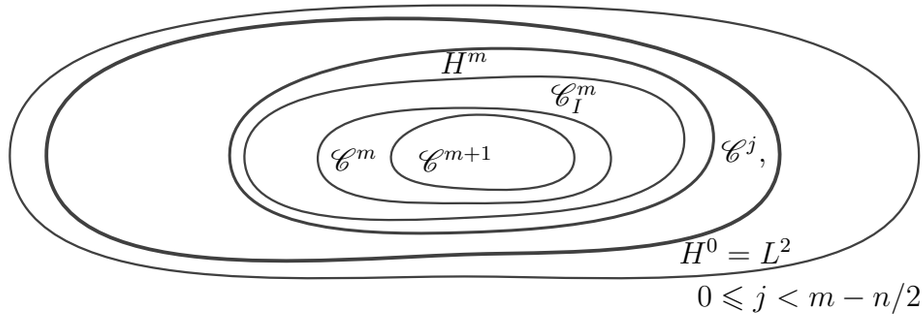
avec $F = \partial^m(\tau f)/\partial r^m$ et $G = r^{m-n}$:

$$|f(c)| \leqslant \frac{1}{\omega(m-1)!}\sqrt{\int_{\text{cône}}\left(\frac{\partial^m[\tau(r)f(\boldsymbol{x})]}{\partial r^m}\right)^2 d\boldsymbol{x}\int_{\text{cône}} r^{2m-2n}\,d\boldsymbol{x}}.$$

La deuxième intégrale fait intervenir r^{2m-n-1}, elle existe [et vaut $\omega\rho^{2m-n}/(2m-n)$] SI $m > n/2$. La première intégrale est bornée par des plus grandes valeurs prises par des dérivées de τ (facteurs géométriques, ne dépendant que de Ω) multipliées par des intégrales de carrés de dérivées d'ordre $\leqslant m$ de f. □

On déduit $H^m(\Omega) \subset \mathscr{C}(\overline{\Omega})$ si $m > n/2$ (et propriété de cône) : toute suite de Cauchy définissant un élément de $H^m(\Omega)$ est encore Cauchy pour la norme du maximum $\| \ \|_\infty$, donc converge dans $\mathscr{C}(\overline{\Omega})$, complet selon $\| \ \|_\infty$.

On a aussi $H^m(\Omega) \subset \mathscr{C}^j(\overline{\Omega})$ si $m > j + n/2$, en reprenant avec une dérivée d'ordre j.



On vient de voir un échantillon des ***théorèmes d'immersion de Sobolev***.

On dit qu'un espace topologique X est immergé dans un espace topologique Y si $X \subseteq Y$ et si l'injection canonique (identité) $X \to Y$ est continue : pour des normés, si $u \in X$, donc $u \in Y$, on a $\|u\|_Y \leqslant C\|u\|_X$.

On vient de voir un cas avec $X = H^m(\Omega)$ et $Y = \mathscr{C}(\overline{\Omega})$, si $m > n/2$, aussi $X = H^m(\Omega)$ et $Y = \mathscr{C}^j(\overline{\Omega})$, si $m > j + n/2$, et si la propriété de cône est valable dans $\overline{\Omega}$.

Les énoncés du ***théorème d'immersion de Sobolev*** sont [Adams, chap. V] :
Si le domaine, borné ou non, $\Omega \subset \mathbb{R}^n$ vérifie la propriété de cône,

(1) Si $mp < n$,
$H^{j+m,p}(\Omega) \subset H^{j,q}(\Omega), \qquad p \leqslant q \leqslant np/(n-mp)$,
$j = 0 : H^{m,p}(\Omega) \subset L^q(\Omega), \qquad p \leqslant q \leqslant np/(n-mp)$,

(2) si $mp = n$,
$H^{m,p}(\Omega) \subset L^q(\Omega), \qquad p \leqslant q < \infty$,

(3) si $mp > n$,
$H^{j+m,p}(\Omega) \subset \mathscr{C}^j_B(\Omega)$,

où $\mathscr{C}^j_B(\Omega)$ désigne l'espace des fonctions $\mathscr{C}^j(\Omega)$ de dérivées d'ordre $\leqslant j$ bornées dans Ω.

Pour les ***immersions compactes des espaces de Sobolev*** on a

Lemme de compacité de Rellich. *Si Ω est un domaine de frontière lipschitzienne, l'injection canonique $H^{m+1}(\Omega) \to H^m(\Omega)$ est compacte.*

Preuves dans [Raviart & Thomas] § 1.5 , [Adams] chap. 6 (théorème de Rellich-Kondrachov), Showalter[3] chap. 2, th. 5.8.

Enoncé équivalent : la boule unité [fermée] de $H^{m+1}(\Omega)$ est une partie compacte de $H^m(\Omega)$.

Une partie fermée bornée d'un espace de dimension infinie n'est normalement pas compacte...Ainsi, la boule unité d'un espace de Hilbert U de dimension infinie ne peut être compacte (selon la norme de U) : d'une famille orthonormale totale $\{\varphi_k\}$, on devrait pouvoir extraire une suite $\{\varphi_{k_i}\}$ convergente dans U. La limite, devant être orthogonale à tous les φ_k, ne peut être que l'élément nul de U, impossible, puisque les normes, toutes égales à l'unité, devraient tendre vers zéro.

[3] R.E. Showalter, *Hilbert Space Methods for Partial Differential Equations*, 1994, `http ://ejde.math.swt.edu/Monographs/01-Showalter/`

En fonctions continues, selon la norme $\| \ \|_\infty$, on a le ***théorème d'Arzela et Ascoli*** : un ensemble P de fonctions continues sur un compact I de $\mathbb{R}^n$ est compact si et seulement si

(1) il est fermé

(2) il est borné

(3) il est constitué d'une famille ***équicontinue*** de fonctions : $\forall \varepsilon > 0$, $\exists \delta$, $\forall f \in P$, $|f(x) - f(y)| \leqslant \varepsilon$ si $\|x - y\| \leqslant \delta$. Le même δ sert donc pour tous les $f \in P$. Ou encore : le module de continuité $\omega(h) = \sup\limits_{\substack{x,y\in I \\ \|x-y\|\leqslant h}} |f(x) - f(y)|$ est indépendant de $f \in P$.

On peut établir [Adams, *op.cit.*] le lemme de Rellich à partir de l'énoncé pour des fonctions continûment dérivables :

La boule unité de $\mathscr{C}^{m+1}(\overline{\Omega})$ est une partie compacte de $\mathscr{C}^m(\overline{\Omega})$.
En effet, il suffit de considérer une dérivée $m^{\text{ème}}$:

$$|D^\alpha f(x) - D^\alpha f(y)| = \left| \int_0^1 \mathbf{grad}\, D^\alpha f(x + t(y - x)) \cdot (y - x)\, dt \right|,$$

où $|\alpha| = m$, est évidemment bornée indépendamment de f puisque les dérivées $(m+1)^{\text{èmes}}$ de f sont bornées par l'unité. Ω doit être borné, et deux points de Ω doivent pouvoir être joints par un arc rectifiable de longueur $\leqslant$ const. $\|y - x\|$ ([Adams], § 1.31).

2.2. Formes bilinéaires et opérateurs ; dérivées faibles.

De même, une forme bilinéaire

$$a(f, g) := \sum_{|\alpha|\leqslant m} \int_\Omega q_\alpha(\boldsymbol{x}) D^\alpha f(\boldsymbol{x})\, D^\alpha g(\boldsymbol{x})\, d\boldsymbol{x}$$

se prolonge par continuité à $H^m(\omega)$ si $|a(f, f)| \leqslant$ const. $\|f\|_m^2$, ce qui est le cas si les fonctions q_α sont bornées en norme du maximum [en fait, dans L^∞ : mesurables et essentiellement bornées, $\|q_\alpha\|_\infty \leqslant$ const].

Enfin, un ***opérateur*** défini sur $\mathscr{C}_I^m(\overline{\Omega})$ et à valeurs dans $\mathscr{C}_I^k(\overline{\Omega})$ se prolonge par continuité en un opérateur $H^m(\Omega) \to H^k(\Omega)$ s'il est borné :

$$\{f_N\} \text{ Cauchy } \Rightarrow \{Af_N\} \text{ Cauchy si } \|A\| := \sup_{f\neq 0} \frac{\|Af\|_k}{\|f\|_m} < \infty.$$

C'est le cas de toute dérivée partielle d'ordre $\leqslant m$:

$$\|D^\alpha f\|_{m-|\alpha|} \leqslant \|f\|_m,$$

donc $D^\alpha : H^m(\Omega) \to H^{m-|\alpha|}(\Omega)$ est borné [de norme $\leqslant 1$].

On appelle ***dérivée faible*** $D^\alpha f$ cet élément de $H^{m-|\alpha|}(\Omega)$ obtenu par continuité.

Une dérivée faible est un exemple de ***dérivée distributionnelle*** : on a

$$f \in H^m(\Omega), |\alpha| \leqslant m \Rightarrow \quad \int_\Omega \psi(\boldsymbol{x}) D^\alpha f(\boldsymbol{x})\, d\boldsymbol{x} = (-1)^{|\alpha|} \int_\Omega f(\boldsymbol{x}) D^\alpha \psi(\boldsymbol{x})\, d\boldsymbol{x} \tag{43}$$

pour tout $\psi \in \mathscr{D}(\Omega) = \mathscr{C}_0^\infty(\Omega)$, l'espace des fonctions indéfiniment dérivables dans $\mathbb{R}^n$ et de support compact dans Ω (donc, sans point commun avec la frontière de Ω !).

(Vérification de (43) : l'égalité est vraie si $f \in \mathscr{C}^m_I(\overline{\Omega})$, et les deux membres de l'égalité sont des formes bornées, donc continues sur $H^m(\Omega)$).

On appelle $W^m(\Omega)$ l'espace des fonctions de $L^2(\Omega)$ admettant des dérivées distributionnelles d'ordre $\leqslant m$ dans $L^2(\Omega)$. On a donc $H^m(\Omega) \subseteq W^m(\Omega)$. On montre que $H^m(\Omega) = W^m(\Omega)$ ([Adams], p. 52).

Une fonction admettant des dérivées classiques d'ordre m est évidemment dans $H^m(\Omega)$. Le test (43) permet d'éclaircir des cas de fonctions régulières presque partout, et d'éviter quelques pièges :

(1) Voici une fonction qui n'est pas dans un H^1 :

Si $a < c < b$, $f =$ sign $(x-c)$ n'est pas dans $H^1(a,b)$:

$$\forall g \in \mathscr{D},\ (f,g')_0 = \int_a^b f(x)g'(x)\,dx = -(g(c)-g(a)) + g(b) - g(c) = -2g(c)$$

qui devrait valoir $-(g,f')_0 = -\displaystyle\int_a^b g(x)f'(x)\,dx$, ce qui est borné par $\|g\|_0\|f'\|_0 \leqslant$ constante $\|g\|_0$. Impossible : pour un $g(c)$ donné, on peut trouver des fonctions g telles que $\|g\|_0$ soit arbitrairement petite.

(2) **Si $a < c < b$, quand $f = |x-c|^\rho$ est elle dans $H^1(a,b)$?**

La dérivée "ordinaire" $\rho\,\mathrm{sign}(x-c)|x-c|^{\rho-1}$ est dans H^0 si l'intégrale $\displaystyle\int_a^b (x-c)^{2\rho-2}\,dx$ a un sens, soit $2\rho - 2 > -1$: $\rho > 1/2$ (et $\rho = 0$).

Il y a donc des fonctions continues ($0 < \rho < 1/2$) qui ne sont pas dans $H^1(a,b)$.

Pour établir $f \in H^m(\Omega)$ à partir de dérivées faibles "devinées", on a utilisé le **Théorème.** $H^m(\Omega)$ est la partie de $H^0(\Omega)$ admettant des dérivées faibles d'ordre $\leqslant m$ dans $H^0(\Omega)$: $H^m(\Omega) = W^m(\Omega)$.

(3) Dans $\mathbb{R}^n$: soit $c \in \Omega$ et examinons $f(x) = \|x-c\|^\rho$ (norme euclidienne). La somme des carrés des normes $\|\ \|_0$ des dérivées partielles premières de f est $\displaystyle\int_\Omega \|\ \mathrm{grad}\ f\|^2\,dx$ existe si l'intégrale de $r^{2\rho-2}r^{n-1}\,dr$ existe ($r = \|x-c\|$), donc si $2\rho + n - 3 > -1$:
$\|x-c\|^\rho \in H^1(\Omega)$ si $\rho > 1 - n/2$.
$\|x-c\|^\rho \in H^m(\Omega)$ si $\rho > m - n/2$.

(4) **Séries de Fourier.** $H_0(0,1)$ est isomorphe à l'ensemble des suites (complexes) $\{c_k\}_{k=-\infty}^{\infty}$ de carré sommable $\displaystyle\sum_{-\infty}^{\infty} |c_k|^2 < \infty$ par $(f,g)_0 = \sum_{-\infty}^{\infty} c_k\,\overline{d_k}$ (Riesz-Fischer).

Pour $f \in \mathscr{C}^1_I$, avec $f(0) = f(1)$, on a $f(x) = \displaystyle\sum_{-\infty}^{\infty} c_k \exp(2\pi i k x)$, $\forall x \in [0,1]$.

Quand f est-elle dans $H^1(0,1)$? Soit $g \in \mathscr{D}$. Si $\{d_k\}$ sont les coefficients de Fourier de g, on a $g'(x) = \displaystyle\sum_{-\infty}^{\infty} 2\pi i k d_k \exp(2\pi i k x)$, ponctuellement dans $[0,1]$, donc les $2\pi i k d_k$ sont les coefficients de g'.

Exprimons maintenant $(f, g')_0 = -(f', g)_0$:
$(f, g')_0 = \sum_{-\infty}^{\infty} c_k(-2\pi ik)\overline{d_k} = -\sum_{-\infty}^{\infty} 2\pi ikc_k\,\overline{d_k}$. f' doit donc être associée aux coefficients de Fourier $2\pi ikc_k$, qui est de carré sommable si $\sum_{-\infty}^{\infty} k^2\,|c_k|^2 < \infty$, ce qui *caractérise* les éléments (périodiques) de $H^1(0,1)$.

$H^m(0,1)$ périodique : $\sum_{-\infty}^{\infty} k^{2m}\,|c_k|^2 < \infty$.

(5) **Transformée de Fourier.** Si $f \in H^0(\Omega) = L^2(\Omega)$, on peut prolonger f en une fonction de $L^2(\mathbb{R}^n)$ et considérer sa transformée de Fourier

$$\hat{f}(\boldsymbol{\xi}) := (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, f(\boldsymbol{x})\, d\boldsymbol{x}.$$

On a $\hat{f} \in L^2(\mathbb{R}^n)$ et $\|\hat{f}\|_0 = \|f\|_0$ (*Plancherel*).

Alors, on montre que

$$H^1(\mathbb{R}^n) = \{f \in L^2(\mathbb{R}^n) : \sqrt{1+|\boldsymbol{\xi}|^2}\ \hat{f}(\boldsymbol{\xi}) \in L^2(\mathbb{R}^n)\}.$$

De plus, si Ω est de frontière suffisamment régulière, toute fonction de $H^1(\Omega)$ peut être prolongée en une fonction de $H^1(\mathbb{R}^n)$ [Raviart & Thomas] Th. 1.5-1.

2.3. Traces.

Une formulation variationnelle $a(u,v) = \varphi(v)$ initialement formulée dans $\mathscr{C}^m_I(\overline{\Omega})$ est progressivement définie sur $H^m(\Omega)$. Nous avons déjà vu comment interpréter des intégrales sur Ω de $D^\alpha u\, D^\alpha v$ (énergie de déformation, énergie potentielle), des valeurs ponctuelles de u et de dérivées de u si m est assez grand (forces ou liaisons se portant sur un seul point). Il faut encore considérer des intégrales sur la *frontière* de Ω associées à des conditions aux limites sur la dite frontière.

Pour cela, on fait des hypothèses supplémentaires sur la frontière de Ω, hypothèses plus fortes que celles déjà rencontrées (propriété de cône) :

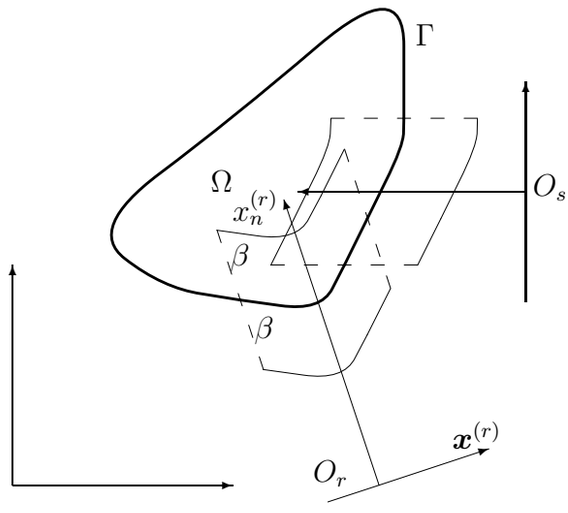
Le domaine Ω est dit de frontière $\Gamma = \partial\Omega$ ***lipschitzienne*** [4], lorsque
Il existe $\alpha > 0$ et $\beta > 0$, et un nombre fini de systèmes d'axes $\{x_1^{(r)}, \ldots, x_{n-1}^{(r)}, x_n^{(r)}\} = \{\boldsymbol{x}^{(r)}, x_n^{(r)}\}, r = 1, \ldots, R$ où

(1) $\Gamma = \bigcup_{r=1}^{R} \{\{\boldsymbol{x}^{(r)}, x_n^{(r)}\} \in \mathbb{R}^{n-1} \times \mathbb{R} : x_n^{(r)} = \varphi_r(\boldsymbol{x}^{(r)}), \|\boldsymbol{x}^{(r)}\| < \alpha\}$,
où les fonctions φ_r sont *lipschitziennes*, c'est-à-dire continues et $|\varphi_r(\boldsymbol{y}) - \varphi_r(\boldsymbol{x})| \leqslant C\|\boldsymbol{y} - \boldsymbol{x}\|$, $\forall \boldsymbol{x}$ et $\boldsymbol{y} \in \mathbb{R}^{n-1}, \|\boldsymbol{x}\| < \alpha, \|\boldsymbol{y}\| < \alpha$.

(2) $\{\{\boldsymbol{x}^{(r)}, x_n^{(r)}\} : \|\boldsymbol{x}^{(r)}\| < \alpha \text{ et } \varphi_r(\boldsymbol{x}^{(r)}) < x_n^{(r)} < \varphi_r(\boldsymbol{x}^{(r)}) + \beta\} \subset \Omega,\ r = 1, \ldots, R,$

(3) $\{\{\boldsymbol{x}^{(r)}, x_n^{(r)}\} : \|\boldsymbol{x}^{(r)}\| < \alpha \text{ et } \varphi_r(\boldsymbol{x}^{(r)}) - \beta < x_n^{(r)} < \varphi_r(\boldsymbol{x}^{(r)})\} \subset \complement\overline{\Omega},\ r = 1, \ldots, R,$

[4] [Ciarlet] p. 13, d'après J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson, 1967.



La frontière ne peut revenir sur elle-même ; aucune partie de Γ ne peut être entourée entièrement de points de Ω.
On accepte des points anguleux (sommets polyédriques ou coniques à 3 dimensions), mais pas de point de rebroussement. Un domaine de frontière lipschitzienne est nécessairement borné.

Ces conditions sur la frontière serviront à établir quelques résultats de densité et, surtout, à bien préciser les conditions aux limites de problèmes à résoudre dans des espaces de Hilbert.

Montrons par exemple que, si $\Omega \subset \mathbb{R}^2$, $F(f) = \displaystyle\int_a^b f(x)\,dx$ a un sens dans $H^1(\Omega)$ si $[a,b]$ est entièrement contenu dans $\overline{\Omega}$ (de frontière lipschitzienne) :
en effet, $[a,b]$ est un côté d'un rectangle de hauteur β entièrement contenu dans $\overline{\Omega}$, par exemple $[a,b]\times[0,\beta]$, donc,
pour x fixé,

$$\begin{aligned} f(x,0) &= \left[(1-y/\beta)\,f(x,y)\right]_{y=0} \\ &= -\int_0^\beta \frac{d}{dy}\left[(1-y/\beta)\,f(x,y)\right]\,dy \\ &= \int_0^\beta \left(\frac{f}{\beta} + \left(\frac{y}{\beta}-1\right)\frac{\partial f}{\partial y}\right)\,dy, \end{aligned}$$

$$|f(x,0)| \leqslant \left[\int_0^\beta f^2(x,y)\,dy\right]^{1/2}\beta^{-1/2} + \left[\int_0^\beta \left(\frac{\partial f}{\partial y}\right)^2\,dy\right]^{1/2}(\beta/3)^{1/2},$$

$$f^2(x,0) \leqslant \left(1+\max(\beta,\beta^{-1})\right)\left[\int_0^\beta f^2(x,y)\,dy + \int_0^\beta \left(\frac{\partial f}{\partial y}\right)^2\,dy\right]$$

et intégrons en x

$$\int_a^b f^2(x,0)\,dx \leqslant \left(1+\max(\beta,\beta^{-1})\right)\left[\int_a^b\int_0^\beta f^2(x,y)\,dxdy + \int_a^b\int_0^\beta \left(\frac{\partial f}{\partial y}\right)^2\,dxdy\right]$$

enfin,

$$\left[\int_a^b f(x,0)\,dx\right]^2 \leqslant (b-a)\int_a^b f^2(x,0)\,dx \leqslant (b-a)\left(1+\max(\beta,\beta^{-1})\right)\|f\|_1^2.$$

Montrons que, $\forall f \in \mathscr{C}^1_I(\Omega)$ de frontière lipschitzienne au sens *supra*,

$$\|f\|_{L^2(\partial\Omega)} \leqslant C(\Omega)\|f\|_1.$$

La frontière $\partial\Omega$ est une réunion de morceaux de surface $x_n = \varphi(x_1,\ldots,x_{n-1})$ sur lesquels l'intégrale s'écrit

$$\int_S f^2(x)\,dS = \int_{S'} f^2(x_1,\ldots,x_{n-1},\varphi)\,\sqrt{1+|\mathbf{grad}\,\varphi|^2}\,dx_1\cdots dx_{n-1}.$$

Soit ψ une fonction continûment dérivable d'une variable qui vaut 1 en φ et 0 en $\varphi+\beta$. On a donc

$$\int_S f^2(x)\,dS = -\int_{S'}\int_\varphi^{\varphi+\beta} \frac{\partial f^2(x_1,\ldots,x_{n-1},x_n)\psi(x_n)}{\partial x_n}\,\sqrt{1+|\mathbf{grad}\,\varphi|^2}\,dx_1\cdots dx_{n-1}dx_n.$$

On a maintenant une intégrale sur un morceau de volume entièrement situé dans $\overline{\Omega}$, d'une combinaison de f^2 et $f\partial f/\partial x_n$. Par Cauchy-Schwarz et Hölder, l'intégrale est majorée par $\|f\|_1^2$ fois des facteurs géométriques. □

On peut donc étendre par continuité l'opérateur de restriction à la frontière en un opérateur borné $H^1(\Omega) \to L^2(\partial\Omega)$: c'est l'opérateur ***trace*** γ_0.

Dans $H^m(\Omega)$, on définit aussi les traces des dérivées normales $\dfrac{\partial^j u}{\partial n^j}$ sur la frontière, $\gamma_j, j = 0,1,\ldots,m-1$.

$H^m_0(\Omega)$ est le sous espace des éléments de $H^m(\Omega)$ vérifiant $\gamma_0 u = \cdots = \gamma_{m-1}u = 0$.
$H^m_0(\Omega)$ est aussi l'adhérence de $\mathscr{D}(\Omega)$ dans $H^m(\Omega)$.

En ***résumé***, on a donc vu trois types de formes ayant un sens, par continuité, et sous conditions, sur $H^m(\Omega)$:

Forme	Conditions	Espace
$(D^\alpha u)(P)$	$0 \leqslant \lvert\alpha\rvert < m - n/2$ propr. de cône	$H^m(\Omega) \subseteq \mathscr{C}^{\lvert\alpha\rvert}(\overline{\Omega})$
$\int_\Omega w D^\alpha u\,d\boldsymbol{x}$	$\lvert\alpha\rvert \leqslant m$ $w \in L^2(\Omega)$	$D^\alpha u \in L^2(\Omega)$
$\int_{\partial\Omega} w D^\alpha u\,dS$	$\lvert\alpha\rvert \leqslant m-1$ $\partial\Omega$ lipschitzienne	$D^\alpha u\big\rvert_{\partial\Omega} \in H^{m-\lvert\alpha\rvert-1}(\partial\Omega)$

3. Coercivité de $\int_\Omega$ **grad** $u\cdot$ **grad** $v\,d\boldsymbol{x}$ dans $H^1_0(\Omega)$.

Rappelons que $H^1_0(\Omega) = \overline{\mathscr{D}}$ dans $H^1(\Omega)$ = éléments de $H^1(\Omega)$ de trace nulle (sur $\Gamma = \partial\Omega$), quand Ω est de frontière lipschitzienne.

Soit Ω de frontière lipschitzienne, montrons qu'il existe $c > 0$:

$$a(u,u) = \int_\Omega \textbf{grad } u\cdot \textbf{ grad } u\,d\boldsymbol{x} \geqslant c\|u\|_1^2,$$

pour tout $u \in H^1_0(\Omega)$. La coercivité de $\int_\Omega$ **grad** $u\cdot$ **grad** $v\,d\boldsymbol{x}$ est évidemment équivalente à ***l'inégalité de Poincaré-Friedrichs***

$$\forall u \in H^1_0(\Omega), \qquad \|u\|_0 \leqslant c'\|\textbf{grad } u\|_0,$$

avec $c = \dfrac{1}{c'^2+1}$.

Démonstration directe de l'inégalité de Poincaré-Friedrichs [Raviart & Thomas] Th. 1.2-5. Il suffit de considérer u dans l'espace $\mathscr{D}(\overline{\Omega})$ dense dans $H^1_0(\Omega)$. On prolonge cette fonction u dans tout $\mathbb{R}^n$ par $u = 0$ hors de Ω. Isolons la composante x_n de $\boldsymbol{x} = (x_1,\ldots,x_n)$. Le domaine borné Ω est donc situé entièrement dans $a < x_n < b$, avec $-\infty < a < b < \infty$. Comme $u = 0$ en $x_n = a$,

$$u(\boldsymbol{x}) = \int_a^{x_n} \frac{\partial u}{\partial x_n}(x_1,\ldots,x_{n-1},t)\,dt.$$

Par Cauchy-Schwarz,

$$u^2(\boldsymbol{x}) \leqslant (x_n - a)\int_a^{x_n}\left(\frac{\partial u}{\partial x_n}\right)^2(x_1,\ldots,x_{n-1},t)\,dt \leqslant (x_n - a)\int_{-\infty}^{\infty}\left(\frac{\partial u}{\partial x_n}\right)^2(x_1,\ldots,x_{n-1},t)\,dt,$$

intégrons sur tous les $(n-1)-$uples $(x_1,\ldots,x_{n-1})$:

$$\int_{\mathbb{R}^{n-1}} u^2(x_1,\ldots,x_{n-1},x_n)dx_1\ldots dx_{n-1} \leqslant (x_n - a)\int_{\mathbb{R}^n}\left(\frac{\partial u}{\partial x_n}\right)^2(\boldsymbol{x})\,d\boldsymbol{x},$$

et intégrons enfin en x_n entre a et b :

$$\|u\|_0^2 = \int_{\mathbb{R}^n} u^2(\boldsymbol{x})d\boldsymbol{x} \leqslant \frac{(b-a)^2}{2}\left\|\frac{\partial u}{\partial x_n}\right\|_0^2 \leqslant \frac{(b-a)^2}{2}\|\textbf{grad } u\|_0^2,$$

ce qui donne bien l'inégalité cherchée, avec $c' \leqslant (b-a)/\sqrt{2}$. □

Démonstration de coercivité utilisant le lemme de compacité de Rellich :

Montrer que la forme bilinéaire a est coercive revient à dire que $a(u,u)$ est minorée par $c > 0$ sur la sphère unité :

$$a(u,u) \geqslant c\|u\|^2 \Leftrightarrow a\left(\frac{u}{\|u\|},\frac{u}{\|u\|}\right) \geqslant c > 0.$$

Supposons a non coercive, il existe alors une suite $\{u_i\}_i$, avec $\|u_i\| = 1$ et $a(u_i,u_i) \to 0$. Tout fermé borné de $H^{m+1}(\Omega)$ étant un ***compact*** de $H^m(\Omega)$ (Rellich, cf. p.94), on peut extraire $\{u_{i_j}\}_j$ convergente dans $H^0(\Omega)$. Cette suite est également Cauchy, donc convergente dans $H^1(\Omega)$: si j et k sont assez grands, $\|u_{i_j} - u_{i_k}\|_0$ et $\|\textbf{grad } (u_{i_j} - u_{i_k})\|_0$ sont arbitrairement petits (pour la dernière norme : parce que $\|\textbf{grad } u_i\|_0 \to 0$),

donc $u_{i_j} \to u$ dans $H^1(\Omega)$, en fait dans $H^1_0(\Omega)$ (complet). u est donc un élément de $H^1_0(\Omega)$ de gradient nul $\Rightarrow u$ doit être la constante nulle, impossible puisque $\|u_{i_j}\|_1 = 1$. □

Autre exemple de forme coercive : $\displaystyle\int_\Omega \Delta u\, \Delta v\, d\boldsymbol{x}$ dans $H^2_0(\Omega)$ [Ciarlet, p. 14].

Comment trouver c (ou c') dans des cas concrets ? Minimiser $a(u,u)$ sur la sphère unité $(u,u)=1$ revient à minimiser $a(u,u)/(u,u)$ (*quotient de Rayleigh*) sur tous les u non nuls de $H^1_0(\Omega)$, donc à chercher u tel que

$$\frac{a(u+\varepsilon v, u+\varepsilon v)}{(u+\varepsilon v, u+\varepsilon v)} = \frac{a(u,u)+2\varepsilon a(u,v)+\varepsilon^2 a(v,v)}{(u,u)+2\varepsilon(u,v)+\varepsilon^2(v,v)}$$

soit minimum en $\varepsilon v = 0$, donc en $\varepsilon = 0$ pour tout $v \in U$. La dérivée en ε doit être nulle en $\varepsilon = 0$, ce qui donne $a(u,v) = \lambda(u,v), \forall v \in U$. Rappelons que, si a est bilinéaire continue symétrique sur l'espace de Hilbert U, on a $a(u,v) = (Au,v)_U$ (représentant de Riesz), avec A autoadjoint borné. On voit donc que les valeurs extrémales de $a(u,u)/(u,u)$ sont les ***valeurs propres*** de A : $c = \lambda_{\min}(A)$.

Pour Poincaré-Friedrichs, on préfère utiliser $\mathcal{A}$ tel que $(\mathcal{A}u, v)_0 = (\mathbf{grad}\ u, \mathbf{grad}\ v)_0$ pour tout u et v dans une partie dense de $H^1_0(\Omega)$. $\mathcal{A}$ est non borné mais plus "commode" que A.

Enfin, $\lambda(A) = \dfrac{\lambda(\mathcal{A})}{\lambda(\mathcal{A})+1}$: si $\mathcal{A}u = \lambda(\mathcal{A})u$, $a(u,v) = (\mathcal{A}u,v)_0 = \lambda(\mathcal{A})(u,v)_0 = \lambda(\mathcal{A})[(u,v)_1 - a(u,v)]$. Donc, en résumé,

$$c = \lambda_{\min}(A) = \frac{\lambda_{\min}(\mathcal{A})}{\lambda_{\min}(\mathcal{A})+1}, \quad c' = \frac{1}{\sqrt{\lambda_{\min}(\mathcal{A})}}.$$

Avec $\Omega = (0,L)$, $\displaystyle\int_0^L (\mathcal{A}u)v dx = \int_0^L u'v' dx = \int_0^L (-u'')v dx$, donc $\mathcal{A}u = -u''$, les fonctions propres sont $\sin(x\sqrt{\lambda})$ qui doivent encore s'annuler en $x = L \Rightarrow \lambda = k^2\pi^2/L^2, k = 1,2,\ldots$

$$c = \frac{\pi^2}{\pi^2+L^2}, \quad c' = \frac{L}{\pi}.$$

On voit l'importance de $L < \infty$.

Les $\sin(k\pi x/L)$, $k = 1,2,\ldots$ forment une suite orthogonale complète à la fois dans $H^0(0,L)$ et dans $H^1_0(0,1)$.

$H^1_0(0,L)$ est l'ensemble des séries de Fourier $\sum_1^\infty c_k \sin(k\pi x/L)$ avec $\sum_1^\infty k^2 c_k^2 < \infty$: si $u \in H^1_0(0,1)$, sa dérivée faible Du admet une série de Fourier $\sum_0^\infty d_k \cos(k\pi x/L)$ [considérer le prolongement pair de Du sur $(-L,L)$] avec $\sum_0^\infty d_k^2 < \infty$. De $(Du,v)_0 = -(u,Dv)_0$ pour tout $v \in H^1(0,1)$, on obtient, avec $v = \cos(k\pi x/L)$, le $k^{\text{ème}}$ coefficient de Fourier de u $c_k = -Ld_k/(k\pi), k = 1,2,\ldots (u \in H^1_0(0,L) \Rightarrow d_0 = 0)$.

Pour le disque de rayon R dans $\mathbb{R}^2$:

$$\mathcal{A} = -\Delta = -\frac{\partial^2}{\partial r^2} - \frac{1}{r}\frac{\partial}{\partial r} - \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2},$$

fonctions propres $= J_m(r\sqrt{\lambda})\exp(\pm im\theta), m = 0,1,\ldots \Rightarrow \lambda = (\text{zéros de } J_m/R)^2$,

$$c = \frac{5.78\ldots}{5.78\ldots + R^2} \qquad c' = \frac{R}{2.4048\ldots}$$

[me rappeler : "peut-on entendre la forme d'un tambour ?"]
Boule de rayon R dans $\mathbb{R}^3$:

$$\mathcal{A} = -\Delta = -\frac{\partial^2}{\partial r^2} - \frac{2}{r}\frac{\partial}{\partial r} - \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2} - \frac{1}{r^2}\frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial \theta} - \frac{1}{r^2}\frac{1}{\sin^2\theta}\frac{\partial^2}{\partial \varphi^2},$$

fonctions propres $= r^{-1/2} J_{n+1/2}(r\sqrt{\lambda}) P_n^m(\cos\theta)\exp(\pm im\varphi), m, n = 0, 1, \ldots, \Rightarrow \lambda =$ (zéros de $J_{n+1/2}/R)^2$.

$$c = \frac{\pi^2}{\pi^2 + R^2}, \quad c' = \frac{R}{\pi}.$$

Voir J. Meinguet, From Dirac distributions to multivariate representation formulas, pp. 225-248 *in* Z. Ziegler, editor, *Approximation Theory and Applications*, Ac. Press, 1981.

4. Erreur d'approximation de la méthode de Ritz.

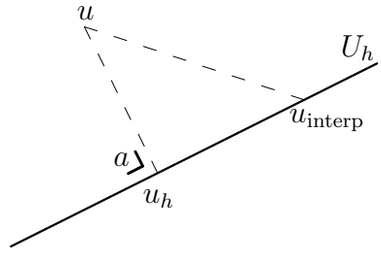
On a donc le problème variationnel

$$?u \in H^m(\Omega) : \ a(u, v) = \varphi(v), \ \forall v \in H^m(\Omega),$$

que l'on sait avoir une solution unique si a est bilinéaire symétrique *continue coercive* (Lax Milgram).

On approche u en résolvant ce même problème variationnel dans un sous espace U_h de dimension finie (méthode de Ritz)

$$?u_h \in U_h : \ a(u_h, v) = \varphi(v), \ \forall v \in U_h,$$



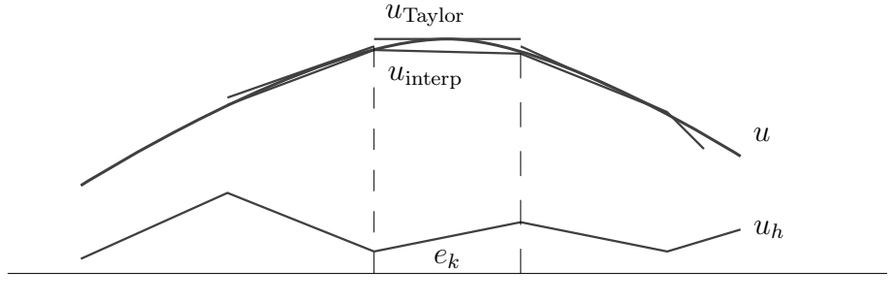
qui se traduit, pour un espace d'éléments finis, par un système linéaire avec matrice (*matrice de rigidité*) symétrique définie positive

$$[(L_i, L_j)_m]_{i,j=1,\ldots,N} [F_j(u_h)]_j = [\varphi(L_i)]_i,$$

où les F_i sont les formes attachées à des points de $\overline{\Omega}$, L_i (*base de Lagrange*) est la fonction de U_h dont la restriction à l'élément e_k vérifie $L_i|_{e_k} \in V_k$ et $F_j(L_i) = \delta_{i,j}$, pour $j \in Q_k$.
Soit u_{interp} l'élément de U_h dont la restriction à e_k est l'élément de V_k vérifiant

$$F_i(u_{\text{interp}}) = F_i(u), \quad i \in Q_k.$$

On ne pourra pas discuter directement la distance entre u et u_h, on estimera cette distance à partir de u_{interp} et d'autres fonctions intermédiaires dont le rôle séclaircira progressivement :



On a $u_{\text{interp}}(x) = \sum_{i \in Q_k} F_i(u) L_i(x)$ pour $x \in e_k$. N.B. : u_{interp} doit être dans $H^m(\Omega)$: on demandera ***Hypothèse 1*** $U_h \subset \mathscr{C}^m_I$: importance de la continuité globale des espaces d'éléments finis (***conformité***).

Estimons $\|u - u_h\|_m$ (***lemme de Céa***), par ***Hypothèse 2*** : a coercive et continue sur H^m :

$$\begin{aligned}
\|u - u_h\|_m^2 &\leqslant \frac{1}{c} a(u - u_h, u - u_h) \qquad \text{coercivité de } a \\
&\leqslant \frac{1}{c} a(u - u_{\text{interp}}, u - u_{\text{interp}}) \quad u_h \text{ est la } a\text{−meilleure approximation de } u \text{ dans } U_h \\
&\leqslant \frac{d}{c} \|u - u_{\text{interp}}\|_m^2 \qquad \text{continuité de } a
\end{aligned}$$

Ensuite :

$$\|u - u_{\text{interp}}\|_m^2 = \sum_{|s| \leqslant m} \sum_{e_k} \int_{e_k} \left(D^s u - D^s u_{\text{interp}}\right)^2 dx.$$

On apprécie $D^s u - D^s u_{\text{interp}}$ en remarquant que l'on a encore $D^s u - D^s u_{\text{interp}} = D^s(u - p) - D^s(u - p)_{\text{interp}}$, où p est un élément quelconque de V_k (les éléments de V_k sont leurs propres interpolants !) (lemme de Bramble).

Prenons une hypothèse (très) forte : ***Hypothèse 3***, $u \in \mathscr{C}^{m+1}(\overline{\Omega})$, (en fait, $u \in \mathscr{C}^m(\overline{\Omega})$, de dérivées $(m+1)^{\text{èmes}}$ bornées dans $\overline{\Omega}$), et prenons $p = u_{\text{Taylor}}$, le développement de Taylor de u de degré m autour d'un point $c_k \in e_k$. Alors,

$$|D^s u - D^s u_{\text{interp}}| \leqslant |D^s(u - u_{\text{Taylor}})| + |D^s((u - u_{\text{Taylor}})_{\text{interp}})|.$$

Lemme. *Si* **Hypothèse 4**, *e_k est étoilé autour de c_k, c'est-à-dire si le segment $[c_k, x]$ est entièrement dans $e_k \subset \mathbb{R}^n$ pour tout $x \in e_k$, si $u \in \mathscr{C}^{m+1}(e_k)$, et si* **Hypothèse 5**, *V_k contient $\mathscr{P}_m$,*

$$|D^s(u - u_{\text{Taylor}})(x)| \leqslant (nh_k)^{m+1-|s|} \|D^{m+1} u\|_\infty, \tag{44}$$

tant que $|s| \leqslant m+1$, où h_k est le diamètre de e_k, et $\|D^{m+1}u\|_\infty$ est la plus grande valeur absolue atteinte par une dérivée d'ordre $m+1$ de u dans e_k.

En effet, le lemme est trivial si $|s| = m+1$ ($D^{m+1} u_{\text{Taylor}} = 0$). Si le lemme est vrai pour toutes les dérivées d'ordre $|s|+1$, montrons qu'il est vrai pour D^s :

$$D^s(u - u_{\text{Taylor}})(x) = \int_{c_k}^{x} \mathbf{grad}\ (D^s(u - u_{\text{Taylor}}))(y) \cdot \vec{\boldsymbol{\tau}}\, d\ell,$$

(toutes les dérivées d'ordre $\leqslant m$ de $u - u_{\text{Taylor}}$ s'annulent en $x = c_k$), où $\vec{\boldsymbol{\tau}}$ est le vecteur unitaire dirigé de c_k vers x et $d\ell$ l'élément de longueur sur le segment $[c_k, x]$. Chaque composante du gradient est une dérivée d'ordre $|s|+1$, donc bornée par $(nh_k)^{m-|s|}\|D^{m+1}u\|_\infty$, chaque composante de $\vec{\boldsymbol{\tau}}$ est de valeur absolue $\leqslant 1$, et l'intégrale de $d\ell$ est la longueur du segment $\leqslant h_k$. □

Passons maintenant à $(u - u_{\text{Taylor}})_{\text{interp}}$, qui s'exprime dans la base de Lagrange de V_k par

$$(u - u_{\text{Taylor}})_{\text{interp}} = \sum_{i \in Q_k} F_i(u - u_{\text{Taylor}}) L_i,$$

$$D^s(u - u_{\text{Taylor}})_{\text{interp}} = \sum_{i \in Q_k} F_i(u - u_{\text{Taylor}}) D^s(L_i). \tag{45}$$

Si F_i fait appel à au plus r_i valeurs de dérivées d'ordre $\leqslant \ell_i$, on a, par (44)
$|F_i(u - u_{\text{Taylor}})| \leqslant r_i (nh_k)^{m+1-\ell_i} \|D^{m+1}u\|_\infty$.

On suppose encore ***Hypothèse 6***, que les éléments de la base de Lagrange possèdent la propriété de régularité suivante :

$$|D^s L_i| \leqslant q h_k^{\ell_i - |s|},$$

(vérifications dans deux cas simples :

(1) e_k = triangle de $\mathbb{R}^2$, $V_k = \mathscr{P}_1$, F_i = valeur en un sommet (espace de Courant : $\ell_i = 0$), alors $0 \leqslant L_i(x) \leqslant 1$ dans e_k, et $|\mathbf{grad}\ L_i| \leqslant 1/\text{plus petite hauteur} \leqslant (2/|\sin\theta|_{\min}) h_k^{-1}$ (ici, h_k est le plus grand côté du triangle e_k). L'hypothèse est donc vérifiée si *les angles ne sont proches ni de 0 ni de* π.

(2) $e_k = [x_k, x_{k+1}] \subset \mathbb{R}$, $V_k = \mathscr{P}_3$, F_i = valeur de la dérivée première en x_k (interpolant d'Hermite). Alors, $L_i(x) = (x - x_k)(x_{k+1} - x)^2/h_k^2$, où $h_k = x_{k+1} - x_k$, on trouve $L_i'(x) = (x_{k+1} - x)(x_{k+1} + 2x_k - 3x)/h_k^2$, $|L - i(x)| \leqslant 4h_k/27$ et $|L_i'(x)| \leqslant 1$, ce qui correspond bien à $\ell_i = 1$.

alors, (45) est bornée par $qrn^{m+1} h_k^{m+1-|s|} \|D^{m+1}u\|_\infty$, où $r = \sum_{i \in Q_k} r_i$.

Enfin,

$$\|D^s u - D^s u_{\text{interp}}\|_0 = \left[\sum_{e_k} \int_{e_k} \left(D^s u - D^s u_{\text{interp}}\right)^2 dx \right]^{1/2} \leqslant \sqrt{\mu(\Omega)}(1+qr) n^{m+1} h^{m+1-|s|} \|D^{m+1}u\|_\infty.$$

et $\|u - u_h\|_m \leqslant$ constante $h \to 0$ quand $h \to 0$.

Fin du théorème . □

Exemple d'estimation plus fine (Nitsche, cf. [Ciarlet] § 3.3) :
si $-\Delta u = f \in H^0(\Omega)$, $\Omega \subset \mathbb{R}^2$ de frontière lipschitzienne, $u \in H_0^1(\Omega)$ et si u a ses dérivées secondes bornées dans $\overline{\Omega}$), on a

$$|u(x) - u_h(x)| \leqslant Ch^2 |\log h|^{3/2} \|D^2 u\|_\infty,$$
$$|\mathbf{grad}(u - u_h)(x)| \leqslant Ch |\log h| \|D^2 u\|_\infty,$$

où C ne dépend pas de $x \in \overline{\Omega}$ ni de h.

5. Méthodes d'éléments finis non conformes, "crimes variationnels".

On peut encore calculer une approximation de Ritz-Galerkin de la solution de $a(u, v) = \varphi(v), \forall v \in U$ par

$$?u_h \in U_h : \qquad a(u_h, v) = \varphi(v), \qquad \forall v \in U_h,$$

même si $U_h \not\subseteq U$! Il faut d'abord vérifier que la matrice de rigidité est non singulière, ce qui est le cas si a est encore définie positive sur U_h.

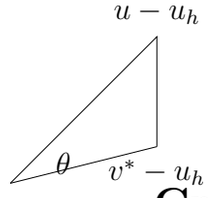
Exemples. ([Thomée], p. 26) : Rectangles de Wilson, $V_k = \mathscr{P}_2$ (dimension 6), $F_i(v)$ = valeurs aux sommets et valeurs des deux dérivées secondes $\partial^2 v/\partial x^2$ et $\partial^2 v/\partial y^2$ (constantes dans chaque rectangle e_k). Alors, si $v = 0$ sur $\Gamma = \partial\Omega$, la forme $a(u, v) = \int_\Omega \text{grad}u \cdot \text{grad}v \; dxdy$ est bien définie positive : $a(v, v) = 0 \Rightarrow v$ est constante dans chaque e_k, la même constante partout puisque v est bien définie aux sommets, donc $v = 0$ dans tout Ω puisque $v = 0$ sur la frontière (N.B., on évalue l'intégrale sur Ω comme somme des intégrales sur les e_k).

Le lemme de Céa devient : pour a bilinéaire symétrique définie positive,

$$\sqrt{a(u-u_h,u-u_h)} \leqslant \inf_{v\in U_h} \sqrt{a(u-v,u-v)} + \sup_{\substack{w\in U_h\\ w\neq 0}} \frac{|a(u,w)-\varphi(w)|}{\sqrt{a(w,w)}}$$

(Strang). En effet ([Ciarlet], sect. 9), $a(u-u_h,w) = a(u,w)-\varphi(w)$ pour tout $w \in U_h$ puisque u_h est la solution du problème de Ritz-Galerkin dans U_h. Donc,

$$\sup_{\substack{w\in U_h\\ w\neq 0}} \frac{|a(u,w)-\varphi(w)|}{\sqrt{a(w,w)}} = \sup_{\substack{w\in U_h\\ w\neq 0}} \frac{|a(u-u_h,w)|}{\sqrt{a(w,w)}}$$



ce dernier supremum est atteint quand w est aligné sur la projection $a-$orthogonale de $u-u_h$ dans U_h, et vaut $\|u-u_h\|_a \cos\theta = \sqrt{\|u-u_h\|_a^2 - \|u-v^*\|_a^2}$, où $\|.\|_a = \sqrt{a(.,.)}$.

Séminaire GSNA

Crime, concomitant et châtiment.

pour des opérateurs de type laplacien.

Alphonse Magnus,
Institut de Mathématique Pure et Appliquée,
Université Catholique de Louvain,
Chemin du Cyclotron,2,
B-1348 Louvain-la-Neuve
(Belgium)
(0)(10)473157 , magnus@anma.ucl.ac.be , http ://www.math.ucl.ac.be/~magnus/

This version : April 2004 (incomplete and unfinished)

l'organisation du crime en Amérique rapporte quarante billions de dollars. Revenu d'autant plus lucratif que la Mafia dépense très peu en frais de bureaux.
Woody Allen

Abstract :

Le ***concomitant bilinéaire*** $B(u,v)$ d'une forme différentielle L est l'expression qui apparaît dans les termes aux limites provenant de la succession d'intégrations par parties joignant $\int_D vLu\,dx$ à $\int_D uL^*v\,dx$, par exemple

$$\int_a^b u''(x)v(x)\,dx = (B(u,v))(b) - (B(u,v))(a) - \int_a^b u(x)v''(x)\,dx,$$

avec $B(u,v) = u'v - uv'$. Pour le laplacien :

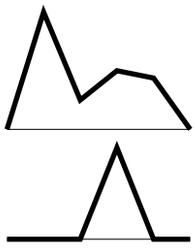
$$\int_D v(x)\Delta u(x)\,dx = \int_{\partial D} (B(u,v))(x)\cdot\vec{n}\,dS - \int_D u(x)\Delta v(x)\,dx,$$

avec $B(u,v) = v\nabla u - u\nabla v$.

1. Méthodes non criminelles.

Il n'y a pas crime lorsque la traduction de le forme bilinéaire $a(u,v)$ est distributionnellement correcte.

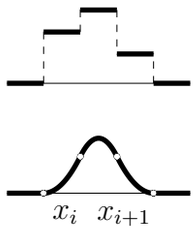
1.1. Galerkin vrai.



u et v sont dans le même espace, un espace de fonctions de dérivées de carré intégrable. On prend donc des fonctions continues pour éviter des produits de δ de Dirac.
Les produits d'intégrations par parties aux points de jonction intérieurs se réduisent (si u' est continue) :

$$\int_{x_i}^{x_{i+1}} u''(x)v(x)\,dx = (u'v)(x_{i+1}) - (u'v)(x_i) - \int_{x_i}^{x_{i+1}} u'(x)v'(x)\,dx.$$

1.2. Petrov.



On peut très bien prendre u discontinu, si cela est compensé par v très continu, mais on risque de casser la symétrie.

Par exemple, u constant par morceaux, et v spline quadratique $\in \mathscr{C}^1$. L'intégrale de $u'v'$ est la somme des sauts de u multipliés par les valeurs bien définies de v' aux points de jonction :

$$\int_a^b u'v'\,dx = \sum [u]_{x_i} v'(x_i)$$

Si les inconnues sont les $u_i = u(x_i)_+$, les équations sont

$$[u]_i v_i'(x_i) + [u]_{i+1} v'i(x_{i+1}) = -u_{i-1}v_i'(x_i) + u_i(v_i'(x_i) - v_i'(x_{i+1})) + u_{i+1}v_i'(x_{i+1}) = \int_{x_{i-1}}^{x_{i+1}} f(x)v_i(x)\,dx.$$

Ça ressemble déjà un peu à un concomitant

1.3. Forme faible distributionnelle.

Bien sûr ! Si on accentue encore la disymétrie, en portant toutes les dérivations des fonctions d'essai u vers les fonctions test v :

$$\int_{x_i}^{x_{i+1}} u''(x)v(x)\,dx = (u'v)(x_{i+1}) - (u'v)(x_i) - (uv')(x_{i+1}) + (uv')(x_i) + \int_{x_i}^{x_{i+1}} u(x)v''(x)\,dx.$$

D'où une somme résiduelle $\sum_i [uv' - u'v]_i$.

1.4 Un peu d'algèbre : saut d'un produit.

$$\begin{aligned}
[ab]_i &= a_i^+ b_i^+ - a_i^- b_i^- \\
&= (a_i^+ - a_i^-) b_i^+ + a_i^- (b_i^+ - b_i^-) \\
&= [a]_i \left(\{b\}_i + \frac{[b]_i}{2} \right) + [b]_i \left(\{a\}_i - \frac{[a]_i}{2} \right) \\
&= [a]_i \{b\}_i + [b]_i \{a\}_i.
\end{aligned} \tag{46}$$

2. Crimes sans concomitant.

La forme bilinéaire est sommée sur les éléments. On ne fait pas intervenir des valeurs de discontinuités aux frontières intérieures, mais les contraintes doivent être telles que la matrice de rigidité soit non singulière (Ciarlet, Strang, années 70). C'est toujours du Galerkin, on garde la symétrie.

cf. Thomée [Th], pp.26-27 :

“ In some situations one may want to use finite element spaces S_h defined by piecewise polynomial approximating functions on a partition $\mathscr{T}_h$ of Ω which are not continuous across interelement boundaries, so called nonconforming elements. Assuming Ω polygonal so that it is exactly a union of elements τ, one may introduce a discrete bilinear form by $D_h(\psi,\chi) = \sum_{\tau\in\mathscr{T}_h}(\nabla\psi,\nabla\chi)_\tau$. Provided S_h is such that $\|\chi\|_{1,h} = D_h(\chi,\chi)^{1/2}$ is a norm on S_h, a unique nonconforming finite element solution u_h of $-\Delta u = f$ in Ω with $u = 0$ on $\partial\Omega$ is now defined by $D_h(u_h,\chi) = (f,\chi)$ for $\chi \in S_h$, and it was shown in Strang (1972) that

$$\|u_h - u\|_{1,h} \leqslant C \inf_{\chi\in S_h} \|u-\chi\|_{1,h} + C \sup_{\chi\in S_h} \frac{|D_h(u,\chi) - (f,\chi)|}{\|\chi\|_{1,h}}, \tag{5.12}$$

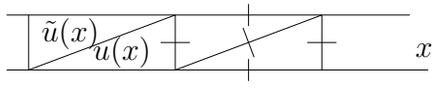
As an example, consider an axes parallel rectangular domain, partitioned into smaller such rectangles with longest edge $\leqslant h$, and let S_h be piecewise quadratics which are continuous at the corners of the partition. Then $\|.\|_{1,h}$ is a norm on S_h. In Wilson's rectangle, the six parameters involved on each small rectangle are determined by the values at the corners plus the (constant) values of $\partial^2\chi/\partial x_l^2$, $l = 1,2$. The functions in S_h are not in $C(\overline{\Omega})$ but using (5.12) one may still show $\|u_h - u\|_{1,h} \leqslant C(u)h$.

The analysis above assumes that all inner products are calculated exactly. An analysis where quadrature errors are permitted was also worked out by Strang (1972). For instance, if (f,χ) is replaced by a quadrature formula $(f,\chi)_h$, a term of the form $C\sup_{\chi\in S_h}|(f,\chi) - (f,\chi)_h|/\|\nabla\chi\|$ has to be added to the bound for $\|\nabla(u_h - u)\|$. For example, if the quadrature formula is exact on each element for constants and if $f \in W_q^1(\Omega)$ with $q > 2$, then the $O(h)$ error for $\|\nabla(u_h - u)\|$ is maintained. The situations when curved boundaries, nonconforming elements, or quadrature errors occur, so that the basic assumptions of the variational formulation are not satisfied, are referred to in Strang (1972), as variational crimes. ” fin de citation.

Et Braess [Bra] chap. 3 dit bien que l'analyse des méthodes non conformes conduit aussi aux méthodes de différences et de volume fini !

2.1. Un crime à une dimension.

Méthodes initialement prévues pour des problèmes d'ordre élevé, par exemple plaque avec triangles de Morley (degré 2, valeurs aux sommets et dérivées normales aux milieux des côtés [Bre] § 8.x.



Réduction à une dimension : on prend un long rectangle avec conditions naturelles sur les longs côtés, donc $\partial u/\partial y = 0 \Rightarrow u$ ne dépend plus que de x, ce qui est le but de la manœuvre (Atkinson [Atki] § 8.3).

On se rapproche effectivement d'une méthode aux différences :

Sur le même intervalle (x_{i-1}, x_i), on considère les deux fonctions du second degré

$u(x) = u_{i-1}(x_i - x)/h_{i-1} + u_i(x - x_{i-1})/h_{i-1} + \alpha_i(x - x_{i-1})(x - x_i)$,

$\tilde{u}(x) = u_{i-1}(x_i - x)/h_{i-1} + u_i(x - x_{i-1})/h_{i-1} + \beta_i(x - x_{i-1})(x - x_i)$.

La continuité de la dérivée demande que la valeur de $u'(x_i)$ coïncide avec celle de $\tilde{u}'$ provenant de l'intervalle (x_i, x_{i+1}) :

$$\frac{u_i - u_{i-1}}{h_{i-1}} + \alpha_i h_{i-1} = \frac{u_{i+1} - u_i}{h_i} - \beta_{i+1}h_i.$$

On voit poindre la différence (divisée) seconde $[x_{i-1}, x_i, x_{i+1}]_u := 2\dfrac{\dfrac{u_{i+1}-u_i}{h_i} - \dfrac{u_i - u_{i-1}}{h_{i-1}}}{h_{i-1} + h_i}$ de u !

Pour résoudre le problème de poutre $u'''' = f$, on minimise
$\sum_i \int_{x_{i-1}}^{x_i} \dfrac{(u''(x))^2 + (\tilde{u}''(x))^2}{2} - [u(x) + \tilde{u}(x)]f(x)\, dx$, soit $\sum_i \left[2(\alpha_i^2 + \beta_i^2)h_{i-1} - (u_{i-1} + u_i)f_{i+1/2}h_{i-1}\right]$.

Minimiser $\alpha_i^2 h_{i-1} + \beta_{i+1}^2 h_i$ sachant que $\alpha_i h_{i-1} + \beta_{i+1}h_i = s \Rightarrow \alpha_i = \beta_{i+1} = [x_{i-1}, x_i, x_{i+1}]/2$.

Enfin, $\dfrac{\partial}{\partial u_i}\left[\sum_i \dfrac{1}{2}(h_{i-1} + h_i)[x_{i-1}, x_i, x_{i+1}]^2 - u_i(f_{i+1/2}h_{i-1} + f_{i+3/2}h_i)\right] =$

$$[x_{i-2}, x_{i-1}, x_i]\frac{1}{h_{i-1}} - [x_{i-1}, x_i, x_{i+1}]\left(\frac{1}{h_{i-1}} + \frac{1}{h_i}\right)[x_i, x_{i+1}, x_{i+2}]\frac{1}{h_i} - f_{i+1/2}h_{i-1} - f_{i+3/2}h_i = 0$$

pratiquement un traitement de différences.

3. Convergence par consistance et stabilité.

A partir du moment où on s'écarte de la logique éléments finis, on se rapproche d'une analyse de méthode aux différences : l'opérateur L étant approché par L_h, on apprécie l'écart entre la solution u_h du problème numérique $L_h u_h = f$ et la solution théorique de $Lu = f$ par

$$\begin{aligned} \|u_h - u\| &= \|L_h^{-1} f - u\| \\ &= \|L_h^{-1}[f - L_h u]\| \\ &\leqslant \underbrace{\|L_h^{-1}\|}_{\substack{\text{stabilité} \\ \text{numérique}}} \underbrace{\|(L - L_h)u\|}_{\text{consistance}} \end{aligned} \tag{47}$$

4. L'article de Süli *et al.*.

4.1. Le schéma.

Pour le problème $-u'' = f$:

on admet pour u et v des polynômes par morceaux sans aucune contrainte, mais une espèce de concomitant est introduit :

$$\int_a^b f(x)v(x)\, dx = \sum_i \int_{x_{i-1}}^{x_i} u'v'\, dx + [v]_i\{u'\}_i - [u]_i\{v'\}_i$$

à quoi il est prudent d'ajouter un terme de pénalité-régularisation $\sigma[u][v]$.

Allons bon. Naïvement, on prendrait :

$$-\int_{x_i}^{x_{i+1}} u" v = \int_{x_i}^{x_{i+1}} u'v' - (u'v)(x_{i+1}) + (u'v)(x_i),$$

soit une somme $\sum_i [u'v]_i$, qui se développe, selon (46), en $[u']_i\{v\}_i + [v]_i\{u'\}_i$. Si u est une solution $\mathscr{C}^1$, il n'y a pas de saut de u', reste donc $[v]_i\{u'\}_i$.

Pourquoi retirer $[u]_i\{v'\}_i$? Si u' est continue, u l'est encore plus, si j'ose dire, il n'y a donc pas de saut de u non plus, c'est dingue !

4.2. Tentative d'explication.

Ce que nous venons de discuter est ce qui se passe quand on introduit la solution théorique dans le schéma numérique : c'est une discussion de ***consistance*** $L_h u - f = (L_h - L)u$. Nous avons vu que la suppression du terme $[u']_i\{v\}_i$ ne détruit pas la qualité de l'erreur de consistance. Il en est de même pour l'introduction de $-[u]_i\{v'\}_i$. Donc, opération blanche. Mais alors, où est l'intérêt ?

Des termes aux frontières intérieures de type concomitant ***s'annulent*** quand $u = v$. Donc, la forme quadratique $(v, L_h v)$ se confond avec (v, Av), où A est la matrice de rigidité ordinaire d'un problème de Galerkin, matrice symétrique et définie positive, donc, en adaptant la propriété de ***coercivité*** d'un problème bien posé de type laplacien :

$$c\|v\|^2 \leqslant (v, L_h v),$$

References.

Atki K. Atkinson, W. Han, *Theoretical Numerical Analysis, A Functional Analysis Framework*, Springer-Verlag, 2001. Lectures in Mathematics - ETH Zürich

Bra D. Braess, *Finite Elements. Theory, Fast Solvers and Applications in Solid Mechanics* , 2nd edition, Cambridge University Press 2001.

Bre Brenner, Susanne C., Scott, L. Ridgeway, *The Mathematical Theory of Finite Element Methods (Texts in Applied Mathematics*, Vol **15**), 2nd Ed 385 pages, Springer-Verlag New York Inc., Jul. 2002.

Nica S. Nicaise, *Analyse numérique et équations aux dérivées partielles. Cours et problèmes résolus*, Dunod, Paris, 2000.

Po K. Polthier, Unstable Periodic Discrete Minimal Surfaces, *in* : *Nonlinear Partial Differential Equations*, S. Hildebrandt and H. Karcher (Eds.) Springer Verlag (2002), pp. 127-143.

`http ://www-sfb288.math.tu-berlin.de/~konrad/articles/alignment/alignmentFinal.pdf`

`http ://www.zib.de/polthier/articles/alignment/alignmentFinal.pdf`

Suli Süli, Endre ; Schwab, Christoph ; Houston, Paul : *hp*-DGFEM for partial differential equations with nonnegative characteristic form. *in* Cockburn, Bernardo (ed.) et al., *Discontinuous Galerkin methods. Theory, computation and applications. 1st international symposium on DGM, Newport, RI, USA, May 24-26, 1999*. Berlin : Springer. Lect. Notes Comput. Sci. Eng. **11**, 221-230 (2000). preprint in `http ://caeconsulting.be/gsna/protected/NA-99-02.ps.gz`

Th V. Thomée, From finite differences to finite elements. A short history of numerical analysis of partial differential equations, *J. Comp. Appl. Math.* **128** (2001) 1-54.

Chapitre 4

Méthodes numériques d'obtention de l'approximation de Ritz.

(cf. Strang, chap. 5)

1. Élaboration et résolution des équations. Conditionnement.

$a(u_h, v_h) = \varphi(v_h), \forall v_h \in U_h$ devient $A_h x = b$, avec $a_{i,j} = a(u_i, u_j)$, l'inconnue $x_j =$ coefficient de u_h dans la base de Lagrange $\{u_i\}$, $b_i = \varphi(u_i)$.

Si a est bilinéaire symétrique coercive, A_h est symétrique définie positive. De plus, dans les problèmes d'éléments finis, A_h est creuse (et bande, mais de largeur de bande d assez importante).

Résolution : factorisation de Cholesky (1875-1918) $A_h = L_h L_h^T$.

Si $j < i$, tous les $\ell_{j,k}$ connus, $k = 1, \ldots, j$, et $\ell_{i,1}, \ldots \ell_{i,j-1}$ connus $\Rightarrow a_{i,j} = \sum_{k=1}^{j} \ell_{i,k}\ell_{j,k}$ donne $\ell_{i,j}$;

si $j = i$: $\ell_{i,i} = \left[a_{i,i} - \sum_{k=1}^{i-1} \ell_{i,k}^2 \right]^{1/2}$.

Matrice bande : si $a_{i,j} = 0$ quand $j < i - d$, $\ell_{i,j} = 0$ quand $j < i - d$.
Donc, nombre d'opérations pour une matrice bande d'ordre N de largeur de bande $d \approx$ const. Nd^2.
Problème de Laplacien à 2 dimensions avec $h \approx 0.01 : N \approx 10^4$, $d \approx 100 \Rightarrow \approx 10^8$ opérations.

$A_h =$ matrice de rigidité ; $M_h(m_{i,j} = (u_i, u_j)_0) =$ matrice de masse.
Soit μ_h la plus petite valeur propre (nécessairement > 0) de M_h.

Si a est bilinéaire symétrique coercive de constante de coercivité $c > 0$, A_h est symétrique définie positive de plus petite valeur propre $\lambda_{min} \geqslant c\mu_h$:
$x^T A_h x = a(u_h, u_h) \geqslant c\|u_h\|_m^2 \geqslant c\|u_h\|_0^2 = c(u_h, u_h)_0 = cx^T M_h x \geqslant c\mu_h x^T x$.

Conditionnement d'un opérateur : mesure relative de la sensibilité de la solution d'un problème de résolution à des perturbations sur les données :

$$\kappa(A_h) = \max_{\substack{A_h x = b \\ A_h(x+\delta x) = b + \delta b}} \frac{\|\delta x\|/\|x\|}{\|\delta b\|/\|b\|} = \|A_h\| \, \|A_h^{-1}\|.$$

Avec la norme euclidienne des vecteurs, $\kappa(A_h) = \lambda_{\max}(A_h)/\lambda_{\min}(A_h)$.

Pour un problème aux limites d'ordre $2m$ (formulation variationnelle semi-faible faisant intervenir des dérivées d'ordre $\leqslant m$), on peut montrer (Strang, p.209) que

$$\kappa(A_h) \leqslant \text{ constante } h_{\min}^{-2m}.$$

On verra quelques illustrations.

Importance du conditionnement en méthodes itératives : une méthode utilisant une multiplication par A_h à chaque pas aboutit nécessairement à une $n^{\text{ème}}$ approximation de la forme

$x = R_n(A_h)x_0 + Q_n(A_h)b$, avec $b = A_h x_0 \Rightarrow x = x_0 : R_n(A_h) = I - A_h Q_n(A_h) : R_n(0) = 1$. La norme de l'erreur dépend de $\|R_n(A_h)\| = \max_j |R_n(\lambda_j)|$. Elle est d'autant plus grande que $\kappa(A_h)$ est grand.

Exemple : méthode des gradients conjugués, on minimise au $(n+1)^{\text{ème}}$ pas $J_h(x_{n+1}) = \frac{1}{2}x_{n+1}^T a_h x_{n+1} - x_{n+1}^T b$ sur

$$x_{n+1} = x_n + p_n(x_n - x_{n-1}) + q_n(b - A_h x_n),$$

$(x_{-1} = 0, p_0 = 0)$. On montre que $\|R_n(A_h)\| \leqslant \text{const.} \left(1 - \frac{1}{\sqrt{\kappa(A_h)}}\right)^n$.

Problème de Laplacien à 2 dimensions avec $h \approx 0.01 : N \approx 10^4$, $\approx 10^4$ opérations par itération, $\kappa \approx 10^4$, il faut ≈ 100 pas pour gagner une décimale $\Rightarrow$ moins de 10^7 opérations pour avoir 3 ou 4 décimales correctes.

Préconditionnement : si on sait résoudre efficacement $B_h C_h y = c$, $\forall c$, on examine $(B_h^{-1} A_h C_h^{-1})(C_h x) = B_h^{-1} b$, avec $\kappa(B_h^{-1} A_h C_h^{-1})$ raisonnablement petit.

L'importance du conditionnement apparaît dès qu'il s'agit d'estimer la qualité d'une solution approchée d'un problème :

Au $k^{\text{ème}}$ pas de l'itération, on dispose d'une estimation $x^{(k)}$ de la solution de $Ax = b$. Le degré de validité de $x^{(k)}$ ne peut s'évaluer que par l'examen du résidu $Ax^{(k)} - b$:

$$x - x^{(k)} = A^{-1}(b - Ax^{(k)}).$$

On ne peut donc comparer l'amplitude (inconnue) de l'erreur à celle du résidu que via la norme de A^{-1} :

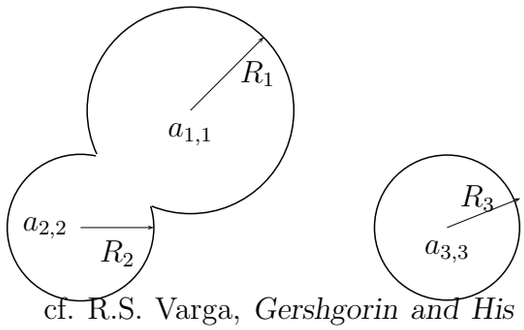
$$\|x - x^{(k)}\| \leqslant \|A^{-1}\| \, \|b - Ax^{(k)}\|.$$

Bien entendu, on a aussi $\|b - Ax^{(k)}\| = \|A(x - x^{(k)}\| \leqslant \|A\| \, \|x - x^{(k)}\|$.

Au fil des itérations, on a donc

$$\frac{\|x - x^{(k)}\|}{\|x - x^{(0)}\|} \leqslant \|A^{-1}\| \, \|A\| \, \frac{\|b - Ax^{(k)}\|}{\|b - Ax^{(0)}\|}.$$

Théorème de Hadamard et disques de Gershgorin. Théorème (Hadamard). *Une matrice carrée A vérifiant $|a_{i,i}| > \sum_{j \neq i} |a_{i,j}|$, $i = 1, 2, \ldots, n$ (***matrice fortement dominante [par lignes]***) est non singulière.* En effet, si A était singulière, il existerait un vecteur $v \neq 0$ tel que $Av = 0$. Soit v_m une des composantes de plus grande valeur absolue de v : $|v_m| \geqslant |v_i|, i = 1, \ldots, n$. La $m^{\text{ème}}$ équation $a_{m,m} v_m = -\sum_{\substack{j=1 \\ j \neq m}}^{n} a_{m,j} v_j$ de $Av = 0$ est alors impossible puisque $\left|\sum_{j \neq m} a_{m,j} v_j\right| \leqslant |v_m| \sum_{j \neq m} |a_{m,j}| < |v_m a_{m,m}|$. □



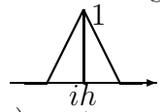
Théorème (Gershgorin). *Les valeurs propres d'une matrice carrée A sont toutes situées dans la réunion des disques fermés de centres $a_{i,i}$ et de rayons $R_i = \sum_{j\neq i} |a_{i,j}|$.*

En effet, $\lambda I - A$ est non singulière si λ est hors de chaque disque : $|\lambda - a_{i,i}| > R_i$, $\forall i \Rightarrow \det(\lambda I - A) \neq 0 \Rightarrow \lambda$ n'est pas valeur propre.

cf. R.S. Varga, *Gershgorin and His Circles*, Springer Series in Computational Mathematics, Band **36**, 2004.

Exemples d'estimation de nombre de condition de la matrice de rigidité :

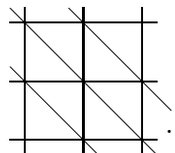
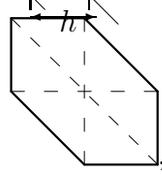
(1) Problème unidimensionnel $a(u,v) := \int_0^L u'(x)v'(x)\,dx$ avec u et v nuls en 0 et L.

U_h = fonctions continues linéaires par morceaux sur les intervalles de longueur $h = L/N$. L'élément b_i de la base de Lagrange n'est non nul que sur un support formé de deux intervalles : 

. On voit que seuls les trois éléments $a(b_i, b_{i-1})$, $a(b_i, b_i)$ et $a(b_i, b_{i+1})$ sont non nuls sur la $i^{\text{ème}}$ ligne de A. Les valeurs sont $-1/h$, $2/h$ et $-1/h$ (les dérivées sont $\pm 1/h$, à multiplier et intégrer sur un ou deux intervalles de longueur h).

Gershgorin : $\lambda_{\max}(A) \leqslant 4/h$, mais on ne peut rien conclure sur $\lambda_{\min}$. Passage par la matrice de masse M : les intégrales $\int b_i b_j$ sont maintenant $h/6$, $2h/3$ et $h/6$, d'où $\lambda_{\min}(M) \geqslant 2h/3 - 2h/6 = h/3$ et

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \leqslant \frac{4/h}{ch/3} = O(h^{-2}).$$

(2) Problème bidimensionnel $a(u,v) = \int_\Omega \mathbf{grad}\, u \cdot \mathbf{grad}\, v\; dxdy$ sur un carré. U_h = espace de Courant sur les triangles 
. . .Le support d'une fonction b_i de la base de Lagrange est un hexagone 

, où le gradient vaut $(\pm 1/h, 0)$, $(0, \pm 1/h)$ ou $(\pm 1/h, \pm 1/h)$. Les seules intégrales $a(b_i, b_j) = \int_\Omega \mathbf{grad}\, b_i \cdot \mathbf{grad}\, b_j$ susceptibles d'être non nulles correspondent aux 6 fonctions b_j de support centré en un sommet de l'hexagone de b_i. On obtient une intégrale $2(-1/h^2)$ fois l'aire d'un petit triangle $= -1$ dans les 4 cas où l'intersection est un quadrilatère non carré ;

une intégrale nulle dans les deux cas où l'intersection est un carré (les gradients sont orthogonaux). Enfin, $a(b_i, b_i) = (4 \times 1/h^2 + 2 \times 2/h^2)h^2/2 = 4$.

On a encore une fois |élément diagonal| = somme des valeurs absolues des éléments non diagonaux. C'est normal : si $a(u, v) = 0$ sur les constantes, somme des éléments d'une ligne $= a\left(b_i, \sum_{j=1}^{N} b_j\right) = a(b_i, 1) = 0$.

Pour estimer la plus petite valeur propre de A, on repasse par la matrice de masse M. On trouve $m_{i,i} = h^2/2$ et les 6 éléments non diagonaux non nuls tous égaux à $h^2/12$. Aïe. La différence de Gershgorin ne suffit plus à estimer la plus petite valeur propre non plus. Par la théorie des matrices de Toeplitz [Strang], on peut cependant établir $\lambda_{\min}(M) \geqslant \min_{\theta_1,\theta_2} h^2/2 + (h^2/12)(e^{i\theta_1} + e^{i\theta_2} + e^{-i\theta_1+i\theta_2} + e^{-i\theta_1} + e^{-i\theta_2} + e^{i\theta_1-i\theta_2}) = h^$
d'où

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \leqslant \frac{8}{ch^2/4} = O(h^{-2}).$$

2. Un exemple de traitement en matlab.

On considère le problème de Poisson $-\Delta u = 1$ sur un secteur d'angle $3\pi/2$ avec $u = 0$ sur le bord.

On utilise la toolbox "pde[1]" de `matlab` :

```
% Partial Differential Equation Toolbox.
% Version 1.0.1   15-Nov-1996
%
% New Features.
%   Readme      - Important release information about the PDE Toolbox
%                 (double click on Readme, or type "whatsnew directoryname",
%                  i.e. "whatsnew pde" to display this file).
%
% PDE algorithms.
%   adaptmesh   - Adaptive mesh generation and PDE solution.
%   assema      - Assemble area integral contributions.
%   assemb      - Assemble boundary condition contributions.
%   assempde    - Assemble a PDE problem.
%   hyperbolic  - Solve hyperbolic problem.
%   parabolic   - Solve parabolic problem.
%   pdeeig      - Solve eigenvalue PDE problem.
%   pdenonlin   - Solve nonlinear PDE problem.
%   poisolv     - Fast solution of Poisson's equation on a rectangular grid.
%
% User interface algorithms and utilities.
%   pdecirc     - Draw circle.
%   pdeellip    - Draw ellipse.
%   pdemdlcv    - Convert Matlab 4.2c Model M-files for use with Matlab 5.
%   pdepoly     - Draw polygon.
%   pderect     - Draw rectangle.
%   pdetool     - PDE Toolbox graphical user interface (GUI).
%
% Geometry algorithms.
%   csgchk      - Check validity of Geometry Description matrix.
%   csgdel      - Delete borders between minimal regions.
%   decsg       - Decompose Constructive Solid Geometry into minimal regions.
```

[1] Cf. `http ://www.comsol.se/Matlab/pde_tb.html`,
voir aussi `http ://www.indiana.edu/~statmath/math/matlab/pde.html`

```
%   initmesh    - Build an initial triangular mesh.
%   jigglemesh  - Jiggle internal points of a triangular mesh.
%   pdearcl     - Interpolation between parametric representation and arc length.
%   poimesh     - Make regular mesh on a rectangular geometry.
%   refinemesh  - Refines a triangular mesh.
%   wbound      - Write boundary condition specification data file.
%   wgeom       - Write geometry specification data file.
%
% Plot functions.
%   pdecont     - Shorthand command for contour plot.
%   pdegplot    - Plot a PDE geometry.
%   pdemesh     - Plot a PDE triangular mesh.
%   pdeplot     - Generic PDE Toolbox plot function.
%   pdesurf     - Shorthand command for surface plot.
%
% Utility algorithms.
%   dst         - Discrete sine transform.
%   idst        - Inverse discrete sine transform.
%   pdeadgsc    - Pick bad triangles using a relative tolerance criterion.
%   pdeadworst  - Pick bad triangles relative to the worst value.
%   pdecgrad    - Compute the flux of a PDE solution.
%   pdeent      - Indices of triangles neighboring a given set of triangles.
%   pdegrad     - Compute the gradient of the PDE solution.
%   pdeintrp    - Interpolate function values to triangle midpoints.
%   pdejmps     - Error estimates for adaption.
%   pdeprtni    - Interpolate function values to mesh nodes.
%   pdesde      - Indices of edges adjacent to a set of subdomains.
%   pdesdp      - Indices of points in a set of subdomains.
%   pdesdt      - Indices of triangles in a set of subdomains.
%   pdesmech    - Compute structural mechanics tensor functions.
%   pdetrg      - Triangle geometry data.
%   pdetriq     - Measure the quality of mesh triangles.
%   poiasma     - Boundary point matrix contributions for Poisson's equation.
%   poicalc     - Fast solution of Poisson's equation on a rectangular grid.
%   poiindex    - Indices of points in canonical ordering for rectangular grid.
%   sptarn      - Solve generalized sparse eigenvalue problem.
%   tri2grid    - Interpolate from PDE triangular mesh to rectangular mesh.
%
% User defined algorithms.
%   pdebound    - Boundary M-file.
%   pdegeom     - Geometry M-file.
%
% Demonstrations.
%   pdedemo1    - Exact solution of Poisson's equation on unit disk.
%   pdedemo2    - Solve Helmholtz's equation and study the reflected waves.
%   pdedemo3    - Solve a minimal surface problem.
%   pdedemo4    - Solve PDE problem using subdomain decomposition.
%   pdedemo5    - Solve a parabolic PDE (the heat equation).
%   pdedemo6    - Solve a hyperbolic PDE (the wave equation).
%   pdedemo7    - Adaptive solution with point source.
%   pdedemo8    - Solve Poisson's equation on rectangular grid.

% Copyright (c) 1994-96 by The MathWorks, Inc.
%       $Revision: 1.6 $  $Date: 1996/11/01 20:31:16 $
```

Programme pdesect.m :

```
%PDEsect         FEM Solution sur secteur

echo on
clc
%       Solve Poisson's equation
%        -div(grad(u))=1
%       sur secteur      with u=0 on the boundary.
```

```
%       Compare with exact solution.
pause % Strike any key to continue.
clc

%       Problem definition
g='secteur'; % The unit circle
b='sectb1'; % 0 on the boundary
c=1;
a=0;
f=1;
%       A coarse initial mesh
[p,e,t]=initmesh(g,'hmax',1);
pause % Strike any key to continue.
clc
%       Do iterative regular refinement until the error is acceptable
error=[]; er=1;
%while er > 0.001,...
for cnt=1:3,
p
e
t
  [matK,phi]=assempde(b,p,e,t,c,a,f);...
matK
phi'
  u=matK\phi;...
% exact=(1-p(1,:).^2-p(2,:).^2)'/4;...
% er=norm(u-exact,'inf');...
% error=[error er];...
% fprintf('Error: %e. Number of nodes: %d\n',er,size(p,2));...

szt=size(t,2)  ,   szu=size(u,1)    ,  u'

 for cntt=1:szt,
   for cntp=1:3,
    intp=t(cntp,cntt);xp(cntp)=p(1,intp);yp(cntp)=p(2,intp);
  xp1(cntp)=3*(2+xp(cntp)-yp(cntp));yp1(cntp)=1.4*(2+xp(cntp)+yp(cntp));
                           yp2(cntp)=yp1(cntp)+2+20*u(intp);
   end
   fprintf('%cdrawline(%f,%f)(%f,%f)(%f,%f)(%f,%f)\n',...
         92,xp1(1),yp1(1),xp1(2),yp1(2),xp1(3),yp1(3),xp1(1),yp1(1));
   if xp(1)+yp(1) >= 0,
     fprintf('%cthicklines%cdrawline(%f,%f)(%f,%f)(%f,%f)(%f,%f)%cthinlines\n',...
         92,92,xp1(1),yp2(1),xp1(2),yp2(2),xp1(3),yp2(3),xp1(1),yp2(1),92);
   end
  end
  [p,e,t]=refinem(g,p,e,t);...
end

%       The solution
%pdesurf(p,t,u); pause % Press any key after plot

% The error
%pdesurf(p,t,u-exact);

pause % Strike any key to end.

echo off
```

La routine principale de résolution est assempde.m :

```
function [K,M,F,Q,G,H,R]=assempde(b,p,e,t,c,a,f,u,time,sdl)
%ASSEMPDE Assemble the stiffness matrix and right hand side a PDE problem.
%
%       U=ASSEMPDE(B,P,E,T,C,A,F) assembles and solves the PDE problem
%       -div(c*grad(u))+a*u=f, on a mesh described by P, E, and T,
```

```
%       with boundary conditions given by the function name B.
%       It eliminates the Dirichlet boundary conditions from the
%       system of linear equations when solving for U.
%
%       [K,F1]=ASSEMPDE(B,P,E,T,C,A,F) gives assembled matrices by
%       approximating the Dirichlet boundary condition with stiff springs.
%       K and F are the stiffness matrix and right-hand side vector
%       respectively. The solution to the FEM formulation of the PDE
%       problem is u=K\F1.
%
%       [K,F1,B1,UD]=ASSEMPDE(B,P,E,T,C,A,F) assembles the PDE problem by
%       eliminating the Dirichlet boundary conditions from the sytem of
%       linear equations. UN=K\F1 returns the solution on the non-Dirichlet
%       points.  The solution to the full PDE problem can be obtained by
%       the MATLAB command U=B1*UN+UD.
%
%       [K,M,F1,Q,G,H,R]=ASSEMPDE(B,P,E,T,C,A,F) gives a split
%       representation of the PDE problem.
%
%       U=ASSEMPDE(K,M,F,Q,G,H,R) collapses the split representation
%       into the single matrix/vector form, and then solves the PDE
%       problem by eliminating the Dirichlet boundary conditions
%       from the system of linear equations.
%
%       [K1,F1]=ASSEMPDE(K,M,F,Q,G,H,R) collapses the split representation
%       into the single matrix/vector form, by fixing the Dirichlet boundary
%       condition with large spring constants.
%
%       [K1,F1,B,UD]=ASSEMPDE(K,M,F,Q,G,H,R) collapses the split representation
%       into the single matrix/vector form by eliminating the Dirichlet
%       boundary conditions from the system of linear equations.
%
%       The geometry of the PDE problem is given by the triangle
%       data P, E, and T.  See either INITMESH or PDEGEOM for details.
%
%       For the scalar case the solution u is represented as a column
%       vector of solution values at the corresponding node points from
%       P.  For a system of dimension N with NP node points, the first
%       NP values of U describe the first component of u, the
%       following NP values of U describe the second component of u,
%       and so on.  Thus, the components of u are placed in the vector
%       U as N blocks of node point values.
%
%       B describes the boundary conditions of the PDE problem.  B
%       can either be a Boundary Condition Matrix or the name of Boundary
%       M-file. See PDEBOUND for details.
%
%   ...
%
%       PDE COEFFICIENTS FOR SCALAR CASE
%
%       The coefficients c, a and f of the PDE problem can
%       be given in a wide variety of ways:
%       - A constant.
%       - A row vector of representing values at the triangle centers
%       of mass. A MATLAB text expression for computing coefficient values
%       at the triangle centers of mass. The expression is evaluated in a
%       context where the variables X, Y, SD, U, UX, UY, and T are row
%       vectors representing values at the triangle centers of mass.
%       (T is a scalar). The row vectors contain x- and y-coordinates,
%       subdomain label, solution with x and y derivatives and
%       time. U, UX,  and UY can only be used if U0 have been  passed to
%       ASSEMPDE. The same applies to the scalar T, which is passed to
%       ASSEMPDE as TIME.
%       - A sequence of MATLAB text expressions separated by exclamation
```

```
%       marks !.  The syntax of each of the text expressions must
%       be according to the above item.  The number of expressions in
%       the sequence must equal the number of subdomain labels in the
%       triangle list t.
%       - The name of a user-defined MATLAB function that accepts the
%       arguments  (P,T,U,TIME). P and T are mesh data, U is the U0
%       input argument and T is the TIME input argument to ASSEMPDE.
%       U will be the empty matrix, and TIME will be NaN if the
%       corresponding parameter was not passed to ASSEMPDE.
%
%       If C contains two rows with data according to any of the above
%       items, they are the c(1,1), and c(2,2), elements of a 2-by-2
%       diagonal matrix. If c contains three rows with data according to
%       any of the above items, they are the c(1,1),  c(1,2),
%       and c(2,2) elements of a 2-by-2 symmetric matrix. If C contains four
%       rows with data according to any of the above items, they are the
%       c(1,1), c(2,1), c(1,2), and c(2,2) elements of a 2-by-2 matrix.
%
%       PDE COEFFICIENTS FOR SYSTEM CASE
%  ...
```

L'utilisateur doit créer un fichier spécifiant Ω à partir de représentations paramétriques de parties de la frontière. Voici secteur.m (pompé sur un fichier de démonstration) :

```
function [x,y]=secteur(bs,s)
%CIRCLEG        Gives geometry data for the circleg PDE model
%
%   NE=CIRCLEG gives the number of boundary segment
%
%   D=CIRCLEG(BS) gives a matrix with one column for each boundary segment
%   specified in BS.
%   Row 1 contains the start parameter value.
%   Row 2 contains the end parameter value.
%   Row 3 contains the number of the left hand region.
%   Row 4 contains the number of the right hand region.
%
%   [X,Y]=CIRCLEG(BS,S) gives coordinates of boundary points. BS specifies the
%   boundary segments and S the corresponding parameter values. BS may be
%   a scalar.

nbs=5;

if nargin==0,
  x=nbs; % number of boundary segments
  return
end

d=[  0 0 0 0 0 % start parameter value
     1 1 1 1 1 % end parameter value
     1 1 1 1 1 % left hand region
     0 0 0 0 0 % right hand region
];

bs1=bs(:)';

if find(bs1<1 | bs1>nbs),
  error('Non existent boundary segment number')
end

if nargin==1,
  x=d(:,bs1);
  return
end
```

```
x=zeros(size(s));
y=zeros(size(s));
[m,n]=size(bs);
if m==1 & n==1,
  bs=bs*ones(size(s)); % expand bs
elseif m~=size(s,1) | n~=size(s,2),
  error('bs must be scalar or of same size as s');
end

if ~isempty(s),

% boundary segment 1
ii=find(bs==1);
x(ii)=1*cos((pi/2)*s(ii)-pi);           y(ii)=1*sin((pi/2)*s(ii)-pi);

% boundary segment 2
ii=find(bs==2);
x(ii)=1*cos((pi/2)*s(ii)-(pi/2));       y(ii)=1*sin((pi/2)*s(ii)-(pi/2));

% boundary segment 3
ii=find(bs==3);
x(ii)=1*cos((pi/2)*s(ii));              y(ii)=1*sin((pi/2)*s(ii));

% boundary segment 4
ii=find(bs==4);
x(ii)=0*s(ii);                          y(ii)=1-s(ii);

% boundary segment 5
ii=find(bs==5);
x(ii)=-s(ii);                           y(ii)=0*s(ii);

end
```

Et on exécute :

```
diary on

pdesect

echo on
clc

...
```

Construction d'une première grille, 16 points dont 11 sur la frontière ; triangulation de Ω en 19 triangles :

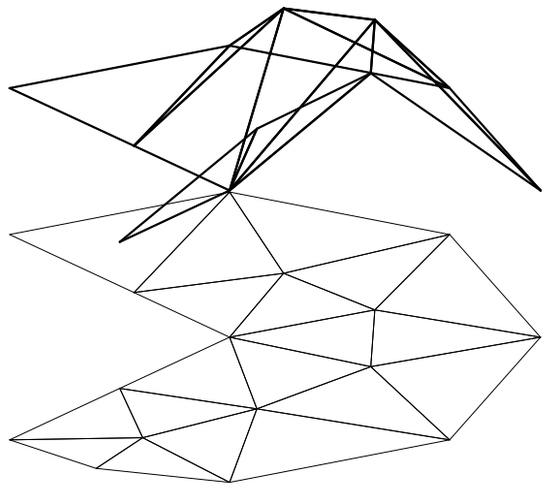
```
p =

  Columns 1 through 7

   -1.0000    0.0000    1.0000    0.0000         0   -0.7071    0.7071
    0.0000   -1.0000         0    1.0000         0   -0.7071   -0.7071

  Columns 8 through 14

    0.7071   -0.5000   -0.9423         0    0.4343   -0.2862    0.4635
    0.7071         0   -0.3349    0.4348    0.1885   -0.4107   -0.1971

  Columns 15 through 16

   -0.6869    0.1772
   -0.2910   -0.4639


e =   Columns 1 through 7
```

```
    1.0000   10.0000    6.0000    2.0000    7.0000    3.0000    8.0000
   10.0000    6.0000    2.0000    7.0000    3.0000    8.0000    4.0000
         0    0.2174    0.5000         0    0.5000         0    0.5000
    0.2174    0.5000    1.0000    0.5000    1.0000    0.5000    1.0000
    1.0000    1.0000    1.0000    2.0000    2.0000    3.0000    3.0000
    1.0000    1.0000    1.0000    1.0000    1.0000    1.0000    1.0000
         0         0         0         0         0         0         0

  Columns 8 through 11

    4.0000   11.0000    5.0000    9.0000
   11.0000    5.0000    9.0000    1.0000
         0    0.5652         0    0.5000
    0.5652    1.0000    0.5000    1.0000
    4.0000    4.0000    5.0000    5.0000
    1.0000    1.0000    1.0000    1.0000
         0         0         0         0


t =   Columns 1 through 12

     8     6    13     7    11     9    10    12    14     3     8     5
     4     2     2     3     5     1     6     5     5     8    11     9
    11    13    16    14    12    15    15    14    16    12    12    13
     1     1     1     1     1     1     1     1     1     1     1     1

  Columns 13 through 19

    13     3     2     1     6     5     7
     9    12     7    10    13    13    14
    15    14    16    15    15    16    16
     1     1     1     1     1     1     1
```

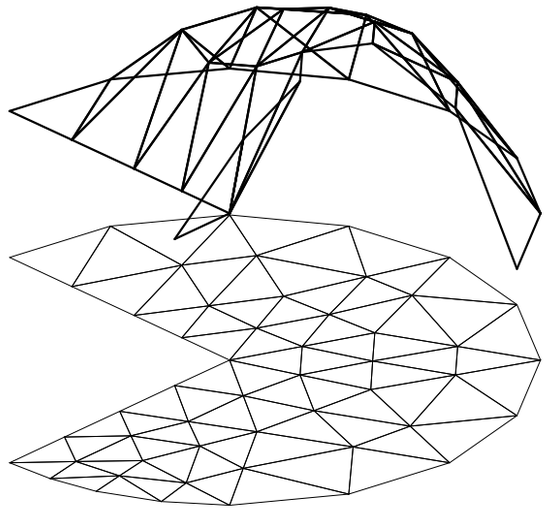
Aperçu de la matrice de rigidité, dont seuls les éléments non nuls sont mis en mémoire dans une structure appropriée. On voit que les éléments diagonaux sont nettement plus grands que les autres : excellent nombre de condition (ça ne durera pas)

```
matK =

 (1,1)  16.7501    (9,1) -0.0899   (10,1) -0.3814  (15,1) -0.3210
 (2,2)  17.2973    (6,2) -0.0814   (7,2)  -0.0558  (13,2) -0.4207  (16,2)  -0.7815
 (3,3)  17.3040    (7,3) -0.0469   (8,3)  -0.0820  (12,3) -0.5578  (14,3)  -0.6595
 (4,4)  16.6761    (8,4) -0.1926   (11,4) -0.5257
 (5,5)  18.5266    (9,5) -0.2618   (11,5) -0.3765  (12,5) -0.5423  (13,5) -0.5200 (14,5)-0.3762 (16,5)-0.4
 (2,6)  -0.0814    (6,6) 17.4381   (10,6) -0.1111  (13,6) -0.6867 (15,6)  -0.6010
 (2,7)  -0.0558    (3,7) -0.0469   (7,7)  17.3158  (14,7) -0.6666 (16,7)  -0.5887
 (3,8)  -0.0820    (4,8) -0.1926   (8,8)  17.2994  (11,8) -0.2229 (12,8)  -0.8441
 (1,9)  -0.0899    (5,9) -0.2618   (9,9)  17.8251  (13,9) -0.4925 (15,9)  -1.0230
(1,10)  -0.3814    (6,10)-0.1111   (10,10)17.8190 (15,10) -1.3686
(4,11)  -0.5257    (5,11)-0.3765   (8,11) -0.2229 (11,11) 17.8714 (12,11) -0.7884
(3,12)  -0.5578    (5,12)-0.5423   (8,12) -0.8441 (11,12) -0.7884 (12,12)  3.8332 (14,12)-1.1007
(2,13)  -0.4207    (5,13)-0.5200   (6,13) -0.6867  (9,13) -0.4925 (13,13)  3.6499 (15,13)-0.6758 (16,13)-0
(3,14)  -0.6595    (5,14)-0.3762   (7,14) -0.6666 (12,14) -1.1007 (14,14)  3.8849 (16,14)-1.0819
(1,15)  -0.3210    (6,15)-0.6010   (9,15) -1.0230 (10,15) -1.3686 (13,15) -0.6758 (15,15) 3.9895
(2,16)  -0.7815    (5,16)-0.4920   (7,16) -0.5887 (13,16) -0.8542 (14,16) -1.0819 (16,16) 3.7983

phi =

  Columns 1 through 7
    0.0389    0.1530    0.1465    0.0666    0.1588    0.1012    0.1433

  Columns 8 through 14
    0.1728    0.0816    0.0322    0.1468    0.2019    0.2183    0.1831
```

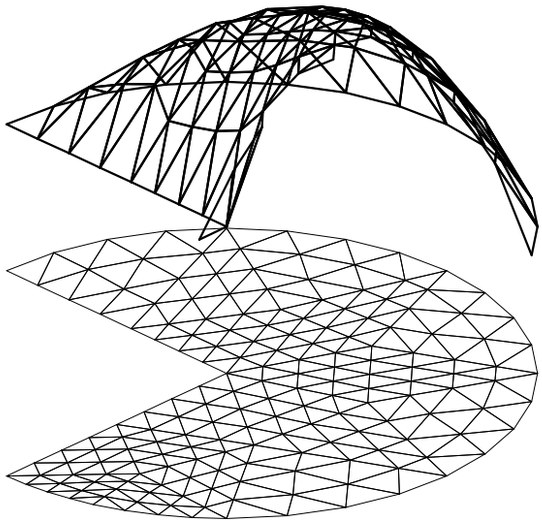
```
  Columns 15 through 16
    0.1078    0.1970


u   =

  Columns 1 through 7
         0         0         0         0         0         0         0

  Columns 8 through 14
         0         0         0         0    0.0808    0.0911    0.0979

  Columns 15 through 16
    0.0425    0.1002
```



Plus fin : 50 points, 22 segments sur le bord, 76 triangles.



Encore plus fin, il y a maintenant 175 points, 44 segments sur le bord, et 304 triangles !



```
%       The solution
%pdesurf(p,t,u); pause % Press any key after plot

% The error
%pdesurf(p,t,u-exact);

pause % Strike any key to end.

echo off
exit

 145048 flops.
```

Exemple de spécification d'un secteur parabolique : `http ://www.math.ucl.ac.be/ magnus/num2/pdepa` , `parab.m` , `bparab.m`

Chapitre 5

Schémas de différences finies : problèmes elliptiques.

1. Opérateurs de prolongement et de restriction.

Les approximations u_h de la solution dans U de $Lu = f$ sont dans des espaces U_h qui ne sont plus nécessairement des parties de U. En différences finies, U_h est l'espace des fonctions définies sur un ensemble (fini) de points $\{x_i^{(h)}\}$ de $\overline{\Omega}$. S'il y a N_h points $x_i^{(h)}$, on a $U_h = \mathbb{R}^{N_h}$.

On ne peut plus comparer directement des éléments de U et de U_h. On considère :

(1) L'opérateur de **prolongement** $p_h : U_h \longrightarrow U$ associe une fonction de U à partir d'un vecteur de U_h.

(2) L'opérateur de **restriction** $r_h : U \longrightarrow U_h$ associe une fonction de U_h à partir d'une fonction de U.

(Ce ne sont d'ailleurs pas de vrais opérateurs de prolongement ou de restriction si $U_h \not\subset U$...mais U_h est en quelque sorte "beaucoup plus petit" que U [la dimension de U_h est le nombre de points $x_i^{(h)}$]. Cf. [Temam, p. 27] pour une présentation de l'approximation externe.).

Par exemple, on prendra $(r_h u)_i = u(x_i^{(h)})$...si des valeurs ponctuelles de u ont un sens (rappelons que ce n'est pas vrai pour un élément quelconque de L^2 ou de certains espaces de Sobolev). Mais des considérations de normes vont faire préférer l'usage de pondérations :

1.1. Normes. On adoptera la norme euclidienne

$$\|[\xi_1, \ldots, \xi_N]\| = \sqrt{\xi_1^2 + \cdots \xi_N^2}$$

des vecteurs de U_h, tout en veillant à ce que la norme de $r_h u$ se rapproche de la norme de u dans $L^2(\Omega)$ à mesure que h se rapproche de 0 : pour une grille régulière, on prendra $r_h u = [\ldots \sqrt{|\omega_h|}\, u(x_i) \ldots]$, où $|\omega_h|$ est l'étendue (longueur, aire, volume) d'une maille.

Pour une grille quelconque, on prendra donc $(r_h u)_i = \sqrt{|\omega_i^{(h)}|}\, u(x_i^{(h)})$, ou $(r_h u)_i = \dfrac{1}{\sqrt{|\omega_i|}} \displaystyle\int_{\omega_i} u(t)\, dt$, où ω_i est un voisinage de $x_i^{(h)}$.

Quant à $p_h u_h$, il sera souvent un interpolant de u_h, y compris un interpolant de type élément fini.

L'interpolation devrait converger : il est bon d'avoir $p_h r_h u \to u$ quand $h \to 0$ pour tout $u \in U$, ou au moins dans une partie dense de U.

2. Approximation d'opérateurs. Consistance.

Les opérateurs différentiels sont approchés par des combinaisons linéaires de valeurs. Ainsi, $\dfrac{u(x_{i+1}) - u(x_i)}{h_i = x_{i+1} - x_i}$ est une approximation vraisemblable de $\partial u / \partial x$.

Si u est suffisamment dérivable, on vérifie l'approximation par développement de Taylor :

$$\frac{u(x_i)-u(x_{i-1}}{h_{i-1}=x_i-x_{i-1}}=\frac{u(x_i)-u(x_i-h_{i-1})}{h_{i-1}}=u'(x_i)-\frac{h_{i-1}}{2}u''(x_i)+\cdots \tag{48}$$

2.1. Définition. *Les opérateurs L_h sont des approximations* ***consistantes*** *de L si*

$$\|L_h r_h u - r_h L u\| \to 0 \text{ quand } h \to 0, \tag{49}$$

pour tout u dans une partie dense de U.

N.B. L_h est opérateur sur $U_h=\mathbb{R}^N$, donc représenté par une ***matrice*** ; la norme dans (49) est la norme euclidienne des vecteurs.

On vérifie la consistance dans $\mathscr{C}^m(\overline{\Omega})$, avec m assez grand, par développement de Taylor de u, par exemple pour la dérivée première plus haut :

Exemple. Dérivée première. Soit U un espace vectoriel de fonctions sur $[0,1]$ s'annulant en $x=0$, et des points $0<x_1<x_2<\cdots<x_N=1$, $h_i=x_{i+1}-x_i$. Si r_h fait correspondre le vecteur $[\xi_1,\ldots,\xi_N]=[\sqrt{h_0}\,u(x_1),\ldots,\sqrt{h_{N-1}}\,u(x_N)]$ à la fonction $u\in U$, on discrétise l'opérateur de dérivation en un opérateur sur $U_h=\mathbb{R}^N$, c'est-à-dire une matrice D_h, en veillant à ce que $D_h r_h u$ soit proche du vecteur des $\sqrt{h_{i-1}}u'(x_i)$, ce qui donne la matrice

$$D_h=\begin{bmatrix} 1/h_0 & & & \\ -1/\sqrt{h_0h_1} & 1/h_1 & & \\ & \ddots & \ddots & \\ & & -1/\sqrt{h_{N-2}h_{N-1}} & 1/h_{N-1} \end{bmatrix}$$

Comment apprécier si cet opérateur approche bien la dérivation ? On applique D_h à $r_h u$, pour obtenir le vecteur des $(u(x_i)/\sqrt{h_{i-1}}-u(x_{i-1})/\sqrt{h_{i-1}}$, $i=1,2,\ldots,N$, à comparer au vecteur des $\sqrt{h_{i-1}}\,u'(x_i)$. On retrouve évidemment l'analyse faite en (48), avec une erreur en $h_{i-1}^{3/2}$ par composante, d'où une norme d'erreur en $\max_i h_i$.

Exemple. Dérivée seconde. Pour vérifier que

$$\frac{\dfrac{u_{i+1}-u_i}{h_i}-\dfrac{u_i-u_{i-1}}{h_{i-1}}}{(h_i+h_{i-1})/2} \tag{50}$$

est une approximation consistante de $u''(x_i)$[1], prenons u dans $\mathscr{C}^3$:

$u(x)=u_i+u_i'(x-x_i)+(u_i''/2)(x-x_i)^2+O(|x-x_i|^3)$,

$u_{i+1}=u_i+h_iu_i'+h_i^2u_i''/2+O(h_i^3)$, $\qquad u_{i-1}=u_i-h_{i-1}u_i'+h_{i-1}^2u_i''/2+O(h_{i-1}^3)$,

$\dfrac{u_{i+1}-u_i}{h_i}=u_i'+h_iu_i''/2+O(h_i^2)$, $\qquad \dfrac{u_i-u_{i-1}}{h_{i-1}}=u_i'-h_{i-1}u_i''/2+O(h_{i-1}^2)$,

et on a bien $(50)=u_i''$ avec une erreur $O(\max(h_i,h_{i-1}))$ par composante.

Pour une vérification soignée,prenons U un espace vectoriel de fonctions sur $[0,1]$ s'annulant en $x=0$, en $x=1$, et des points $0<x_1<x_2<\cdots<x_{N-1}<1$, $h_i=x_{i+1}-x_i$. Soit $r_hu=[\sqrt{|\omega_0|}\,u(x_1),\ldots,\sqrt{|\omega_{N-2}|}\,u(x_{N-1})]$, où $|\omega_i|$ devra être proche de h_i, mais reste libre

[1] On aura peut être reconnu une ***différence divisée***...

(on cherchera à vérifier une autre propriété). On applique (50) à u, mais en cherchant la matrice qui donne un vecteur proche de $r_h u''$ par multiplication par $r_h u$:

$$D_h^{(2)} = \begin{bmatrix} -\frac{2}{h_0 h_1} & \sqrt{\frac{|\omega_0|}{|\omega_1|}}\frac{2}{h_1(h_0+h_1)} & & & \\ \sqrt{\frac{|\omega_1|}{|\omega_0|}}\frac{2}{h_1(h_1+h_2)} & -\frac{2}{h_1 h_2} & \sqrt{\frac{|\omega_1|}{|\omega_2|}}\frac{2}{h_2(h_1+h_2)} & & \\ \ddots & \ddots & \ddots & & \\ & & \sqrt{\frac{|\omega_{N-2}|}{|\omega_{N-3}|}}\frac{2}{h_{N-2}(h_{N-2}+h_{N-1})} & -\frac{2}{h_{N-2}h_{N-1}} \end{bmatrix}$$

Avec $|\omega_i| = (h_i + h_{i+1})/2$, on obtient une matrice ***symétrique***, comme il se doit lorsque l'on approche un problème autoadjoint, ce qui sera précieux par la suite.

Cette propriété de symétrie, qui était réalisée automatiquement dans les méthodes de Ritz-Galerkin[2], doit maintenant être vérifiée cas par cas.

Pour un maillage uniforme $h_i = h$, on retrouve la matrice

$$L_h u_h = r_h f \quad :$$

$$h^{-2}\begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{N-2} \\ u_{N-1} \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_{N-2} \\ f_{N-1} \end{bmatrix} \tag{51}$$

déjà souvent rencontrée auparavant (ici, $Lu = -u''$).

2.2. Discrétisations du laplacien.

Le ***laplacien*** à deux dimensions $\partial^2/\partial x^2 + \partial^2/\partial y^2$ sera approché sur une grille de pas h_x et h_y par le fameux schéma à 5 points

$$\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h_x^2} + \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{h_y^2} \tag{52}$$

Pour $h_x = h_y = h$, on a donc $(L_h u)_{i,j} = h^{-2}\begin{pmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{pmatrix} u_{i,j}$, qui vaut $-\Delta u$ avec une erreur bornée par const. h^2 :

$u(x\pm h, y) = u(x,y) \pm h\partial u(x,y)/\partial x + (h^2/2)\partial^2 u(x,y)/\partial x^2 \pm (h^3/6)\partial^3 u(x,y)/\partial x^3 + (h^4/24)\partial^4 u(x,y)/\partial x^4 \pm \cdots$

$u(x, y\pm h) = u(x,y) \pm h\partial u(x,y)/\partial y + (h^2/2)\partial^2 u(x,y)/\partial y^2 \pm (h^3/6)\partial^3 u(x,y)/\partial y^3 + (h^4/24)\partial^4 u(x,y)/\partial y^4 \pm \cdots$

et on trouve $L_h u = -\Delta u - (h^2/12)[\partial^4 u(x,y)/\partial x^4 + \partial^4 u(x,y)/\partial y^4] + \cdots$

On augmente la précision[3] en faisant intervenir $u(x \pm 2h, y)$ et $u(x, y \pm 2h)$:

[2] Puisque l'on obtenait $L_h =$ ***matrice de rigidité*** par $(L_h\varphi_i, \varphi_j) = (L\varphi_i, \varphi_j) = a(\varphi_i, \varphi_j)$, où a est bilinéaire ***symétrique*** si L est auto-adjoint, $\{\varphi_1, \ldots, \varphi_M\}$ est une base de U_h.

[3] Abramowitz, 25.3.31

$$\frac{1}{12h^2}\begin{pmatrix} & & 1 & & \\ & & -16 & & \\ 1 & -16 & 60 & -16 & 1 \\ & & -16 & & \\ & & 1 & & \end{pmatrix} u_{i,j} = -\Delta u + O(h^4).$$

Pour l'opérateur *biharmonique*[4], $\frac{1}{h^4}\begin{pmatrix} & & 1 & & \\ & 2 & -8 & 2 & \\ 1 & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{pmatrix} u_{i,j} = \Delta^2 u + O(h^2).$

On améliore la précision de la discrétisation de *l'équation de Poisson* $-\Delta u = f$ en faisant intervenir plusieurs valeurs de f (schéma à plusieurs points (*Mehrstellenverfahren*) de Collatz) :

$\frac{1}{6h^2}\begin{pmatrix} -1 & -4 & -1 \\ -4 & 20 & -4 \\ -1 & -4 & -1 \end{pmatrix} u_{i,j} = \frac{1}{12}\begin{pmatrix} & 1 & \\ 1 & 8 & 1 \\ & 1 & \end{pmatrix} f_{i,j}$[5]. En effet, on reconnaît dans le second membre la discrétisation, à h^4 près, de $f + h^2\Delta f/12$. Ce qui vaut, *pour la solution de l'équation de Poisson*, $f - h^2\Delta^2 u/12$. Quant au membre de gauche, il représente la discrétisation, toujours à h^4 près, de $-\Delta u - h^2\Delta^2 u/12$.

3. Solubilité des équations discrètes. Stabilité numérique.

Matrices : comme U_h est essentiellement $\mathbb{R}^M$, L_h se représente par une matrice. Avec l'exemple $Lu = f$: $-u'' = f$ sur $[0, L]$ avec $u(0) = u(L) = 0$, et $h = L/N$, on obtient $M = N - 1$ équations.

Pour un problème de laplacien sur le rectangle $[0, L_1] \times [0, L_2]$, avec $h_x = L_1/N_1$ et $h_y = L_2/N_2$ dans (52), on a une matrice d'ordre $M = (N_1 - 1)(N_2 - 1)$, nombre de points intérieurs.

La moindre des choses dans l'étude de $L_h u_h = r_h f$ est de veiller à l'existence de l'inverse de L_h. En fait, on ira plus loin en cherchant à borner $\|L_h^{-1}\|$ indépendamment de h (*stabilité*).

Si L_h est réelle symétrique, la norme d'une fonction rationnelle réelle $F(L_h)$ est alors le maximum des $|F(\lambda)|$ où λ parcourt le spectre de L_h.

En effet, $F(L_h)$ est encore symétrique réelle, de valeurs propres $F(\lambda)$, et de vecteurs propres $\boldsymbol{v}^{(k)}$, $k = 1, 2, \ldots, M$ orthogonaux entre eux. Soit $\boldsymbol{v} = \sum_1^M \xi_k \boldsymbol{v}^{(k)}$ un vecteur quelconque de $\mathbb{R}^M$, alors

$$\|F(L_h)\|^2 = \sup_{\{\xi_k\}} \frac{\|F(L_h)\boldsymbol{v}\|^2}{\|\boldsymbol{v}\|^2} = \sup_{\{\xi_k\}} \frac{\sum_k (F(\lambda_k))^2 \xi_k^2}{\sum_k \xi_k^2}$$

vaut effectivement $\max_k (F(\lambda_k))^2$. □

D'où l'intérêt d'examiner le spectre de L_h. La méthode de Ritz appliquée à une forme coercive livrait spontanément une matrice de rigidité définie positive, donc de spectre > 0. Ici, il faut de nouveau tout examiner cas par cas :

[4] Abramowitz, 25.3.32

[5] P. Henrici, *Applied and Complex Computational Analysis*, vol. 3, Wiley, 1986, § 15.12

3.1. Spectres de laplaciens discrétisés.

Les valeurs propres et vecteurs propres de (51) *(problème de Dirichlet 1-D) sont*

$$\lambda_k = 2(1-\cos(\pi kh/L))/h^2\ , \qquad \boldsymbol{v}_j^{(k)} = \sin(\pi jkh/L), j = 1,\dots,N-1\ ; k = 1,\dots,N-1, \tag{53}$$

où $L = Nh$.
En effet, avec $\boldsymbol{v}_0 = \boldsymbol{v}_N = 0$, $L_h\boldsymbol{v} = \lambda\boldsymbol{v}$ signifie

$$h^{-2}\left[-\boldsymbol{v}_{j-1} + 2\boldsymbol{v}_j - \boldsymbol{v}_{j+1}\right] = \lambda\boldsymbol{v}_j, \qquad j = 1,2,\dots,N-1.$$

Cette ***relation de récurrence*** à coefficients constants se résout par $\boldsymbol{v}_j = A\rho_1^j + B\rho_2^j$, $j = 0,1,\dots,N$, où ρ_1 et ρ_2 sont les deux solutions[6] de $\rho^2 + (\lambda h^2 - 2)\rho + 1 = 0$, et où A et B sont déterminés par les conditions aux limites $\boldsymbol{v}_0 = \boldsymbol{v}_N = 0$. Comme $\rho_1\rho_2 = 1$, on a $\rho_1 = \rho, \rho_2 = 1/\rho$ et $\boldsymbol{v}_0 = 0 \Rightarrow \boldsymbol{v}_j = A(\rho^j - \rho^{-j})$. La condition $\boldsymbol{v}_N = 0$ implique alors $\rho^{2N} = 1$: ρ est à choisir parmi les racines (complexes) $2N^{\text{èmes}}$ de l'unité $\rho = \exp(ik\pi/N), k = 1,2,\dots$? Le $k^{\text{ème}}$ vecteur propre est bien $\boldsymbol{v}_j^{(k)} =$ constante $\sin(jk\pi/N)$, $j = 1,2,\dots,N-1$, pour $k = 1,2,\dots,N-1 : k > N$ reproduit des éléments déjà trouvés, si $k = 2N - k', \sin(j(2N - k')\pi/N) = -\sin(jk'\pi/N)$.

Enfin, la $k^{\text{ème}}$ valeur propre est $(2-\rho_1-\rho_2)/h^2 = (2-2\cos(k\pi/N))/h^2$. □

On a donc la satisfaction de constater que les valeurs propres sont positives, entre

$$\lambda_{\min} \approx \frac{\pi^2}{N^2h^2} = \frac{\pi^2}{L^2} \qquad \text{et} \qquad \lambda_{\max} \approx \frac{4}{h^2}. \tag{54}$$

En fait, les petites valeurs propres sont proches de $k^2\pi^2/L^2$ qui sont les valeurs propres de $L = -d^2/(dx^2)$, et on a $\|L_h^{-1}\| \approx \|L^{-1}\| = L^2/\pi^2$. Par contre, $\|L_h\| \approx 4/h^2 \to \infty$ (L est non borné sur $L^2[0,L]$).

Problèmes de Neumann et Robin 1-D. On donne à la frontière, non pas u, mais des combinaisons linéaires $u'(0) - \alpha u(0)$ (Robin ; $\alpha = 0$:Neumann ; cas limite $\alpha \to \infty$: Dirichlet) et $u'(L) + \beta u(L)$. Nous avons maintenant $N+1$ équations à $N+1$ inconnues $u_0, u_1, \dots, u_N$. Les membres de gauche sont encore $h^{-2}(-u_{i-1}+2u_i-u_{i+1}$, mais on utilise la connaissance de $(2h)^{-1}(u_1-u_{-1})-\alpha u_0$ pour éliminer u_{-1} de la première équation qui devient $h^{-2}[2(1+\alpha h)u_0 - 2u_1] = \dots$; et $h^{-2}[2(1+\beta h)u_N - 2u_{N-1}]$ pour la dernière équation.
Valeurs et vecteurs propres : les composantes d'un vecteur propre sont $v_j = \sin(j\theta + \varphi)$, avec $h^2\lambda = 2 - 2\cos\theta$, pour $j = 0,\dots,N$ prolongées en $j = -1$ et $j = N+1$. On doit alors vérifier les conditions aux frontières $(2h)^{-1}(v_1 - v_{-1}) - \alpha v_0 = 0$ et $(2h)^{-1}(v_{N+1} - v_{N-1}) + \beta v_N = 0$. On a $\sin\theta = \alpha h\tan\varphi = -\beta h\tan(N\theta + \varphi)$. Elimination de φ : $(\alpha\beta h^2 - \sin^2\theta)\tan(N\theta) + (\alpha+\beta)h\sin\theta = 0$. En $x = \cos\theta$: $(x^2 + \alpha\beta h^2 - 1)U_{N-1}(x) + (\alpha+\beta)hT_N(x) = 0$. N.B. θ peut être complexe, mais $x < 1$, donc $\lambda > 0$ dès que α et $\beta \geqslant 0$, $\alpha + \beta > 0$.

Si $\alpha = \beta = 0$ (Neumann pur), la matrice L_h est singulière, mais les équations $L_hu = b$ sont encore solubles, et donnent u à un vecteur constant près, si $b_0/2 + b_1 + \dots + b_{N-1} + b_N/2 = 0$.

Problème périodique : $u_0 = u_N$, donc N inconnues $u_0,\dots,u_{N-1}$. Avec $u_{-1} = u_{N-1}$, la première équation est $h^{-2}(2u_0 - u_1 - u_{N-1}) = b_0$, la dernière est $h^{-2}(-u_0 - u_{N-2} + 2u_{N-1}) = b_{N-1}$. Le système est singulier, mais soluble et donne u à un vecteur constant près si $b_0 + b_1 + \dots + b_{N-1} = 0$. Valeurs et vecteurs propres : $v_j = \sin(j\theta + \varphi)$, $h^2\lambda = 2 - 2\cos\theta$, avec $\theta = 2k\pi/N$, $k = 0,1,\dots,N-1$, φ arbitraire (les valeurs propres $\neq 0$ et 4 sont doubles).

Pour le problème de Dirichlet $Lu = -\Delta u = 0$ dans le rectangle $\Omega = R = (0,L_1)\times(0,L_2)$ avec u donnée sur la frontière $\partial\Omega$, ou de Poisson $Lu = -\Delta u = f$ dans R avec $u|_{\partial\Omega} = 0$, on

[6] On devra vérifier que $\rho_1 \neq \rho_2$. Dans des cas où $\rho_1 = \rho_2 = \rho$, la solution est $(Aj + B)\rho^j$.

trouve, pour le Laplacien discrétisé par (52)

$$\begin{aligned}\lambda_{k,\ell} &= 2(1-\cos(\pi k h_x/L_1))/h_x^2 + 2(1-\cos(\pi \ell h_y/L_2))/h_y^2\,,\\ \boldsymbol{v}_{i,j}^{(k,\ell)} &= \sin(\pi i k h_x/L_1)\sin(\pi j \ell h_y/L_2),, i,k = 1,\dots,N_1-1,\ j,\ell = 1,\dots,N_2-1,\end{aligned} \tag{55}$$

où $L_1 = N_1 h_x, L_2 = N_2 h_y$.
(Structure tensorielle de L_h).

Donc, $\lambda_{\min} \approx \dfrac{\pi^2}{L_1^2} + \dfrac{\pi^2}{L_2^2}$ et $\lambda_{\max} \approx \dfrac{4}{h_x^2} + \dfrac{4}{h_y^2}$.

Inégalités si $\Omega \subset R$, pour une même grille recouvrant à la fois Ω et R :

$$\lambda_{\min,\Omega} \geqslant \lambda_{\min,R} \qquad , \qquad \lambda_{\max,\Omega} \leqslant \lambda_{\max,R}.$$

En effet, $\lambda_{\min,\Omega} = \min \xi^T L_{h,\Omega} \xi$ sur les vecteurs vérifiant $\xi^T\xi = 1$. Comme chaque composante de ξ est associée à une valeur de fonction en un des N_Ω points de grille dans Ω, considérons que ξ contient des composantes associées à tous les N_R points de grille de R, mais que les composantes associées à des points de $R \setminus \Omega$ sont nulles. On minimise donc $\xi^T L_{h,R}\xi$ sur une partie de $\mathbb{R}^{N_R}$, le minimum est donc plus grand.

3.2. Déterminant, constante de Catalan.

À ***une*** dimension, le déterminant D_N de L_h se relie à D_{N-1} et D_{N-2} par (**??**) :

$$D_N = (2/h^2)D_{N-1} - (1/h^2)^2 D_{N-2}$$

, ce qui se résout par $D_N = \dfrac{N}{h^{2N}}$.

À ***deux*** dimensions[7], on a, par (55),

$$\begin{aligned} D_{N_1,N_2} &= \prod_{k=1}^{N_1}\prod_{\ell=1}^{N_2} 2(1-\cos(\pi k h_x/L_1))/h_x^2 + 2(1-\cos(\pi \ell h_y/L_2))/h_y^2\\ &= \frac{1}{h_y^{2N_1N_2}} \prod_{k=1}^{N_1} U_{N_2}\left(1 + \frac{h_y^2}{h_x^2}(1-\cos(\pi k h_x/L_1))\right)\\ &\sim \frac{1}{h^{2N^2}} \prod_{k=1}^{N} [2-\cos(\pi k h/L) + \sqrt{(2-\cos(\pi k h/L))^2-1}]^N \quad \text{si } L_1 = L_2\\ &\sim \frac{1}{h^{2N^2}} \exp\left(\frac{N^2}{\pi}\int_0^\pi \log[2-\cos\theta + \sqrt{(2-\cos\theta)^2-1}]\,d\theta\right)\end{aligned}$$

3.3. Consistance et stabilité numérique ⇒ convergence.

Définition. La famille de problèmes $\{L_h u_h = r_h f\}_h$ dans les espaces normés U_h est **numériquement stable** si

$$\|L_h^{-1}\|_h \leqslant C,$$

avec C indépendant de h quand $h \to 0$.

Théorème. *Si*

(1) *Les opérateurs L_h sont des approximations consistantes de L,*

(2) *$Lu = f$ a une solution dans U,*

(3) *la famille $\{L_h u_h = r_h f\}$ est numériquement stable,*

[7] Ph. Ruelle : "Des modèles de piles de sable aux matrices de Toeplitz", *Séminaire non linéaire MAPA, FYMA, CORE*, 17 octobre 2002.

alors, $\|u_h - r_h u\| \to 0$ *quand* $h \to 0$.

Cet énoncé traverse, sous de nombreuses formes (théorèmes de Kantorovitch, Lax, etc. etc.) toute l'analyse numérique

En effet, $u_h - r_h u = L_h^{-1} r_h f - r_h u = L_h^{-1} r_h L u - r_h u = L_h^{-1}(r_h L u - L_h r_h u)$ borné par $C\|r_h L u - L_h r_h u\|$ qui tend vers 0 par (49) (définition de consistance, § 2.1, p. 123). □

Rappelons que l'on a normalement $p_h r_h u \to u$, donc, ici, $p_h u_h \to u$ quand $h \to 0$. On déduit alors ***l'unicité*** de la solution de $Lu = f$ dans U.

4. Méthodes itératives de résolution numérique. Méthodes multigrilles.

Pour h petit, $L_h u_h = r_h f$ représente un nombre énorme d'équations linéaires toutes très simples. Les méthodes directes de résolution (éliminations- factorisations de Gauss) deviennent ruineuses, surtout pour des problèmes tridimensionnels.

Les méthodes ***itératives***[8] de résolution consistent à construire des approximations successives $\ldots, u^{(m)}, u^{(m+1)}, \ldots$ de u_h par des opérations très bon marché, des combinaisons linéaires de vecteurs et des multiplications par L_h.

4.1. Méthode de Jacobi.

Ainsi, une famille de ***méthodes de Jacobi*** (ou ***Richardson***) à un paramètre fixe se présente comme

$$u^{(m+1)} = u^{(m)} - \alpha(L_h u^{(m)} - r_h f). \tag{56}$$

Comme $L_h u_h = r_h f$, on a évidemment

$$u_h = u_h - \alpha(L_h u_h - r_h f),$$

d'où, par soustraction de (56)

$$u^{(m+1)} - u_h = (I - \alpha L_h)(u^{(m)} - u_h) = (I - \alpha L_h)^{m+1}(u^{(0)} - u_h). \tag{57}$$

Si L_h est symétrique définie positive, la méthode converge si $-1 < 1 - \alpha\lambda < 1$ pour λ dans le spectre de L_h, ce qui donne la condition $0 < \alpha < 2/\lambda_{\max}$.

On prend α proche de $1/\lambda_{\max}$. Si $\kappa = \lambda_{\max}/\lambda_{\min}$ (***nombre de condition*** de L_h) est grand, la convergence est *très* médiocre : pour gagner une décimale, $(1 - 1/\kappa)^m \approx 0.1 \Rightarrow$ il faut environ 2.3κ itérations.

Pour des problèmes du second ordre, κ se comporte comme h^{-2}, h^{-4} pour des problèmes du quatrième ordre...

On améliore les choses en recourant à des méthodes plus élaborées : gradients conjugués, relaxation optimale, GMRES, etc. assorties de techniques de préconditionnement sophistiquées.

Voici une technique spécialement adaptée aux systèmes d'équations provenant de discrétisations de problèmes différentiels :

[8] Plus de détails dans le cours MATH2172.

4.2. Méthodes multigrilles.

On fait une analyse modale, ou fréquentielle, des itérés successifs de Jacobi pour un problème modèle $-u'' = f$ ou $-\Delta u = f$: soit $u^{(0)} - u_h = \sum_1^M \xi_k^{(0)} \boldsymbol{v}^{(k)}$ la représentation du vecteur d'erreur initial dans la base des vecteurs propres de L_h. Comme $(I - \alpha L_h)\boldsymbol{v}^{(k)} = (1 - \alpha\lambda_k)\boldsymbol{v}^{(k)}$, (57) montre que

$$u^{(m)} - u_h = \sum_1^M (1 - \alpha\lambda_k)^m \xi_k^{(0)} \boldsymbol{v}^{(k)}.$$

Soit $\alpha = 1/\lambda_{\max}$. On voit que les contributions des vecteurs propres $\boldsymbol{v}^{(k)}$ s'amortissent d'autant plus vite que λ_k est grand, ainsi, l'amortissement est plus rapide que $1/2^m$ si $\lambda_{\max}/2 \leqslant \lambda_k \leqslant \lambda_{\max}$.

Les petites valeurs λ_k correspondent à des contributions de ***basse fréquence*** $\boldsymbol{v}_j^{(k)} = \sin(\pi jkh/L)$: longueur d'onde $= 2L/k \Rightarrow$ fréquence $= k/(2L)$. Ces contributions ne nécessitent pas un pas de discrétisation très fin pour être déterminées de manière satisfaisante, on va les calculer séparément sur une grille plus grossière, par exemple de pas $2h$, c'est l'idée des schémas multigrilles, qui s'applique chaque fois qu'une méthode itérative n'est ralentie que par les contributions de basse fréquence dans le vecteur d'erreur[9].

On procède comme suit :
on applique quelques fois, disons $p-1$ fois, le schéma itératif, on obtient ainsi $u_h^{(pm)}$. On va estimer le vecteur d'erreur pour établir $u_h^{(pm+1)}$:

$$u_h^{(pm+1)} = u_h^{(pm)} + \text{ une approximation de } \ u_h - u_h^{(pm)}.$$

On devrait résoudre $L_h(u_h - u_h^{(pm)}) = r_h f - L_h u_h^{(pm)}$ (vecteur résidu). On va résoudre au moyen de L_{2h} au lieu de L_h. Comme L_{2h} ne s'applique qu'à des fonctions de U_{2h}, on a besoin d'un opérateur de restriction $U_h \to U_{2h}$: on évalue $L_{2h}^{-1} r_{h\to 2h}(r_h f - L_h u_h^{(pm)})$. Souvent, l'opérateur $r_{h\to 2h}$ consiste à prendre une composante sur 2, ou une sur 4 dans des problèmes à deux dimensions, etc.
On revient à la grille initiale par un opérateur de prolongement $p_{2h\to h}$ (souvent une interpolation rudimentaire) :

$$u_h^{(pm+1)} = u_h^{(pm)} + p_{2h\to h} L_{2h}^{-1} r_{h\to 2h}(r_h f - L_h u_h^{(pm)}).$$

En tenant compte des $p-1$ applications de départ, le vecteur d'erreur est multiplié par

$$\left(I - p_{2h\to h} L_{2h}^{-1} r_{h\to 2h} L_h\right)(I - \alpha L_h)^{p-1}$$

au cours d'un cycle complet.

Examinons l'effet sur un vecteur propre $\boldsymbol{v}^{(k)} = \{\sin(j\pi kh/L)\}_{j=1}^{j=N-1}$ de L_h : tout d'abord, $L_h\boldsymbol{v}^{(k)} = 4h^{-2}\sin^2\theta_k \boldsymbol{v}^{(k)}$, avec $\theta_k = \pi kh/(2L) = \pi k/(2N)$. Dès lors, si $\alpha = 1/\lambda_{\max} \approx h^2/4$, $(I - \alpha L_h)^{p-1}\boldsymbol{v}^{(k)} = \cos^{2p-2}\theta_k \boldsymbol{v}^{(k)}$.

[9] C'est aussi le cas avec les méthodes de Gauss-Seidel. Cf : Briggs, William L., *A multigrid tutorial*, Philadelphia, Pa. : Society for Industrial and Applied Mathematics (SIAM). IX, 90 p. ; (1987).

Ensuite, on prend une composante sur deux[10] de $\boldsymbol{v}^{(k)}$, ce qui donne $\{\sin(2j\pi kh/L)\}_{j=1}^{j=N/2-1} = \{\sin(j\pi k(2h)/L)\}_{j=1}^{j=N/2-1} = \boldsymbol{w}^{(k)}$, le $k^{\text{ème}}$ vecteur propre de L_{2h}.

En fait, L_{2h} n'a que $N/2-1$ valeurs et vecteurs propres au lieu de $N-1$. Ce qui précède est valable pour $k = 1, 2, \ldots, N/2-1$ (les basses fréquences) ; pour $k = N/2+1, \ldots, N-1$, on voit que $\{\sin(j\pi k(2h)/L)\}_{j=1}^{j=N/2-1} = \{\sin(2j\pi k/N)\}_{j=1}^{j=N/2-1} = \{-\sin(2j\pi(N-k)/N)\}_{j=1}^{j=N/2-1} = -\boldsymbol{w}^{(N-k)}$ (*aliasing*).

On a donc

$$r_{h\to 2h}L_h\boldsymbol{v}^{(k)} = \begin{cases} 4h^{-2}\sin^2\theta_k\boldsymbol{w}^{(k)} \text{ si } \ k < N/2, \\ -4h^{-2}\sin^2\theta_k\boldsymbol{w}^{(N-k)} \text{ si } \ k > N/2. \end{cases}$$

Ensuite, comme $\boldsymbol{w}^{(r)}$ est un vecteur propre de L_{2h} de valeur propre $4(2h)^{-2}\sin^2(2\theta_r) = 4h^{-2}\sin^2\theta_r\cos^2\theta_r$, la multiplication par L_{2h}^{-1} donne

$$L_{2h}^{-1}r_{h\to 2h}L_h\boldsymbol{v}^{(k)} = \begin{cases} \dfrac{1}{\cos^2\theta_k}\boldsymbol{w}^{(k)} \text{ si } \ k < N/2, \\ -\dfrac{1}{\cos^2\theta_k}\boldsymbol{w}^{(N-k)} \text{ si } \ k > N/2. \end{cases}$$

Enfin, $p_{2h\to h}\boldsymbol{w}^{(r)}$ reprend $\sin(2j\pi r/N)$ en composante d'indice pair $2j$, et $[\sin(2j\pi r/N) + \sin((2j+2)\pi r/N)]/2 = \sin((2j+1)\pi r/N)\cos(\pi r/N)$ en composante d'indice impair $2j+1$. On vérifie que $p_{2h\to h}\boldsymbol{w}^{(r)} = \cos^2\theta_r\boldsymbol{v}^{(r)} - \sin^2\theta_r\boldsymbol{v}^{(N-r)}$, et on arrive à

$$p_{2h\to h}L_{2h}^{-1}r_{h\to 2h}L_h\boldsymbol{v}^{(k)} = \boldsymbol{v}^{(k)} - \mathrm{tg}^2\theta_k\boldsymbol{v}^{(N-k)},$$

(en utilisant $\theta_{N-k} = \pi/2 - \theta_k$), d'où finalement

$$\left(I - p_{2h\to h}L_{2h}^{-1}r_{h\to 2h}L_h\right)(I - \alpha L_h)^{p-1}\boldsymbol{v}^{(k)} = \cos^{2p-2}\theta_k\mathrm{tg}^2\theta_k\boldsymbol{v}^{(N-k)}$$

qui montre donc un facteur < 1 dès que $p \geqslant 3$.

En fait, le problème sur la grille de pas $2h$ est lui-même résolu en passant par des grilles de pas $4h$, $8h$,etc.

Cf. : W.L. Briggs, *A Multigrid Tutorial*, SIAM, 1987 ; nombreux articles et ouvrages de Brandt, Hackbush, Hemker, etc.

`http ://www.math.ucl.ac.be/~magnus/num2/mgrid.m` , `mgrid0.m`, `mglab.readme` .

```
Script started on Thu Jan 30 14:20:33 1997
venus 1 %
venus 1 % ls -l mgri*
-rw-r-----   1 magnus   ganma         27 Jan 30 11:12 mgrid.m
-rw-r-----   1 magnus   ganma        589 Jan 30 14:20 mgrid0.m

venus 2 % cat mgrid.m

f=ones(31,1);
u=mgrid0(f);

venus 3 % cat mgrid0.m

function u=mgrid0(f)
[m,n]=size(f);N=m+1;
u=zeros(N+1,1);
if N==2
    u(2)=f(1)/8;
    N
```

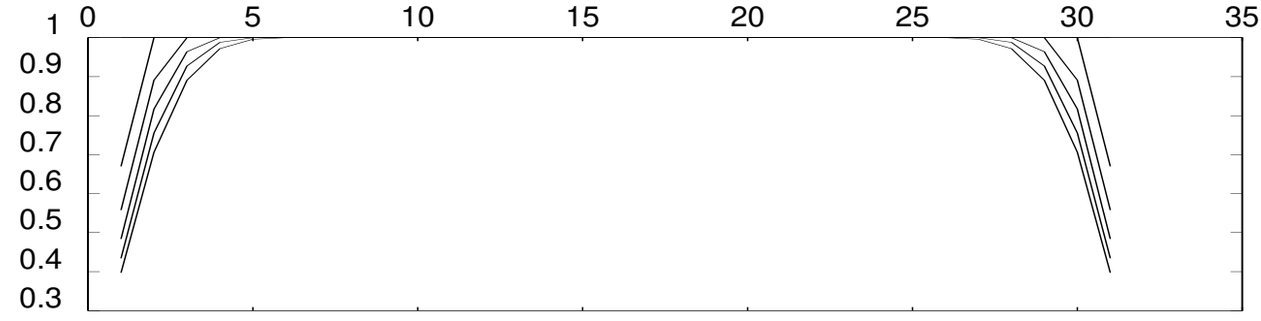
[10] Un meilleur opérateur de restriction est $(r_{h\to 2h}\boldsymbol{x})_j = (x_{2j-1} + x_{2j} + x_{2j+1})/4$, $j = 1, \ldots, N/2-1$. On voit alors s'introduire un précieux facteur $\cos^2\theta_k$ supplémentaire.

```
else
    h=1/N; u=0*ones(N+1,1);  om=0.66;
    r=( u(3:N+1)-2*u(2:N)+u(1:N-1) )/h^2 + f ; nr=norm(r);
    plot(1:m,r);
    while nr > N*0.0002
     for i=1:5
         u(2:N) = u(2:N) +om*h^2*r/2;
         r=( u(3:N+1)-2*u(2:N)+u(1:N-1) )/h^2 + f ; nr=norm(r); nrv(i)=nr;
        hold on;plot(1:m,r);
     end
     hold off; [N nrv],   co=input(' 0 = stop');
     if co==0 print -dps 'mgridf';exit;end
     newplot;
     fgros=r(3:2:N-1);
     ucorr=mgrid0(fgros);
     u(3:2:N-1)=u(3:2:N-1)+ucorr(2:N/2);
     u(2:2:N)=u(2:2:N)+0.5*(ucorr(1:N/2)+ucorr(2:N/2+1) );
     r=( u(3:N+1)-2*u(2:N)+u(1:N-1) )/h^2 + f ; nr=norm(r);
    end
end

venus 4 % matlab
                          < M A T L A B (R) >
                (c) Copyright 1984-94 The MathWorks, Inc.
                            All Rights Reserved
                               Version 4.2c
                                Dec 31 1994

Commands to get started: intro, demo, help help
Commands for more information: help, whatsnew, info, subscribe

>> mgrid
```

On ne garde que les résultats (normes de résidus) correspondant à $h = 1/32, 1/16$ et $1/8$:

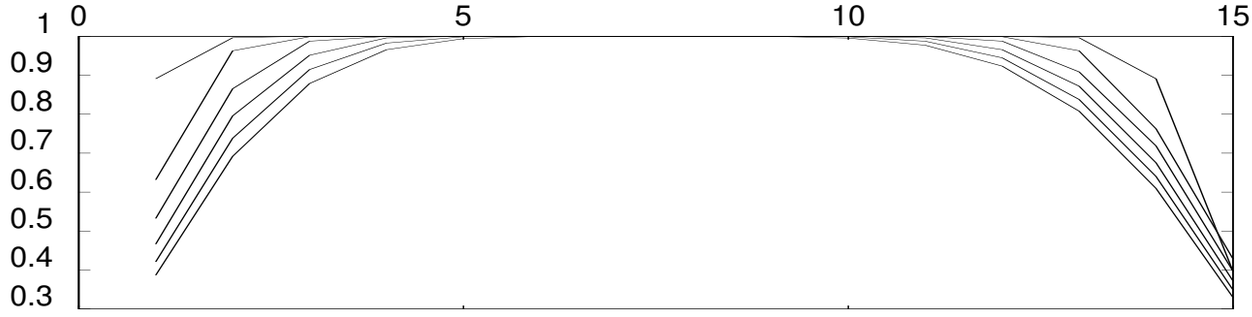
```
ans =      32    5.4679    5.4047    5.3537    5.3096    5.2700
```

Premières itérations du problème à 32 points : le vecteur résidu $r = L_h x - f$ s'amortit très, très lentement. On remarque aussi d'énormes contributions de basse fréquence.



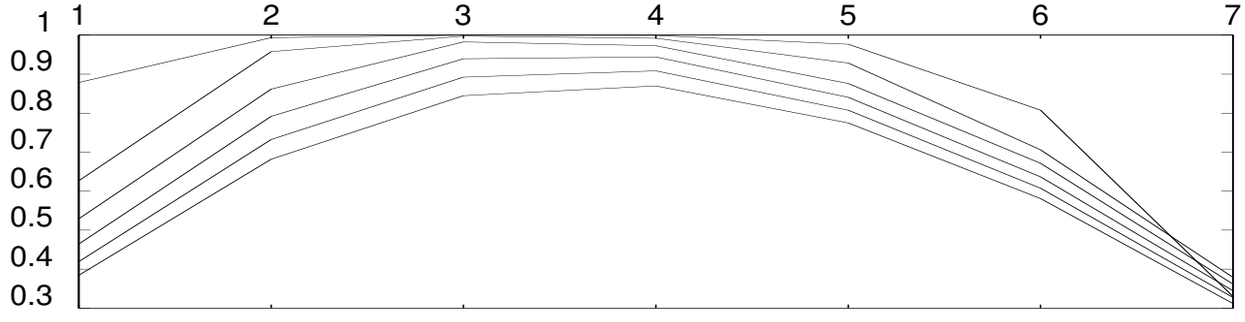
On établit une correction basée sur la grille à 16 points. Même d'esastre, mais on prépare l'avenir :

```
    16    3.6073    3.5326    3.4675    3.4085    3.3537
```



Et une correction à 8 points de la correction à 16 points :

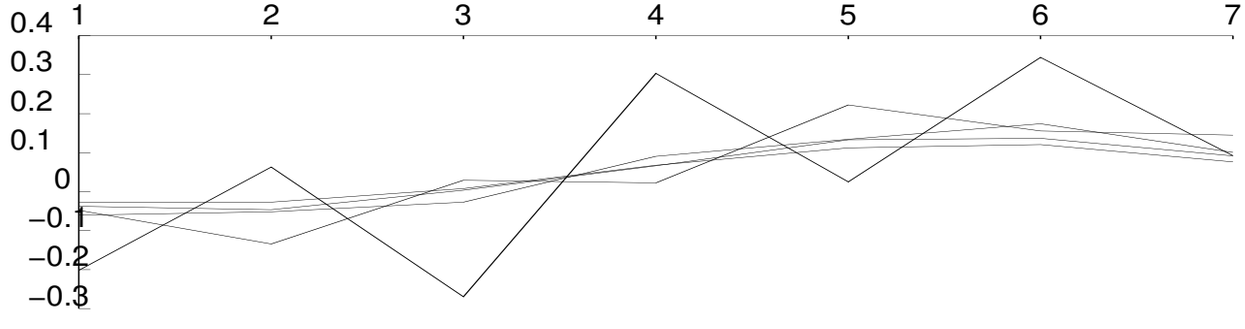
```
8     2.1895     2.0698     1.9615     1.8607     1.7659
```



. . .

Après des passages à des grilles de 4 et 2 points, résolus immédiatement, on retrouve le problème à 8 points, la réduction de la norme du vecteur résidu est bien plus rapide :

```
8     0.5806     0.3410     0.2716     0.2298     0.1978
```



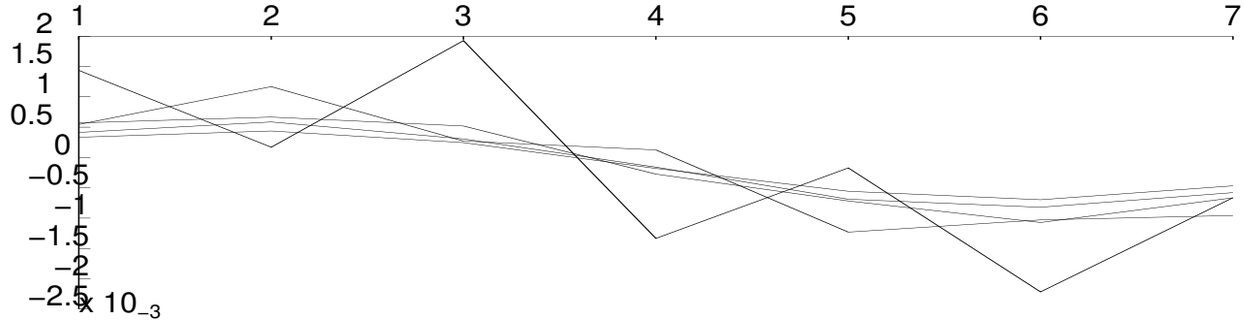
On voit aussi que les contributions de grande amplitude sont maintenant de haute fréquence, ce sont ces contributions qui sont rapidement rabotées.

. . .

Et on finit par résoudre tout-à-fait le problème à 8 points :

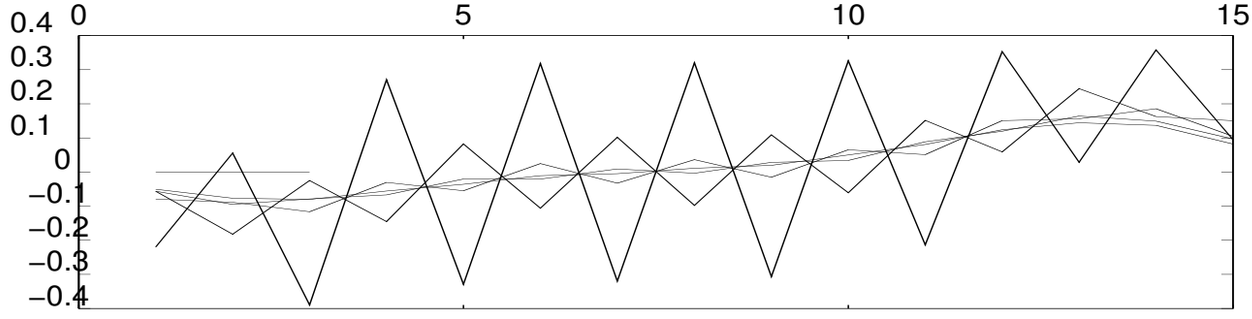
. . .

```
8     0.0036     0.0023     0.0018     0.0015     0.0012
```



La correction sur la grille de 8 points est maintenant interpolée à la grille de 16 points, ajoutée à la dernière estimation à 16 points, et on examine le résidu sur quelques itérations :

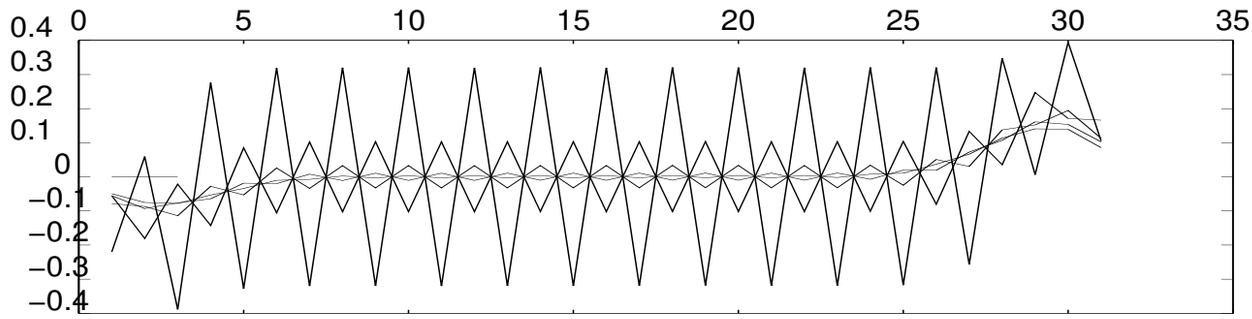
```
16    1.0950    0.4969    0.3665    0.3264    0.3012
```




```
  8    0.1649    0.1384    0.1178    0.1016    0.0888
...
...
  8    0.0016    0.0011    0.0009    0.0008    0.0007
 16    0.1128    0.0730    0.0614    0.0551    0.0509
  8    0.0299    0.0267    0.0245    0.0227    0.0213
...
...
  8    0.0015    0.0010    0.0007    0.0006    0.0005
 16    0.0219    0.0151    0.0131    0.0120    0.0112
  8    0.0062    0.0052    0.0044    0.0039    0.0034
...
  8    0.0017    0.0011    0.0008    0.0007    0.0005
 16    0.0052    0.0041    0.0036    0.0033    0.0031
```

La correction sur la grille à 16 points est maintenant arrivée à un état satisfaisant, on l'interpole sur la grille à 32 points, et on examine le résidu sur cette grille :

```
32    1.6990    0.6459    0.3836    0.3194    0.2898
```



Encore une fois, hautes fréquences qui s'amortissent très rapidement, d'où réduction significative de la norme du résidu. On reprend plusieurs fois tout le cycle :

```
    16    0.1566    0.1330    0.1164    0.1039    0.0940
     8    0.0514    0.0426    0.0361    0.0311    0.0271
   ...
   ...
     8    0.0019    0.0016    0.0015    0.0014    0.0013
    16    0.0388    0.0258    0.0215    0.0191    0.0176
     8    0.0103    0.0094    0.0088    0.0083    0.0079
   ...
     8    0.0027    0.0017    0.0014    0.0012    0.0010
    16    0.0082    0.0061    0.0054    0.0050    0.0046
     8    0.0026    0.0022    0.0018    0.0016    0.0014
   ...
     8    0.0007    0.0004    0.0003    0.0003    0.0002
    16    0.0022    0.0017    0.0015    0.0013    0.0012
    32    0.1096    0.0700    0.0572    0.0497    0.0444
    16    0.0228    0.0185    0.0159    0.0141    0.0127
     8    0.0062    0.0049    0.0041    0.0035    0.0031
   ...
     8    0.0015    0.0009    0.0007    0.0005    0.0004
    16    0.0062    0.0049    0.0043    0.0038    0.0034
     8    0.0015    0.0011    0.0008    0.0007    0.0006
    16    0.0018    0.0015    0.0014    0.0013    0.0012
    32    0.0218    0.0167    0.0147    0.0134    0.0124
    16    0.0068    0.0057    0.0049    0.0042    0.0037
     8    0.0016    0.0011    0.0008    0.0006    0.0005
    16    0.0019    0.0016    0.0015    0.0013    0.0013
    32    0.0065    0.0053    0.0048    0.0045    0.0042
>> exit
 51119 flops.

venus 5 % exit
venus 6 %script done on Thu Jan 30 14:22:15 1997
```

Donc, on décrit un vecteur associé à des points de discrétisation par son contenu ***fréquentiel***. Pour le vecteur d'erreur $\boldsymbol{e}_h := \boldsymbol{u}_h - \boldsymbol{L}_h^{-1}\boldsymbol{f}_h$: $\boldsymbol{e}_h = \sum_\omega C(\omega) \begin{bmatrix} \vdots \\ e^{i\omega x} \\ \vdots \end{bmatrix}$ où ω parcourt des valeurs entre $\pm\omega_0$ et $\pm\pi/h$ (à deux dimensions, on aurait $\exp i(\omega_1 x + \omega_2 y)$).

On distingue les contributions de ***basses fréquences*** $|\omega| < \pi/(2h)$ et celles de ***hautes fréquences*** $|\omega| \geqslant \pi/(2h)$.

La description fréquentielle permet d'étudier commodément la performance d'une méthode itérative.

Par exemples, avec $\boldsymbol{L}_h$ = discrétisation habituelle de $-$ dérivée seconde :

(1) ***Jacobi***.

$$u_{h,new}(x) = \frac{f_h + u_{h,old}(x-h)/h^2 + u_{h,old}(x+h)/h^2}{2/h^2},$$

pour le vecteur d'erreur $\boldsymbol{e}_h$:

$$e_{h,new}(x) = \frac{e_{h,old}(x-h) + e_{h,old}(x+h)}{2},$$

contribution de $\exp(i\omega x)$: $C_{new} = [(e^{-i\omega h} + e^{i\omega h})/2]C_{old} = \cos(\omega h)C_{old}$.
Pas de séparation basses-hautes fréquences.

(2) ***Jacobi-Richardson***.

$$\boldsymbol{u}_{h,new} = \boldsymbol{u}_{h,new,Jacobi} + \mu(\boldsymbol{L}_h\boldsymbol{u}_{h,old} - \boldsymbol{f}_h),$$

vecteur d'erreur : $\boldsymbol{e}_{h,new} = \boldsymbol{e}_{h,new,Jacobi} + \mu \boldsymbol{L}_h \boldsymbol{e}_{h,old}$,
Commes les lignes de $\boldsymbol{L}_h$ sont $[\cdots, h^{-2}, -2h^{-2}, h^{-2}, \cdots]$,

$$\begin{aligned} C_{new} &= C_{new,Jacobi} + \mu h^{-2}[\exp(-i\omega h) - 2 + \exp(i\omega h)]C_{old} \\ &= [(1 + 2\mu h^{-2})\cos(\omega h) - 2\mu h^{-2}]C_{old} \end{aligned}$$

$= \cos^2(\omega h/2)$ avec $\mu = -h^2/4$.

(3) ***Gauss-Seidel***.

$$u_{h,new}(x) = \frac{f_h + u_{h,\boldsymbol{new}}(x-h)/h^2 + u_{h,old}(x+h)/h^2}{2/h^2},$$

pour le vecteur d'erreur $\boldsymbol{e}_h := \boldsymbol{u}_h - \boldsymbol{L}_h^{-1}\boldsymbol{f}_h$:

$$e_{h,new}(x) = \frac{e_{h,new}(x-h) + e_{h,old}(x+h)}{2},$$

contribution de $\exp(i\omega x)$: $C_{new} = e^{-i\omega h}C_{new}/2 + e^{i\omega h})C_{old}/2$,
donc, $C_{new} = \dfrac{e^{i\omega h}}{2 - e^{-i\omega h}}C_{old}$, divise la part des hautes fréquences par au moins $\sqrt{5}$.

Considérons une méthode itérative qui aboutit à $C_{new}(\omega) = \rho(\omega)C_{old}(\omega)$, avec $|\rho(\omega)| \leqslant \sigma < 1$ pour les hautes fréquences.

On applique quelques itérations : $C(\omega) \to \rho^p(\omega)C(\omega)$, et on estime la correction $\boldsymbol{L}_h^{-1}\boldsymbol{f}_h - \boldsymbol{u}_h$ qu'il faudrait ajouter à $\boldsymbol{u}_h$, **en l'évaluant sur une grille de pas $2h$**, ce qui devrait convenir aux contributions de basses fréquences (les seules mal amorties par la méthode itérative).

Donc,

(1) On effectue p itérations, et

(2) on ne retient que les composantes du vecteur résidu $\boldsymbol{L}_h\boldsymbol{u}_h - \boldsymbol{f}_h$ que sur une grille de pas $2h$.

(3) on évalue le vecteur d'erreur sur cette grille en résolvant un système d'équations avec la matrice $\boldsymbol{L}_{2h}$,

(4) on reconstitue un vecteur d'erreur sur la grille initiale par interpolation élémentaire : $e_{2j+1} = (e_{2j} + e_{2j+2})/2$, par exemple

(5) on soustrait ce vecteur $\boldsymbol{e}$ de $\boldsymbol{u}_h$. Les contributions de basses fréquences dans l'erreur devraient être fortement réduites.

Et on reprend en (1).

On obtient

$$\boldsymbol{u}_h \leftarrow \boldsymbol{u}_h - \begin{bmatrix} 1 & 0 & & \\ 1/2 & 1/2 & & \\ & 1 & 0 & \\ & 1/2 & 1/2 & \\ & \vdots & \vdots & \end{bmatrix} \boldsymbol{L}_{2h}^{-1} \begin{bmatrix} 1 & & & & & \cdots \\ 0 & 0 & 1 & & & \cdots \\ & & 0 & 0 & 1 & \cdots \end{bmatrix} (\boldsymbol{L}_h\boldsymbol{u}_h - \boldsymbol{f}_h)$$

Contributions de $\exp(i\omega x)$ lors de ces opérations :

(1) $C(\omega) \to \rho(\omega)^p C(\omega)$ après p itérations, devient $-4h^{-2}\sin^2(\omega h/2)\rho(\omega)^p C(\omega)$ dans le résidu,

(2) quand on ne retient qu'une composante sur deux, on a une contribution en $\exp(i\omega x)$ qui est la somme des contributions initiales en $\exp(i\omega x)$ et $\exp(i(\omega \pm \pi/h)x)$ (***aliasing***) :
$-4h^{-2}\sin^2(\omega h/2)\rho(\omega)^p C(\omega) - 4h^{-2}\cos^2(\omega h/2)\rho(\omega \pm \pi/h)^p C(\omega \pm \pi/h)$
pour $|\omega| \leqslant \pi/(2h)$,

(3) on résout le système avec $\boldsymbol{L}_{2h}$, d'où division par $-h^{-2}\sin^2(\omega h)$:

$$\frac{\rho(\omega)^p C(\omega)}{\cos^2(\omega h/2)} + \frac{\rho(\omega \pm \pi/h)^p C(\omega \pm \pi/h)}{\sin^2(\omega h/2)}$$

(4) Revenons à la grille de pas h. Comme $\exp(i\omega x)$ et $\exp(i(\omega \pm \pi/h)x)$ prennent les mêmes valeurs aux abscisses paires, et des valeurs opposées aux abscisses impaires, la combinaison $\lambda \exp(i\omega x) + (1-\lambda)\exp(i(\omega \pm \pi/h)x)$ qui vaut $[\exp(i\omega(x-h)) + \exp(i\omega(x+h))]/2$
$= \cos(\omega h)\exp(i\omega x)$ aux abscisses impaires est
$\cos^2(\omega h/2)\exp(i\omega x) + \sin^2(\omega h/2)\exp(i(\omega \pm \pi/h)x)$. On a donc en $\exp(i\omega x)$

$$\rho(\omega)^p C(\omega) + \frac{\rho(\omega \pm \pi/h)^p C(\omega \pm \pi/h)}{\tan^2(\omega h/2)}$$

(5) Enfin, la soustraction : $(1-\rho(\omega)^p)C(\omega) - \dfrac{\rho(\omega \pm \pi/h)^p C(\omega \pm \pi/h)}{\tan^2(\omega h/2)}$

Avec Jacobi-Richardson, $\rho(\omega) = \cos^2(\omega h/2)$, on a donc $1 - \cos^{2p}(\omega h/2)$ et $\sin^{2(p-1)}(\omega h/2)\cos^2(\omega h/2)$.

Références

BRAMBLE James H., *Multigrid Methods.* Pitman Research Notes in Mathematics Series **294**, Longman Scientific & Technical, John Wiley & Sons, Inc., 1995.

BRAMBLE J.H. ; COHEN A. ; DAHMEN W., *Multiscale Problems and Methods in Numerical Simulations. Lectures given at the C.I.M.E. Summer School held in Martina Franca, Italy, September 9-15, 2001*, Lecture Notes in Mathematics **1825**, Springer-Verlag, 2003.

réseau : MGNet, `http ://www.mgnet.org/`

MGNet Sites.....

MGNet Pages.....

multigrid animation

MGNet

Introduction

This is a repository for information related to multigrid, multilevel, multiscale, aggregation, defect correction, and domain decomposition methods. These methods are used primarily by scientists and engineers to solve partial differential equations on serial or parallel computers.

...

Tutorials

If you are new to the multigrid or domain decomposition fields, or just wonder what MGNet is all about, look here. There are pointers to some rather nice tutorials on the subject.

Free Software

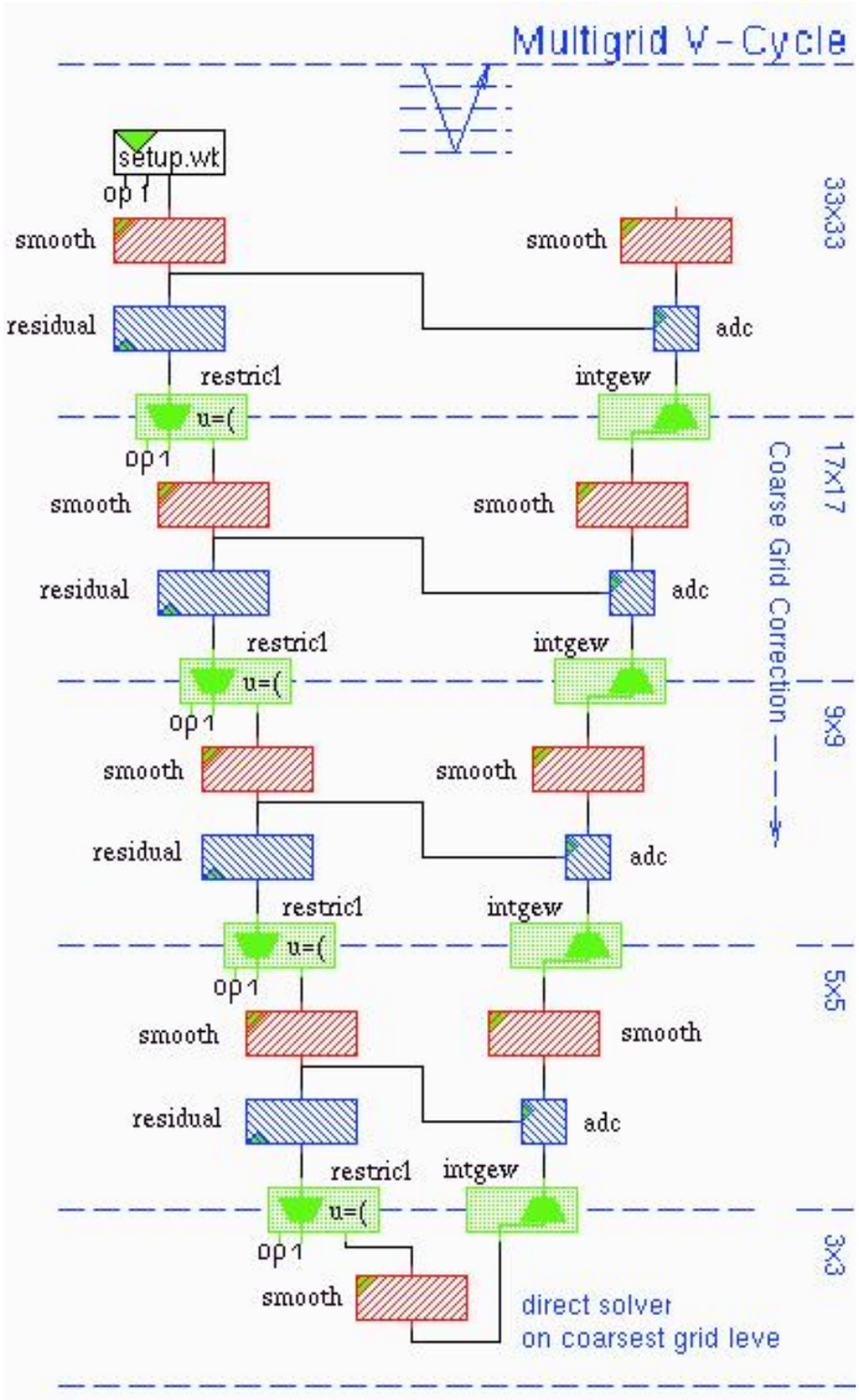
A number of software packages are stored on MGNet. Some are public domain, some are copyrighted, and some may someday be copyrighted. If you have a multigrid, domain decomposition, or parallel code or package you can place a copy in MGNet, too.

Début de Uli Rüde's multigrid workbench, `http ://www.mgnet.org/mgnet/tutorials/xwb.html`

Multigrid Workbench

The following picture displays a V-Cycle of the multigrid method. Each box represents one of the algorithmic components of a multigrid algorithm. The unbroken channels show how data is passed through the algorithm.

This algorithm is applied to an example problem. The boxes labeled smooth in the picture are active. By clicking on them, information about the status of the algorithm at this stage is displayed (graphically). Mosaic users can use the middle mouse button to obtain the pictures in separate windows.


Multigrid V-Cycle
setup.wk
op 1
smooth
residual
restric1
u=(
op 1
smooth
residual
restric1
u=(
op 1
smooth
residual
restric1
u=(
op 1
smooth
residual
restric1
u=(
op 1
smooth
direct solver
on coarsest grid leve
smooth
adc
intgew
smooth
adc
intgew
smooth
adc
intgew
smooth
adc
intgew
33x33
17x17
9x9
5x5
3x3
Coarse Grid Correction

5. Matrices d'inverses positives. Convergence.

On examine ici une idée très intéressante dans le traitement par différences finies de problèmes elliptiques. On ne sépare pas vraiment consistance et stabilité, mais on trouvera un encadrement pour la solution.

Ref. : L. Collatz, *The Numerical Treatment of Differential Equations*, 3ème éd., Springer, 1966, pp. 42-47 et 342-353.

5.1. Matrices positives, $M-$matrices.

Définition *Une matrice $m \times n$ est dite positive si tous ses éléments sont positifs : $a_{i,j} \geqslant 0,\ i = 1, \ldots, m,\ j = 1, \ldots, n$.*
Une matrice carrée est dite d'inverse positive si son inverse est positive : $(\boldsymbol{A}^{-1})_{i,j} \geqslant 0,\ i, j = 1, \ldots, n$.

On munit $\mathbb{R}^N$ de la relation d'ordre (partiel) $\boldsymbol{x} \leqslant \boldsymbol{y}$ si $x_i \leqslant y_i, i = 1, 2, \ldots, N$.

Alors, si $\boldsymbol{A}$ est d'inverse positive, $\boldsymbol{Ax} \geqslant \boldsymbol{Ay} \Rightarrow \boldsymbol{x} \geqslant \boldsymbol{y}$. On a donc une relation de ***monotonie***[11] du second membre vers la solution.

Définition. Une $M-$matrice est une matrice carrée inversible, d'inverse positive, déléments diagonaux > 0, et d'éléments non diagonaux $\leqslant 0$.

On a le

Théorème[12] *Une matrice carrée réelle vérifiant*

(1) $a_{i,i} > 0, a_{i,j} \leqslant 0, i = 1, \ldots, N; j = 1, \ldots, N, j \neq i$,

(2) $a_{i,i} \geqslant \sum_{j \neq i} |a_{i,j}|, i = 1, \ldots, N,$ *(dominance),*

(3) $a_{i,i} > \sum_{j \neq i} |a_{i,j}|,$ *pour au moins une valeur de i,*

(4) $\boldsymbol{A}$ *irréductible,*

est d'inverse positive.

Par ***irréductible***, ou ***indécomposable***, on entend ici une matrice dont on ne peut extraire des lignes d'indices $i_1, \ldots, i_m,\ m < N$ qui n'auraient des éléments non nuls que pour des indices de colonnes $\in \{i_1, \ldots, i_m\}$ (de sorte que, du système d'équations $\boldsymbol{Ax} = \boldsymbol{b}$, on pourrait extraire un système pour les seuls $x_{i_1}, \ldots, x_{i_m}$).

Preuve du théorème 5.1 : examinons $\boldsymbol{b} = \boldsymbol{Ax}$, où $\boldsymbol{x} \in \mathbb{R}^N$, $\boldsymbol{x} \neq 0$, et $\boldsymbol{b} \geqslant 0$.

Montrons d'abord que $\boldsymbol{A}$ est non singulière : $\boldsymbol{Ax} = 0 \Rightarrow \boldsymbol{x} = 0$.

Supposons $\boldsymbol{x} \neq 0$, soit x_{i_1} une des composantes de plus grande valeur absolue de $\boldsymbol{x}$.

L'équation $a_{i_1,i_1} x_{i_1} - \sum_{j \neq i_1} |a_{i_1,j}| x_j = 0$ n'est possible que si $a_{i_1,i_1} = \sum_{j \neq i_1} |a_{i_1,j}|$ et $x_j = x_{i_1}$ pour tous les indices j tels que $a_{i_1,j} \neq 0$. Soient $i_2, i_3, \ldots$ ces nouveaux indices (il y en a au mois un, sinon l'équation i_1 ne contiendrait que l'inconnue x_{i_1} et $\boldsymbol{A}$ serait réductible). On arrive ainsi à un groupe d'équations d'indices $i_1, i_2, \ldots, i_m$ reliées

[11] Collatz parle de "problème de type monotone" ; A. Berman et R.J. Plemmons écrivent , dans le chap. 5 de *Nonnegative Matrices in the Mathematical Sciences*, Ac. Press, 1979, que $\boldsymbol{A}$ est une matrice monotone.

[12] Voir aussi le théorème de Taussky, et les $M-$matrices, dans R.S. Varga, *Matrix Iterative Analysis*, Prentice Hall, 1962 ; 2nd revised and expanded edition Springer-Verlag, 2000.

entre elles et où $a_{i_k,i_k} = \sum_{j\neq i_k} |a_{i_k,j}|$, $k = 1,\dots,m$. On doit avoir $m = N$ car $\boldsymbol{A}$ est irréductible, et $x_1 = x_2 = \cdots = x_N \neq 0$, impossible car un des éléments diagonaux $a_{i,i}$ est *strictement* supérieur à la somme $\sum_{j\neq i} |a_{i,j}|$.

Résolvons maintenant $\boldsymbol{Ax} = \boldsymbol{b} \geqslant 0$ par itérations (de *Jacobi*) $x_i^{(k)} = \dfrac{b_i - \sum_{j\neq i} a_{i,j} x_j^{(k-1)}}{a_{i,i}}$, $i = 1,\dots,N$, en partant du vecteur nul $\boldsymbol{x}^{(0)} = 0$:

(1) $\boldsymbol{x}^{(k)} \geqslant 0$: $\quad a_{i,i} > 0, b_i \geqslant 0, -a_{i,j} \geqslant 0$ quand $j \neq i$,

(2) $\forall i,k, |x_i^{(k)}| \leqslant 2\max_i |x_i|$. En effet, comme $\boldsymbol{b} = \boldsymbol{Ax}$, on a $x_i^{(k)} = x_i + \dfrac{\sum_{j\neq i} a_{i,j}(x_j^{(k-1)} - x_j)}{a_{i,i}}$, $i = 1,\dots,N$, donc $|x_i^{(k)} - x_i| \leqslant \max_j |x_j^{(k-1)} - x_j| \leqslant \dots \leqslant \max_j |x_j^{(0)} - x_j| = \max_j |x_j|$.

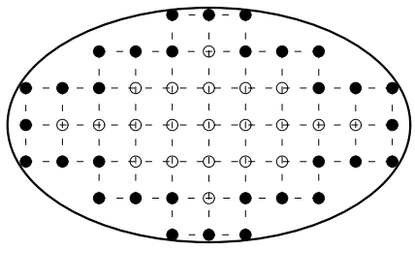
(3) $\forall i, x_i^{(k)} \geqslant x_i^{(k-1)}$. En effet, c'est vrai si $k = 1$; si $k > 1$,
$x_i^{(k)} - x_i^{(k-1)} = -\dfrac{\sum_{j\neq i} a_{i,j}(x_j^{(k-1)} - x_j^{(k-2)})}{a_{i,i}}$, $i = 1,\dots,N$,

Chaque $\{x_i^{(k)}\}_{k=0}^{\infty}$ est donc une suite positive croissante bornée, donc convergente : $\boldsymbol{x}^{(k)} \to \boldsymbol{y} \geqslant 0$ quand $k \to \infty$. Enfin, $y_i = \lim_{k\to\infty} x_i^{(k)} = \dfrac{b - \sum_{j\neq i} a_{i,j} y_j}{a_{i,i}}$, $i = 1,\dots,N$, donc, $\boldsymbol{Ay} = \boldsymbol{b} \Rightarrow \boldsymbol{y} = \boldsymbol{x}$ puisque $\boldsymbol{A}$ est non singulière, et $\boldsymbol{x} \geqslant 0$. □

On peut aussi montrer que la méthode de Seidel converge pour une telle matrice $\boldsymbol{A}$[13].

5.2. Application au laplacien.

Prenons maintenant un problème elliptique type, soit $-\Delta u = f$ dans un domaine borné Ω de $\mathbb{R}^n$, avec conditions de Dirichlet : u donnée sur la frontière $\partial\Omega$ de Ω.



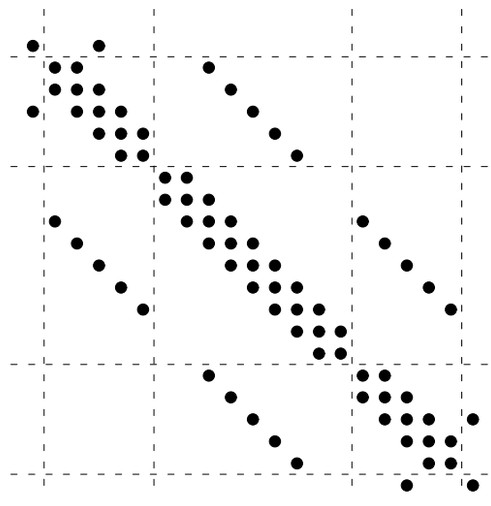
Dans la discrétisation la plus simple, on se limite à la partie de $(h\mathbb{Z})^n$ contenue dans $\overline{\Omega}$. On affecte aux points de la frontière de ce réseau (les points marqués •) la valeur d'un point proche de $\partial\Omega$. Les valeurs aux points intérieurs (marqués ∘) entrent dans les équations (on prend $n = 2$)

$$(L_h u_h)_A = \frac{4u_A - u_B - u_C - u_D - u_E}{h^2} = f_A,$$

(d'après (52) avec $h_x = h_y = h$), où A est un point

[13] Cf. Varga, *op. cit.*

courant du réseau, et B, C, D et E ses 4 voisins : $B - \begin{matrix} D \\ | \\ A \\ | \\ E \end{matrix} - C$. Dans le cas de la figure, on a 21 points intérieurs qui, repris ligne par ligne, contribuent à la silhouette (*skyline*) suivante dans la matrice L_h :



Les éléments diagonaux valent tous $4/h^2$, et les éléments non diagonaux non nuls valent $-1/h^2$. On voit que la matrice est effectivement (faiblement) dominante, et qu'elle vérifie les hypothèses du th. 5.1, du moins si h est assez petit (connexité du graphe).

Reprenons l'analyse de consistance : on a, si $u \in \mathscr{C}^2(\overline{\Omega})$, $|(L_h r_h u)_A - f_A| \leqslant \varepsilon_h$, avec $\varepsilon_h \to 0$ quand $h \to 0$. Comme $f_A = (r_h f)_A = (L_h u_h)_A$, on a $|(L_h(r_h u - u_h)_A| \leqslant \varepsilon_h$.

Considérons maintenant $w(x,y) = R^2 - x^2 - y^2$, où R est le rayon d'un disque de centre 0 contenant entièrement $\overline{\Omega}$. On trouve aisément $L_h r_h w \geqslant 4\boldsymbol{e}$,

où $\boldsymbol{e} = [1, 1, \ldots, 1]^T$ (en effet, on trouve exactement 4 aux points du réseau entourés de 4 points intérieurs ; une valeur plus grande que 4 aux points du réseau entourés de $\leqslant 4$ points intérieurs). On a donc[14]

$$-\frac{\varepsilon_h}{4} L_h r_h w \leqslant -\varepsilon_h \boldsymbol{e} \leqslant L_h(r_h u - u_h) \leqslant \varepsilon_h \boldsymbol{e} \leqslant \frac{\varepsilon_h}{4} L_h r_h w,$$

d'où, par monotonie de L_h,

$$-\frac{\varepsilon_h}{4} r_h w \leqslant r_h u - u_h \leqslant \frac{\varepsilon_h}{4} r_h w^{15}.$$

Raffinements du traitement de la frontière : croix à branches inégales, interpolation linéaire (Collatz). Cas des points anguleux (J.M.-S. Lubuma, S. Nicaise : Finite element method for elliptic problems with edge corners, *Recherche de mathématique* **48**, Institut de mathématique UCL, novembre 1995).

5.3. Processus stochastique.

Le point de vue probabiliste a éclairé et unifié les principes de base de la théorie du potentiel ; réciproquement, les concepts de la théorie du potentiel, appliqués à la théorie des probabilités ont montré

[14] Technique remontant à Courant (1926) et Gerschgorin (1930), cf. Collatz, *op. cit.*

[15] Si $u \in \mathscr{C}^4(\overline{\Omega})$, on trouve $\varepsilon_h \leqslant M_4 h^2/6$, où $M_4 = \max_{(x,y)\in\overline{\Omega}}(|\partial^4 u/\partial x^4|, |\partial^4 u/\partial y^4|)$.

la structure analytique profonde des processus de Markov. Ainsi l'isolation dont souffrait dans le passé la théorie des probabilités a été rompue par l'association de ces deux discipline mathématiques[16].

Reprenons un problème de Dirichlet $\Delta u = 0$ dans $\Omega \subset \mathbb{R}^n$, avec u donnée sur la frontière Γ. Sur une grille uniforme, les équations aux points intérieurs sont

$$u_i = \frac{\text{somme des } 2n \text{ } u \text{ voisins}}{2n},$$

où certains voisins peuvent être des points de la frontière où u est donc connue. Soit $\boldsymbol{u}$ le vecteur de *tous* les u_i, partitionné en valeurs à la frontière $\boldsymbol{u}_\Gamma$ et en valeurs aux points intérieurs $\boldsymbol{u}_\Omega$.

$$\boldsymbol{u} = \boldsymbol{S}\boldsymbol{u} : \left[\begin{array}{c} \boldsymbol{u}_\Gamma \\ \hline \boldsymbol{u}_\Omega \end{array}\right] = \left[\begin{array}{c|c} I & 0 \\ \hline B & I - A \end{array}\right] \left[\begin{array}{c} \boldsymbol{u}_\Gamma \\ \hline \boldsymbol{u}_\Omega \end{array}\right], \tag{58}$$

où la matrice A est la matrice L_h vue auparavant, multipliée par un facteur tel que ses éléments diagonaux sont tous égaux à 1. La matrice $I - A$ a donc des éléments diagonaux nuls, et au plus $2n$ éléments non diagonaux non nuls par ligne, tous égaux à $1/(2n)$. L'équation (58) est bien l'équation $A\boldsymbol{u}_\Omega = B\boldsymbol{u}_\Gamma$ pour les valeurs intérieures. Nous avons vu que A est inversible (c'est important), et que A^{-1} est la série des puissances de $I - A$ (convergence de Jacobi).

Les éléments de chaque ligne de $\boldsymbol{S}$ sont positifs et de somme unité : $\boldsymbol{S}$ est une *matrice stochastique*.

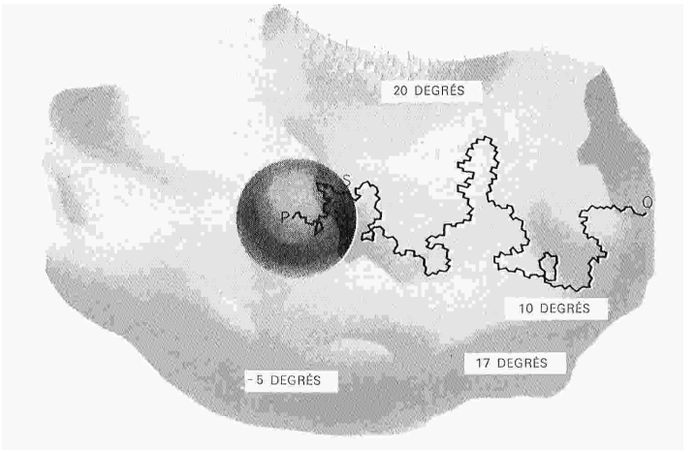


Fig. 6 — LA RÉPARTITION DES TEMPÉRATURES dans un corps solide homogène en équilibre thermique fait également intervenir des fonctions harmoniques. En effet, comme la température en un point P ne varie pas au cours du temps, celle-ci doit être égale à la moyenne des températures des points d'une petite sphère centrée en P. La température T est une fonction harmonique des coordonnées x, y, z de P. Le problème peut être résolu par la théorie probabiliste du mouvement brownien, en plaçant une particule (virtuelle) en P qui atteint la surface en un point aléatoire Q.

On interprète $S_{i,j}$ comme la probabilité de passer du point P_i au point P_j en un pas d'une promenade aléatoire. A partir d'un point P_i du réseau, on peut donc

(1) y rester indéfiniment si P_i est un point frontière (prob. $= 1$: *point absorbant*),

(2) se diriger avec une même probabilité $1/(2n)$ vers un des $2n$ voisins accessibles.

Les puissances de $\boldsymbol{S}$ donnent les probabilités de passer d'un point à un autre en plusieurs pas.

[16] R. Hersh, R.J. Griego : "Brownian Motion and Potential Theory", *Scientific American*, March 1969, Vol. **220**, No. 3, pp. 66-74 = le mouvement brownien et la théorie du potentiel, pp. 74-82 *in Les progrès des mathématiques*, Bibliothèque Pour la science, Belin, 1981.

Que se passe-t-il après un grand nombre de pas ? On a[17]

$$\boldsymbol{S} = \left[\begin{array}{c|c} I & 0 \\ \hline B & I-A \end{array}\right], \boldsymbol{S}^2 = \left[\begin{array}{c|c} I & 0 \\ \hline (I+I-A)B & (I-A)^2 \end{array}\right], \ldots, \boldsymbol{S}^N = \left[\begin{array}{c|c} I & 0 \\ \hline (I+I-A+\cdots+(I-A)^{N-1})B & (I-A)^N \end{array}\right],$$

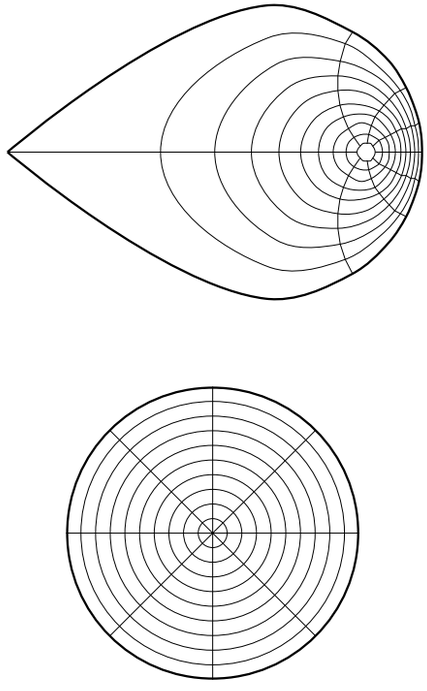
qui converge donc quand $N \to \infty$ vers $\left[\begin{array}{c|c} I & 0 \\ \hline A^{-1}B & 0 \end{array}\right]$. Une promenade aléatoire (ou alcoolique) partant d'un point intérieur P_i aboutit donc à un point frontière P_j avec une probabilité $(\boldsymbol{S}^\infty)_{i,j}$. Enfin,

$$u_i = (A^{-1}B\boldsymbol{u}_\Gamma)_i = \sum_{P_j\in\Gamma} (\boldsymbol{S}^\infty)_{i,j} u_j = \lim_{N\to\infty} \frac{\sum_j N_j u_j}{N},$$

où N_j est le nombre de fois qu'on aboutit à $P_j \in \Gamma$ après N promenades aléatoires issues de P_i (théories de Kakutani, Doob et Wiener).

6. Autres méthodes de traitement du laplacien.

6.1. Transformation conforme (2D).



D est le disque unité et Ω est son image par $z = \Psi(w) = \sqrt{3(1-w)/(2-w)}$.

Soit Φ une fonction analytique de $z = x + iy$ appliquant Ω sur un domaine D canonique (disque ou rectangle) et d'inverse également analytique :

$z = \Psi(w) = x + iy \in \Omega \longleftrightarrow w = \Phi(z) = \xi + i\eta \in D$

$dw = d\xi + id\eta = \Phi'(z)dz =$

$$[\mathrm{Re}\,\Phi' dx - \mathrm{Im}\,\Phi' dy] + i[\mathrm{Im}\,\Phi' dx + \mathrm{Re}\,\Phi' dy],$$

$$du = \frac{\partial u}{\partial \xi} d\xi + \frac{\partial u}{\partial \eta} d\eta = \left[\mathrm{Re}\,\Phi' \frac{\partial u}{\partial \xi} + \mathrm{Im}\,\Phi' \frac{\partial u}{\partial \eta}\right] dx + \left[-\mathrm{Im}\,\Phi' \frac{\partial u}{\partial \xi} + \mathrm{Re}\,\Phi' \frac{\partial u}{\partial \eta}\right] dy,$$

donc, $\begin{bmatrix} \partial u/\partial x \\ \partial u/\partial y \end{bmatrix} = \begin{bmatrix} \mathrm{Re}\,\Phi' & \mathrm{Im}\,\Phi' \\ -\mathrm{Im}\,\Phi' & \mathrm{Re}\,\Phi' \end{bmatrix} \begin{bmatrix} \partial u/\partial \xi \\ \partial u/\partial \eta \end{bmatrix}$,

ou $\partial u/\partial x + i\partial u/\partial y = \overline{\Phi'}(\partial u/\partial \xi + i\partial u/\partial \eta)$.

Appliquons l'opérateur $\partial/\partial x - i\partial/\partial y$:

$\partial^2 u/\partial x^2 + \partial^2 u/\partial y^2 = |\Phi'|^2(\partial^2 u/\partial \xi^2 + \partial^2 u/\partial \eta^2)$, ce qui montre comment le laplacien se transforme dans les nouvelles variables[18]. Une équation de Poisson $-\Delta u = f$ dans Ω se transforme avec le second membre $f/|\Phi'|^2$ dans D.

Démonstration plus "géométrique" : le laplacien étant une divergence (la divergence du gradient), on le récupère en un point par moyenne sur un petit disque d de rayon $r \to 0$:

[17] Cf. Peter Doyle `http ://www.math.dartmouth.edu/~doyle/` , P. Doyle, J.L. Snell : 'Random Walks and Electric Networks' `http ://front.math.ucdavis.edu/math.PR/0001057`

[18] Il manque $(\partial\overline{\Phi'}/\partial\xi - i\partial\overline{\Phi'}/\partial\eta)(\partial u/\partial\xi + i\partial u/\partial\eta)$ qui est nul car, $\overline{\Phi'}$ étant une fonction analytique, donc holomorphe, de $\xi - i\eta$, $\partial\overline{\Phi'}/\partial\xi = i\partial\overline{\Phi'}/\partial\eta = d\overline{\Phi'}/d(\xi - i\eta)$.

$\Delta u = \lim_{r\to 0} \frac{1}{\pi r^2} \int_d \text{div } \text{ grad}\, u \; dxdy = \frac{1}{\pi r^2} \int_\gamma \frac{\partial u}{\partial n}\, ds$, où γ est la frontière du disque. Comme la transformation conforme conserve les angles, les dérivées normales sont transformées en dérivées normales, à un facteur $|\Phi'|$ près. On retrouve le facteur $|\Phi'|^2$ entre les laplaciens dans Ω et dans D.

Conditions frontière : si la frontière de Ω est une courbe de Jordan, Φ est continue et inversible sur les fermés $\overline{\Omega}$ et $\overline{D}$ (théorie de Carathéodory).

cf. COURANT R., *Dirichlet's principle, conformal mapping, and minimal surfaces*; GOOVAERTS M.J.; GRAGG W.B.; WUYTACK L.; TREFETHEN Lloyd N. (Editor), *Journal of Computationl and applied mathematics*, **14** (1986)1 & 2, Special issue on Numerical Conformal Mapping.

6.2. Equations intégrales sur la frontière, fonction de Green.

La solution fondamentale singulière en $P_0 \in \mathbb{R}^n$ de l'équation de Laplace vérifie $\Delta E_{P_0}(P) = \delta_{P_0}(P)$. C'est donc une fonction harmonique sauf en P_0, telle que

$$1 = \int_d \text{div } \text{ grad}\, E_{P_0}(P)\, dP = \int_\gamma \frac{\partial E_{P_0}(P)}{\partial n}\, dS$$

sur n'importe quelle boule d de centre P_0.

A une dimension, $E_{P_0}(P) = |P - P_0|/2$, à deux dimensions, $E_{P_0}(P) = \dfrac{\log|P-P_0|}{2\pi}$, à trois dimensions, $E_{P_0}(P) = -\dfrac{1}{4\pi|P-P_0|}$.

La solution du problème de Dirichlet $\Delta u = 0$ avec u donnée sur Γ est

$$\begin{aligned} u(P) &= \int_\Omega (u(P_0)\Delta E_{P_0}(P) - E_{P_0}(P)\Delta u(P_0))dP_0 = \int_\Omega \text{div}\,(u\,\text{grad}\, E_{P_0} - E_{P_0}\,\text{grad}\, u)dP_0 \\ &= \int_\Gamma \left[u(P_0)\frac{\partial E_{P_0}(P)}{\partial n} - E_{P_0}(P)\frac{\partial u(P_0)}{\partial n} \right] dS_0, \end{aligned}$$

où la fonction $\partial u/\partial n$ est encore inconnue. Si on fait tendre P vers la frontière Γ, on a une équation intégrale pour $\partial u/\partial n$. On évite la fonction singulière $\partial E/\partial n$ en remarquant que la formule donne 0 si P est hors de Ω (pas de singularité). Précisément, appliquons la formule au problème de Dirichlet dans l'extérieur de Ω. La solution u_{ext} vérifie donc

$$0 = \int_\Gamma [u_{\text{ext}}(P_0)\frac{\partial E_{P_0}(P)}{\partial n} - E_{P_0}(P)\frac{\partial u_{\text{ext}}(P_0)}{\partial n}]\, dS_0,$$

si $P \in \Omega$. Différence :

$$u(P) = \int_\Gamma E_{P_0}(P)\sigma(P_0)\, dS_0,$$

où la fonction inconnue est maintenant $\sigma(P_0)$ = saut de la dérivée normale de u à la frontière. Si P parcourt cette frontière, on a une équation intégrale (de Fredholm, première espèce) pour σ.

Résolution approchée par Galerkin ou Petrov-Galerkin : $?\sigma_h \in U_h$,

$$\int_\Gamma v(P)u(P)\,dS_P = \int_\Gamma\int_\Gamma v(P)E_{P_0}(P)\sigma_h(P_0)\,dS_0\,dS_P,\ \forall v \in V_h.$$

Cf. Boundary Integral Techniques `http ://www.rpi.edu/~des/GFEM/lec14.ppt`

Jaswon, M. A. ; Symm, G. T. *Integral equation methods in potential theory and elastostatics. Computational Mathematics and Applications*. Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York, 1977.

La ***fonction de Green*** G_{P_0} de l'équation de Laplace dans Ω, singulière en P_0, est la fonction harmonique hors de P_0, nulle sur la frontière Γ de Ω, et telle que $G_{P_0} - E_{P_0}$ soit régulière ($\mathscr{C}^1$) dans Ω.

Exemples : $\dfrac{\max(P-a)(P_0-b),(P-b)(P_0-a)}{b-a}$ dans (a,b) ;

$\dfrac{1}{2\pi}\log\left|\dfrac{P-P_0}{1-\overline{P_0}P}\right|$ dans le disque unité de $\mathbb{R}^2$;

$\dfrac{1}{4\pi}\left[\dfrac{R/|P_0|}{|P-P_0^*|}-\dfrac{1}{|P-P_0|}\right]$ dans la boule de rayon R de $\mathbb{R}^3$, où P_0^* est l'inverse (*image*) de P_0, situé dans le prolongement de P_0, à la distance $R^2/|P_0|$ de l'origine.

La solution de l'équation de Poisson $-\Delta u = f$ avec $u|_\Gamma = 0$ est $u(P) = -\displaystyle\int_\Omega G_{P_0}(P)f(P_0)\,dP_0$.

Week 8 Lectures, Math 716, Tanveer 1 Greens function for
`http ://www.math.ohio-state.edu/~tanveer/m716/week8.716.pdf`

A deux dimensions, $G_{z_0}(z) = \mathrm{Re}\ \log\left(\dfrac{\Phi(z)-\Phi(z_0)}{1-\overline{\Phi(z_0)}\Phi(z)}\right)$.

6.3. Développements multipolaires.

Discrétisation de l'équation intégrale : on a un système d'équations linéaires de forme

$$\sum_{j=1}^{N}\Phi(P_i,Q_j)x_j = y_i,\quad P_i,Q_j\in\Gamma.$$

On décompose Γ en sous-régions, disons $P_i \in \Gamma_I$ et $Q_j \in \Gamma_J$. Si Γ_J est éloignée de Γ_I, $\Phi(P_i,Q_j)$ ne varie presque pas avec Q_j et les contributions des x_j pour $Q_j \in \Gamma_J$ se réduisent à la somme $\displaystyle\sum_{Q_j\in\Gamma_J} x_j$, d'où une forte réduction du nombre d'opérations. Pour une meilleure précision, on ajoute les contributions du premier ordre, etc., de Q_j autour d'un point de Γ_J.

Exemple : solution fondamentale de l'équation de Laplace autour d'un point $R_J = (X_J, Y_J, Z_J)$:

$$\frac{1}{|P_i - Q_j|} = \frac{1}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}}$$
$$= \frac{1}{\sqrt{[(x_i - X_J) - (x_j - X_J)]^2 + [(y_i - Y_J) - (y_j - Y_J)]^2 + [(z_i - Z_J) - (z_j - Z_J)]^2}}$$
$$= \underbrace{|P_i - R_J|^{-1}}_{\text{terme monopolaire}}$$
$$+ \underbrace{[(x_i - X_J)(x_j - X_J) + (y_i - Y_J)(y_j - Y_J) + (z_i - Z_J)(z_j - Z_J)]|P_i - R_J|^{-3}}_{\text{termes dipolaires}} + \cdots$$

Enfin, on considère simultanément plusieurs niveaux de discrétisation.
`www-theor.ch.cam.ac.uk/people/ross/thesis/node97.html`

The Fast Multipole Method.

The breakthrough was made in 1985 when Greengard and Rokhlin [178,179,180] showed how to compute the Coulomb energy of point charges in only linear work. Greengard's Fast Multipole Method (FMM) belongs to the family of algorithms called tree codes. Tree codes [181] acquire their speed by transforming the information about a cluster of charge into a simpler representation which is used to compute the influence of the cluster on objects at large distances.

The Fast Multipole Method begins by scaling all particles into a box with coordinate ranges $[0 \ldots 1, 0 \ldots 1, 0 \ldots 1]$ to ensure numerical stability of subsequent operations. The parent box is then divided in half along each Cartesian axis. Each child box is then further subdivided, forming a 'computational family tree'. The number of subdivisions is chosen so that the number of particles in the lowest level boxes is approximately independent of the number of particles (this is required to achieve linear scaling).

Each particle is then placed within a box on the lowest level of the tree. Any empty boxes are removed to allow efficiency for inhomogeneous systems. The charges of particles within each lowest level box are then expanded in multipoles about the center of its box. The multipole expansion of each lowest level box is then translated to the center of its parent box via one of three special FMM operators.

The second of the FMMs operators is used for each box to transform the multipole expansions of all well-separated boxes (those that are not nearest neighbours) into Taylor expansions about the center of the current box. However, only those multipole expansions from boxes which are well-separated at the present level and not well-separated at the parent level are interacted. The multipole expansions are also translated to Taylor expansions in the parent box. This allows each transformation to be performed as high up the tree as possible. In practice this pass is the bottleneck of the algorithm, yet, like all the other passes, it is O(N).

The third pass transforms the parent Taylor expansions down the tree to the child boxes, so that each low-level box contains the Taylor expansion representing all well-separated particles. The fourth pass calculates the far-field potential for each particle via the Taylor representation in the particles box. A final pass calculates the interactions between particles that are not well-separated at any level in the tree.

The Continuous Fast Multipole Method.

The FMM has subsequently been applied to problems in astrophysics, plasma physics, molecular dynamics, fluid dynamics, partial differential equations and numerical complex analysis. The FMM was generalized to handle continuous distributions (forming the Continuous FMM, CFMM) in 1994 by White et al. [182,157] after making several improvements to the original FMM [180,183,184]. The main change was the introduction of a well-separated index, describing the distance required before interactions can be calculated via multipoles. This index depends on the diffuseness of the charge distributions involved. Over the last few years CFMM has become a very mature algorithm, and the Q-CHEM implementation is now a highly optimized, efficient code.

Ross D. Adamson 1999-01-27

157 C. A. White, B. G. Johnson, P. M. W. Gill, M. Head-Gordon, Chem. Phys. Lett. 253 (1996) 268.
178 L. Greengard, V. Rokhlin, J. Comput. Phys. 73 (1987) 325.

179 L. Greengard, The Rapid Evaluation of Potential Fields in Particle Systems, MIT, Cambridge, (1987).
180 C. A. White, M. Head-Gordon, J. Chem. Phys. 101 (1994) 6593.
181 L. Greengard, Science 265 (1994) 909.
182 C. A. White, B. G. Johnson, P. M. W. Gill, M. Head-Gordon, Chem. Phys. Lett. 230 (1994) 8.
183 C. A. White, M. Head-Gordon, Chem. Phys. Lett. 257 (1996) 647.
184 C. A. White, M. Head-Gordon, J. Chem. Phys. 105 (1996) 5061.

7. Décomposition de domaine, complément de Schur, substructuring.

On considère une partition de Ω en p sous-domaines.

Que ce soit en éléments finis ou en différences, la plupart des éléments non nuls de la $i^{\text{ème}}$ ligne de $\boldsymbol{Ax} = \boldsymbol{b}$ sont associés à des points appartenant au même sous-domaine que le $i^{\text{ème}}$ point.

En rassemblant les indices associés à chaque sous-domaine, on obtient

$$\begin{bmatrix} \boldsymbol{A}_1 & & & \boldsymbol{F}_1^T \\ & \ddots & & \vdots \\ & & \boldsymbol{A}_p & \boldsymbol{F}_p^T \\ \boldsymbol{F}_1 & \cdots & \boldsymbol{F}_p & \boldsymbol{A}_B \end{bmatrix} \begin{bmatrix} \boldsymbol{x}^{(1)} \\ \vdots \\ \boldsymbol{x}^{(p)} \\ \boldsymbol{x}^{(B)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}^{(1)} \\ \vdots \\ \boldsymbol{b}^{(p)} \\ \boldsymbol{b}^{(B)} \end{bmatrix},$$

où l'ordre du bloc $\boldsymbol{A}_B$ est normalement très inférieur à la somme des ordres des autres blocs.

Donc, $\boldsymbol{x}^{(k)} = \boldsymbol{A}_k^{-1}\boldsymbol{b}^{(k)} - \boldsymbol{A}_k^{-1}\boldsymbol{F}_k^T\boldsymbol{x}^{(B)} \Rightarrow \boldsymbol{S}\boldsymbol{x}^{(B)} = \boldsymbol{b}^{(B)} - \sum_{k=1}^{p} \boldsymbol{F}_k\boldsymbol{A}_k^{-1}\boldsymbol{b}^{(k)}$, avec $\boldsymbol{S} = \boldsymbol{A}_B - \sum_{k=1}^{p} \boldsymbol{F}_k\boldsymbol{A}_k^{-1}\boldsymbol{F}_k^T$.

On appelle ***complément de Schur*** de $\boldsymbol{X}$ dans $\begin{bmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{Z} & \boldsymbol{W} \end{bmatrix}$ la matrice $\boldsymbol{U}$ telle que $\begin{bmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{Z} & \boldsymbol{W} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{a} \\ \boldsymbol{b} \end{bmatrix} \Rightarrow \boldsymbol{U}\boldsymbol{y} = \cdots$ On trouve $\boldsymbol{U} = \boldsymbol{W} - \boldsymbol{Z}\boldsymbol{X}^{-1}\boldsymbol{Y}$. La matrice $\boldsymbol{S}$ est donc une somme de compléments de Schur.

Dans les formules ci-dessus, les expressions utilisant $\boldsymbol{A}_k^{-1}$ s'entendent comme des résolutions de sous-systèmes de matrice $\boldsymbol{A}_k$. Ces résolutions peuvent s'effectuer en ***parallèle*** sur p processeurs.

Voir aussi méthodes frontales.

Cf. B.F. Smith, P.E. Bjørstad, W.D. Gropp, *Domain Decomposition, Parallel multilevel methods for elliptic partial differential equations*, Cambridge Univ. Press, 1996 ;
S.C. Brenner, The condition number of the Schur complement in domain decomposition, *Numer. Math.* **83** (1999) 187-203.

From : Thanh-Ha.Lethi@springer-sbm.com
Date : Wed, 27 Oct 2004 11 :18 :21 +0200
Subject : New Book on Domain Decomposition Methods

Andrea Toselli, Olof Widlund *"Domain Decomposition Methods - Algorithms and Theory"* - Springer Series in Computational Mathematics Vol. **34**, XV, 450 pp. ISBN 3-540-20696-5, Springer-Verlag, October 2004

The purpose of this text is to offer a comprehensive and self-contained presentation of some of the most successful and popular domain decomposition preconditioners for finite and spectral element approximations of partial differential equations. Strong emphasis is placed on both algorithmic and mathematical aspects. Some important methods such FETI and balancing Neumann-Neumann methods and algorithms for spectral element methods, not treated previously in any monograph, are covered in detail.

For further details, please contact

Dr. Martin Peters Excutive Mathematics Editor Springer Tiergartenstrasse 17 D-69121 Heidelberg, Germany Email : Martin.Peters@springer-sbm.com

8. Conditionnement et méthodes itératives, préconditionnement.

On résout $Ax = b$ par itérations ne contenant que des additions et des multiplications par $A \Rightarrow$ le $n^{\text{ème}}$ itéré est $P_n(A)b$, où P_n est un polynôme de degré n. Le ***polynôme résidu*** est $R_{n+1}(x) := 1 - xP_n(x)$. Le $n^{\text{ème}}$ résidu est alors $b - Ax = b - AP_n(A)b = R_{n+1}(A)b$.

Si on se donne des polynômes R_n, avec $R_n(0) = 1$, on décrit une méthode itérative. Par exemple, ***Jacobi-Richardson*** :

$$R_n(z) = (1-\mu z)^n \Rightarrow x = \frac{I - (I - \mu A)^{n+1}}{A}\, b : x_{\text{nouveau}} - x_{\text{ancien}} = \mu(I-\mu A)^n b = \mu(b - Ax_{\text{ancien}}).$$

Il s'agit de contrôler la décroissance de $\|R_{n+1}(A)b\|$ en fonction de n. Si A est symétrique (réelle[19]), on voit, en exprimant b dans la base ***orthonormale*** des vecteurs propres de A, que

$$\|R_{n+1}(A)\| = \max_k |R_{n+1}(\lambda_k(A))|.$$

Si on sait que A est définie positive, $\lambda_k \in [\alpha, \beta]$, avec $0 < \alpha \leqslant \beta$, et il suffit d'examiner $\max_{z\in[\alpha,\beta]} |R_{n+1}(z)|$.

Avec Jacobi-Richardson et $\mu = 2/(\alpha + \beta)$, on obtient $\|R_{n+1}(A)\| \leqslant \left(\dfrac{\beta - \alpha}{\beta + \alpha}\right)^{n+1}$.

Conditionnement de A : $\kappa = \lambda_{\max}(A)/\lambda_{\min}(A)$. Donc, au mieux, $\|R_{n+1}(A)\| \leqslant \left(\dfrac{\kappa - 1}{\kappa + 1}\right)^{n+1}$.

Préconditionnement : extraits de

ITERATIVE METHODS FOR LARGE SCALE SYSTEMS AND EIGENVALUE PROBLEMS.

This seminar series represents the graduate courses "Special Topics in Linear Algebra", P. Van Dooren (UCL) and "Scientific Computing", S. Vandewalle (KULeuven).

...

Friday 6/3 1998, 14h-16h30, K.U.Leuven, sem. room 200B.02.16
Preconditioning techniques for iterative methods (Notay, Univ. Libre Bruxelles)

...

Preconditioning techniques for iterative methods

Y. Notay *
Université Libre de Bruxelles
* "Fonds National de la Recherche Scientifique", Chercheur qualifié.

1. Introduction

Consider a Krylov method to solve the linear system $Au = b$. The convergence depends essentially on the eigenvalue distribution of A. Ideally, all eigenvalues should be clustered around some point on the real axis. If A is Symmetric Positive Definite (SPD), all eigenvalues are real and positive, and the convergence rate of the Krylov method depends mainly on the condition number $\kappa(A) = \dfrac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$. In this case, the conjugate gradient method is the most widely used (optimal with short recurrence).

A ***preconditioner*** is a matrix B such that the preconditioned system $B^{-1}Au = B^{-1}b$ is easier to solve by a Krylov method.

This means that

[19] ou complexe hermitienne

– the eigenvalue distribution of $B^{-1}A$ should be more favourable (better clustered) ;
– solving a linear system $Bx = y$ should be cheap (each iteration requires the multiplication of a vector by B^{-1}).

If A is SPD, it is better to choose B SPD. The CG method can then still be used. The eigenvalues of $B^{-1}A$ are positive real and the convergence depends on $\kappa(B^{-1}A) = \dfrac{\lambda_{\max}(B^{-1}A)}{\lambda_{\min}(B^{-1}A)}$.

In general, it is more important to use a good preconditioner than to select an "optimal" Krylov method.

Illustrations

(1) Discrete Laplacian with Dirichlet boundary conditions, 5 point finite difference, uniform 128×128 grid

prec.	κ	it(s. d.)$^{(*)}$	it(CG)
No	6640.	61157	237
ILU(0)	587.	5407	99
MILU(0)	40.9	377	54
AML	2.70	24	15

$^{(*)}$ steepest descent, theoretical estimates

(2) Non symmetric problems, from Saad (1996).

Mflops to solve the problem :

matrix	GMRES(10)	BICGSTAB	SGS	ILU(0)	ILUT
F2DA	3.8	2.0	2.0	1.5	1.0
D3D	11.9	6.4	4.9	4.0	3.4
ORS	9.2	5.2	6.8	1.2	0.3

(preconditioned cases with GMRES(10))

2. Basic iterative techniques

Reminder : given a splitting $A = M - N$, the basic iteration writes $Mu_{k+1} = b + Nu_k$.

Example : $M = \mathrm{diag}(A)$ (Jacobi) gives $(u_{k+1})_i = a_{ii}^{-1}\left(b_i - \sum_{j\neq i} a_{ij}(u_k)_j\right)$.

Convergence : let $\epsilon_k = u - u_k$; then $\epsilon_k = (I - M^{-1}A)\epsilon_{k-1} = (I - M^{-1}A)^k\epsilon_0$ Fast if $\sigma(M^{-1}A)$ well clustered around $1 \Rightarrow M$ is a good preconditioner for A.

Rmk : the sequence $\{\tilde{u}_k\}$ resulting from the use of a Krylov method is connected to $\{u\}_k$ defined above by $\{\tilde{u}_k\} = \sum_{i=0}^{k} \alpha_k^i u_i$.

3. Non algebraic preconditioners

Physical problem (say, PDE) $\xrightarrow{\text{discretization}} Au = b$

(Simplified) problem $\xrightarrow[\text{discretization}]{\text{(simplified)}} B\tilde{u} = \tilde{b}$

If the "simplifications" entail that $B\tilde{u} = \tilde{b}$ is easier to solve, B can be used as preconditioner for A.

Examples

PDE with smoothly variable coefficient "preconditioned" by a PDE with constant coefficients.

Spectral finite elements ($\rightarrow$ relatively dense matrices) "preconditioned" by linear finite elements ($\rightarrow$ sparse matrices).

4. Two-step preconditioning Inner-outer schemes

"$B\tilde{u} = \tilde{b}$ is easier to solve" may in fact mean that an efficient (algebraic) preconditioning technique is applicable to B. ("easier to solve by a Krylov method").

The resulting preconditioner $\tilde{B}$ can be used directly as preconditioner for A.

Alternatively, instead of applying directly $\tilde{B}$ to precondition A, one may effectively precondition A by B, and solve the subsidiary systems with B by an (inner) iterative process with $\tilde{B}$ as preconditioner.

In general, it does not pay off to make exact solves, so that the preconditioner for A is not exactly B, but some 'near' mapping $\tilde{B}$.

Inner-outer schemes are attractive when

– B is much sparser than A ;
– Both inner and outer processes converge in a few iterations, whereas the direct scheme requires truncation of the Krylov method (non SPD case ; sometimes used with $B = A$).

...

HOWEVER, two assumptions are used which are seldom met in practice :

(1) the relation holds for any w : in practice, the error is controlled only with respect to the current iterate ;

(2) if a Krylov method is applied to solve the systems with B, in general, given $y_1, y_2, \ldots$, no $\widehat{B}$ exists such that the resulting approximate solutions satisfies $x_1 = \widehat{B}^{-1}y_1, x_2 = \widehat{B}^{-1}y_2, \ldots$ For instance, the computed solution for $y = y_1 + y_2$ will generally not be $x_1 + x_2$. Further, even is such a $\widehat{B}^{-1}$ formally exists, it will surely not be SPD.

Therefore :
– one should choose, for the outer process, a Krylov method that explicitly accommodates for variable preconditioning (in particular, avoid the application of an unpreconditioned scheme to the transformed system $\widehat{B}^{-1}A = \widehat{B}^{-1}b$) ;
– It is difficult to predict which ϵ will realize the best compromise.

5. Incomplete factorization preconditioning

Overview of algebraic preconditioning techniques
– From classical iterative methods :

Jacobi, Gauss-Seidel, SOR, SSOR (theory of convergent splittings, see under-graduate courses). Always applicable, often poor convergence except when A itself is well conditioned.
– Incomplete LU (ILU) for general matrices : widely applicable, unpredictable convergence.
– Incomplete LU, specific schemes for (symmetric) $M-$matrices : restricted applicability relatively fast convergence.
– Algebraic multilevel methods : very restricted applicability, very fast convergence.

ILU (general)

LU decomposition of A : $A = LP^{-1}U (P = \mathrm{diag}(L) = \mathrm{diag}(U))$ One seeks an iterative scheme because this decomposition is too costly to compute (L, U much less sparse than A).

Incomplete LU decomposition : $A = LP^{-1}U + R$. The remainder matrix R is nonzero because, when computing the decomposition, 'fill' entries are discarded to keep L, U (nearly) as sparse as A. $B = LP^{-1}U$ is then a preconditioner.

Algorithm

Initialization : $L = \mathrm{low}(A), U = \mathrm{upp}(A), P = \mathrm{diag}(A)$.

For $k = 1, \ldots, n-1$:

$\forall i, j > k : \ell_{ik} \cdot u_{kj} \neq 0$

$f_{ij} = -\dfrac{\ell_{ik} u_{kj}}{p_{kk}}$

if (not discard (f_{ij}) or $i = j$)

$$\begin{cases} \text{if } i < j : u_{ij} & = u_{ij} + f_{ij} \\ \text{if } i = j : p_{ii} & = p_{ii} + f_{ij} \\ \text{if } i > j : \ell_{ij} & = \ell_{ij} + f_{ij} \end{cases}$$

'discarding' rules :
– by position (depends only of i and j, essentially for structured matrices) ;
– by value : $|f_{ij} \leqslant \tau m_{ij}$, where τ is a threshold parameter and $m_{ij} = \min(|a_{ii}|, |a_{jj}|)$ (for instance).

General matrices
– If a pivoting strategy is required for the stability of the exact LU decomposition, a similar strategy should be implemented for ILU.
– If A is SPD, the positive definiteness of B (that is, $p_{ii} > 0\ \forall i$) may be guaranteed by adding the discarded $|f_{ij}|$ to the diagonal, but this has an adverse effect on the conditioning.
– In general, one tries to prevent problems by selecting a small threshold value τ.

However :

– fill-in is then unpredictable (→ strategies that fix the maximumnumber of nonzero per row of U (column of L))

–

$$\frac{(z,Az)}{(z,Bz)} = 1 + \frac{(z,Rz)}{(z,Bz)}$$

if A is ill conditioned (and B close to A), $\left|\frac{(z,Rz)}{(z,Bz)}\right|$ may be large even when $\|R\|$ is small.

Rmk : ordering is important.

M-matrices

M-matrices have nonpositive offdiagonal entries (offdiag$(A) \leqslant 0$).

If A is nonsingular and such that offdiag$(A) \leqslant 0$, then

$$\begin{aligned} A\text{is an M-matrix} &\iff A^{-1} > 0 \\ &\iff \text{real}\,(\lambda) > 0,\ \forall \lambda \in \sigma(A) \\ &\iff \frac{1}{2}(A + A^t) \text{ is SPD} \\ &\iff \exists x > 0 : Ax \geqslant 0 \end{aligned}$$

Basic stable discretization schemes applied to second order PDEs lead to M-matrices with nonnegative row-sum ($Ae \geqslant 0, e = (11\ldots1)^t$).

If a "non basic" scheme is used, the basic scheme may be considered to define a two-step preconditioner.

Rmk : analysis of classical iterative methods is essentially based on M-matrix theory.

Chapitre 6

Schémas de différences finies : problèmes d'évolution.

On aborde des problèmes dont la solution $u(x,t)$ dépend d'une variable temporelle et une ou plusieurs variables spatiales. Il est question de déterminer $u(x,t)$ par pas de Δt en t : $u(x,0)$, $u(x,\Delta t)$, $u(x,2\Delta t)$,...

Complainte du Temps et de sa commère l'Espace
J. Laforgue, 1880

"L'espace est un garçon solide et simple, tout d'une pièce, sans le mondre détour...Il règne sur un domaine que vous avez le droit de parcourir en tout sens...

Le caractère du temps est autrement difficile...C'est un personnage...d'une instabilité maladive..."
J. d'Ormesson, 1996

1. Equation de la chaleur ; de la diffusion.

1.1. Equation de la chaleur.

Problème à une dimension spatiale : étudions la répartition de la température sur un segment $[0,\ell]$ conducteur de chaleur. Si $u_i(t) = u(ih,t)$ est la température du $i^{\text{ème}}$ sous-segment de longueur h, il se crée des flux de chaleur entre sous-segments de températures différentes, par conduction thermique :

densité de flux de chaleur par unité de temps $= -K$ **grad** u.

Le $i^{\text{ème}}$ sous-segment reçoit donc de la part de ses deux voisins une quantité de chaleur par unité de temps égale à $q_i = K\left(\dfrac{u_{i+1}-u_i}{h} + \dfrac{u_{i-1}-u_i}{h}\right)$, à quoi il faut ajouter $hF(x_i,t)$ si le segment est soumis à une source de chaleur de densité F.

D'autre part, le sous segment réagit à la réception de la quantité de chaleur q_i par unité de temps en augmentant sa température selon la loi $q_i\Delta t = Ch(u_i(t+\Delta t) - u_i(t))$, où C est la chaleur spécifique du conducteur.

On a donc

$$\frac{u_i(t+\Delta t)-u_i(t)}{\Delta t} = \sigma\frac{u_{i+1}(t)-2u_i(t)+u_{i-1}(t)}{h^2} + f(x_i,t), \tag{59}$$

avec $\sigma = K/C$, $f = F/C$.

On a obtenu une discrétisation très naturelle de ***l'équation de la chaleur***

$$\frac{\partial u(x,t)}{\partial t} = \sigma\frac{\partial^2 u(x,t)}{\partial x^2} + f(x,t), \tag{60}$$

où $\sigma > 0$.

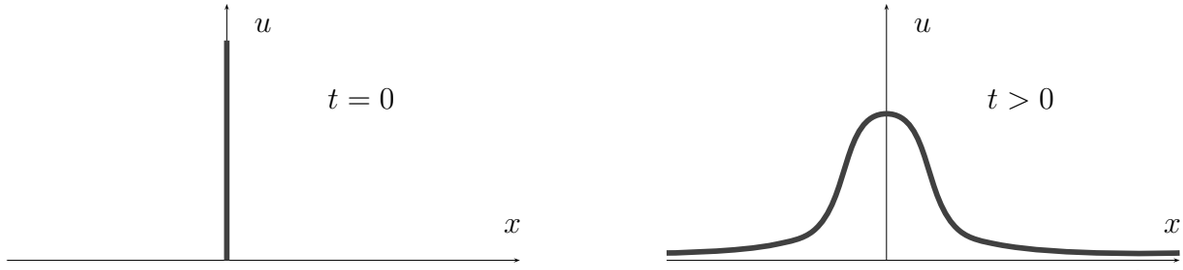
Si K dépend de x : $\dfrac{\partial u(x,t)}{\partial t} = \dfrac{1}{C}\dfrac{\partial}{\partial x}\left(K(x)\dfrac{\partial u(x,t)}{\partial x}\right) + f(x,t)$.

A plusieurs dimensions spatiales : $\dfrac{\partial u(\boldsymbol{x},t)}{\partial t} = \dfrac{1}{C}\,\mathbf{div}\ (K(\boldsymbol{x})\ \mathbf{grad}\ u(\boldsymbol{x},t)) + f(\boldsymbol{x},t)$.

Conditions aux limites : $\partial u/\partial n = 0$ (Neumann) là où le flux de chaleur est nul (isolation) ; u donné (Dirichlet) sur les parties de la frontière de température connue.

1.2. Solution par noyau de Poisson.

Constatons que l'équation (60) homogène admet la solution $\dfrac{1}{\sqrt{4\pi\sigma(t+t_0)}}\ \exp\left(-\dfrac{(x-x_0)^2}{4\sigma(t+t_0)}\right)$ (noyau de Poisson). C'est une gaussienne qui "s'avachit" au cours du temps (cf. p. 24) :



Si $t_0 = 0$, cette solution tend vers le delta de Dirac $\delta(x-x_0)$ quand $t \to 0$ en décroissant[1], d'où une formule pour la solution de l'équation homogène

$$u(x,t) = \frac{1}{\sqrt{4\sigma\pi t}} \int_{-\infty}^{\infty} \exp\left(-\frac{(y-x)^2}{4\sigma t}\right)\ u(y,0)\,dy. \tag{61}$$

Pour $t > 0$ fixé, on voit ainsi un exemple ***d'opérateur solution*** $E(t)$ fournissant la solution au temps t à partir de la condition initiale : $u(.,t) = E(t)u(.,0)$. Bien entendu, on ne disposera pas toujours (équations à coefficients variables, domaines plus compliqués, non-linéarités) d'une formule aussi explicite...

1.3. Equation de la diffusion ; diffusion des euros.

Equation de la diffusion : $u(x,t)$ = concentration d'une substance dans un solvant, K = taux de diffusion par unité de temps. On retrouve (60) ($C = 1 : \sigma = K$).

Nous reprenons le problème (3) de la page 18

$$\frac{\partial y(x,t)}{\partial t} = \sigma \frac{\partial^2 y(x,t)}{\partial x^2}$$

sur tout $\mathbb{R}$, avec $u(x,0) = 1$ si $-L/2 < x < L/2$, $u(x,0) = 0$ ailleurs.

Poisson (61) :

$$y(x,t) = \frac{1}{\sqrt{4\sigma\pi t}} \int_{-L/2}^{L/2} \exp\left(-\frac{(y-x)^2}{4\sigma t}\right)\ dy.$$

[1] La fonction $\dfrac{1}{\sqrt{4\pi\sigma t}}\ \exp\left(-\dfrac{(x-x_0)^2}{4\sigma t}\right)$ est positive, d'intégrale définie sur $\mathbb{R}$ égale à l'unité, et pour tout $\epsilon > 0$, on peut trouver τ tel que l'intégrale sur $|x-x_0| > \epsilon$ soit $< \epsilon$ quand $t \leqslant \tau$ (si $|x-x_0| > \epsilon$, $(x-x_0)^2 > \epsilon|x-x_0|$, donc $\exp(-(x-x_0)^2/(4\sigma t)) < \exp(-\epsilon|x-x_0|/(4\sigma t))$ que l'on intègre aisément sur $|x-x_0| > \epsilon$ pour obtenir une intégrale finale bornée supérieurement par $\dfrac{4}{\epsilon}\sqrt{\dfrac{\sigma t}{\pi}}\exp\left(-\dfrac{\epsilon^2}{4\sigma t}\right)$ croissante en t et nulle en $t = 0$).

$$= \int_{-L/2}^{L/2} \sqrt{\frac{t_0}{2d_0^2\pi t}} \exp\left(-\frac{t_0(x-x_0)^2}{2d_0 t}\right)\, dx_0 = \frac{\operatorname{erf}\left(\sqrt{\frac{1}{4\sigma t}}\left(\frac{L}{2}-x\right)\right) + \operatorname{erf}\left(\sqrt{\frac{1}{4\sigma t}}\left(\frac{L}{2}+x\right)\right)}{2}$$

où $\operatorname{erf}(X) := \frac{2}{\sqrt{\pi}} \int_0^X e^{-u^2}\, du$. Remarquons que

$$\frac{\partial y(x,t)}{\partial x} = \sqrt{\frac{1}{4\sigma\pi t}} \left[\exp\left(-\frac{(L/2+x)^2}{4\sigma t}\right) - \exp\left(-\frac{(L/2-x)^2}{4\sigma t}\right)\right]$$

Au temps $t > 0$, la proportion encore disponible dans le pays d'origine est

$$Y(t) = \frac{1}{L} \int_{-L/2}^{L/2} y(x,t)\, dx$$

Par (3), on a aussi

$$\frac{dY(t)}{dt} = 2\frac{\sigma}{L}\frac{\partial y(L/2,t)}{\partial x} = \sqrt{\frac{\sigma}{L^2\,\pi t}} \left(\exp\left(-\frac{L^2}{4\sigma t}\right) - 1\right)$$

Cette moyenne nationale décroît avec t selon une fonction universelle F de $\sigma t/L^2$, avec

$$\sigma := d_0^2/(2t_0). \tag{62}$$

$$Y(t) = F(\sigma t/L^2); \qquad F(X) = \sqrt{\frac{4X}{\pi}} \left(\exp\left(-\frac{1}{4X}\right) - 1\right) + \operatorname{erf}\left(\sqrt{\frac{1}{4X}}\right).$$

1.3.1. *Effect of adding fresh coins.* After about 6 months, the decreasing trend seems to have died.

We consider that during each exchange period t_0, the national bank makes available a fraction c of fresh national coins. As the number of various coins in each pocket is still supposed constant, the same fraction c of owned coins is removed, so that (2) is now

$$y_i(t+t_0) = (1-c)\frac{y_{i-1}(t) + y_{i+1}(t)}{2} + c, \qquad -L/2 \leqslant x \leqslant L/2, \tag{63}$$

whereas fresh coins brought abroad are of course of THEIR own kind, and our own brand decreases even faster :

$$y_i(t+t_0) = (1-c)\frac{y_{i-1}(t) + y_{i+1}(t)}{2}, \qquad |x| > L/2, \tag{64}$$

In the continuous limit :

$$\frac{\partial y(x,t)}{\partial t} = \sigma\frac{\partial^2 y(x,t)}{\partial x^2} - \gamma\, y(x,t) + \gamma\, \chi_{(-L/2,L/2)}(x), \tag{65}$$

with $\gamma = c/t_0$, and where $\chi_{(a,b)}(x) = 1$ when $a < x < b$ and vanishes elsewhere.

Limit solution when $t \to \infty$:

$$y(x,\infty) = 1 - \exp\left(-\sqrt{\frac{\gamma}{\sigma}}\,\frac{L}{2}\right)\cosh\left(\sqrt{\frac{\gamma}{\sigma}}\,x\right), \qquad -L/2 \leqslant x \leqslant L/2,$$
$$= \sinh\left(\sqrt{\frac{\gamma}{\sigma}}\,\frac{L}{2}\right)\exp\left(-\sqrt{\frac{\gamma}{\sigma}}\,|x|\right), \qquad |x| \geqslant L/2.$$

$$\text{also} = \chi_{(-L/2,L/2)}(x) + \sinh\left(\sqrt{\frac{\gamma}{\sigma}}\frac{L}{2}\right)\cosh\left(\sqrt{\frac{\gamma}{\sigma}}\,x\right)$$
$$-\frac{1}{2}\left[\cosh(\sqrt{\gamma/\sigma}(L/2+x))\text{sign}(L/2+x) + \cosh(\sqrt{\gamma/\sigma}(L/2-x))\text{sign}(L/2-x)\right]$$

Average on $(-L/2, L/2)$:

$$G(\infty) := \frac{1}{L}\int_{-L/2}^{L/2} y(x,\infty)\,dx = 1 - \frac{1-\exp\left(-\sqrt{\frac{\gamma}{\sigma}}\,L\right)}{\sqrt{\frac{\gamma}{\sigma}}\,L} \tag{66}$$

In http ://www.senat.fr/rap/l01-087-344/l01-087-34429.html, it is stated that the French bank intended to release about 6.5 10^9 coins on January 2002, followed by a regular flow of about 3 10^8 coins each month. This means γ about 0.05 month^{-1}. Assuming the same value here, and with σ/L^2 near to 0.002, we have $\gamma L^2/\sigma \approx 25$. The value of γ may decrease with time...

It figures : if $G(\infty) = \dfrac{1}{L}\displaystyle\int_{-L/2}^{L/2} y(x,\infty)\,dx$, we must consider G^2 for a two-dimensional problem, and here are some limit values, from (66) :

$(\gamma L^2/\sigma)$	1	2	3	4	5	10	15	20	25	30	50
$G(\infty)$	0.368	0.465	0.525	0.568	0.601	0.697	0.747	0.779	0.801	0.818	0.859
$G^2(\infty)$	0.135	0.216	0.275	0.322	0.361	0.486	0.558	0.607	0.642	0.669	0.737

Kernel : $\dfrac{1}{\sqrt{4\sigma\pi t}}\exp\left(-\dfrac{(z-x)^2}{4\sigma t} - \gamma t\right)$.

Solution :

$$y(x,t) = y(x,\infty) + \frac{e^{-\gamma t}}{\sqrt{4\sigma\pi t}}\int_{-\infty}^{\infty}\exp\left(-\frac{(z-x)^2}{4\sigma t}\right)\left[\chi_{(-L/2,L/2)}(z) - y(z,\infty)\right]dz$$
$$= y(x,\infty) + \sinh(\sqrt{\gamma/\sigma}\,L/2)\cosh(\sqrt{\gamma/\sigma}\,x) - \frac{1}{4}\left[e^{-\sqrt{\gamma/\sigma}(L/2-x)}\text{erf}\left(\frac{L/2-x}{\sqrt{4\sigma t}} - \sqrt{\gamma t}\right)\right.$$
$$\left. + e^{\sqrt{\gamma/\sigma}(L/2-x)}\text{erf}\left(\frac{L/2-x}{\sqrt{4\sigma t}} + \sqrt{\gamma t}\right) + e^{\sqrt{\gamma/\sigma}(L/2+x)}\text{erf}\left(\frac{L/2+x}{\sqrt{4\sigma t}} + \sqrt{\gamma t}\right) + e^{-\sqrt{\gamma/\sigma}(L/2+x)}\text{erf}\left(\frac{L/2+x}{\sqrt{4\sigma t}} - \sqrt{\gamma t}\right)\right]$$

in $x \in (-L/2, L/2)$.

Average on $(-L/2, L/2)$:

$$G(t) := \frac{1}{L}\int_{-L/2}^{L/2} y(x,t)\,dx = G(\infty) + \sqrt{\frac{\sigma}{4\gamma L^2}}\left[e^{\sqrt{\gamma L^2/\sigma}}\text{erf}\left(\frac{L}{\sqrt{4\sigma t}} + \sqrt{\gamma t}\right) - \text{erf}(\sqrt{\gamma t})\right.$$
$$\left. - e^{-\sqrt{\gamma L^2/\sigma}}\text{erf}\left(\frac{L}{\sqrt{4\sigma t}} - \sqrt{\gamma t}\right) - \text{erf}(\sqrt{\gamma t})\right]$$

Simple discretization of (65) :

$$y_i(t+\Delta t) = y_i(t) + \sigma\Delta t\frac{y_{i-1}-2y_i+y_{i+1}}{h^2} - \gamma\frac{y_{i-1}+y_{i+1}}{2} + \gamma\Delta t\chi_i$$

1.4. Modes et séries de Fourier.

Comme Fourier lui-même[2], examinons comment $A(t)\sin(\pi mx/\ell)$ peut être solution de (60) dans $x \in (0,\ell)$: si $f = 0$, on trouve $dA(t)/dt = -\sigma\pi^2m^2/\ell^2A(t)$, d'où $A(t) = A(0)\exp(-\sigma\pi^2m^2t/\ell^2)$.

On peut donc résoudre (60) sur $(0,\ell)\times[0,\infty)$, avec les conditions aux limites $u(0,t) = u(\ell,t) = 0$ par une série de Fourier

$$u(x,t) = \sum_{m=1}^{\infty} A_m(t)\sin(\pi mx/\ell), \tag{67}$$

à partir des séries du second membre et de la condition initiale $u(x,0)$

$$f(x,t) = \sum_{m=1}^{\infty}\varphi_m(t)\sin(\pi mx/\ell),\ u(x,0) = \sum_{m=1}^{\infty}A_m(0)\sin(\pi mx/\ell),$$

en résolvant les équations différentielles ordinaires

$$\frac{dA_m(t)}{dt} = -\frac{\sigma\pi^2m^2}{\ell^2}A_m(t) + \varphi_m(t),\ m = 1,2,\dots$$

En particulier, si$f = 0$, la série de la solution est

$$u(x,t) = \sum_{m=1}^{\infty}A_m(0)\exp(-\sigma\pi^2m^2t/\ell^2)\sin(\pi mx/\ell),$$

Remarquons que u est indéfiniment dérivable en x dès que $t > 0$!

1.5. Exemples de stabilité et instabilité numérique.

Tentons de résoudre (60) sur $(0,\ell)\times[0,\infty)$ avec $f = 0$, les conditions aux limites $u(0,t) = u(\ell,t) = 0$ et la condition initiale $u(x,0) = 4x(\ell-x)$:

On applique (59) avec $h = 1/10$.

```
10   dim Te(100),Tf(100):print=print+"parabte.out"
20   N=10:H=1/N:for I=1 to N+1:X=(I-1)*H:Te(I)=4*X*(1-X):next I
30   T=0:Dt=0.1
40   print " delta t= ";using(2,4),Dt
50   while T<=1
55    if abs(round(10*T)-10*T)<0.001 then
60     :print using(2,2),T;" ";:for I=2 to 10:print using(2,4),Te(I);" ";:
65     :next I:print
80    for I=2 to N:Tf(I)=Te(I)+Dt*(Te(I+1)-2*Te(I)+Te(I-1))/H^2:next I
85    for I=2 to N:Te(I)=Tf(I):next I
90   T=T+Dt:wend
```

On lance le schéma avec $\Delta t = 0.1$:

[2]J. Fourier, *Théorie analytique de la chaleur*, Didot, 1822, = *Analytical Theory of Heat*, (transl. A. Freeman), Dover, 1955.

```
delta t=  0.1000
 t     x=0.1    0.2     0.3     0.4     0.5     0.6     0.7     0.8     0.9

0.00  0.3600  0.6400  0.8400  0.9600  1.0000  0.9600  0.8400  0.6400  0.3600
0.10 -0.4400 -0.1600  0.0400  0.1600  0.2000  0.1600  0.0400 -0.1600 -0.4400
0.20  6.7600 -0.9600 -0.7600 -0.6400 -0.6000 -0.6400 -0.7600 -0.9600  6.7600
0.30 -?.----  ?.---- -1.5600 -1.4400 -1.4000 -1.4400 -1.5600  ?.---- -?.----
0.40  ?.---- -?.----  ?.---- -2.2400 -2.2000 -2.2400  ?.---- -?.----  ?.----
0.50 -?.----  ?.---- -?.----  ?.---- -3.0000  ?.---- -?.----  ?.---- -?.----
...
```

Que se passe-t-il ? Des valeurs plus petites de Δt sont expérimentées, jusqu'à 0.01 :

```
delta t=  0.0100
0.00  0.3600  0.6400  0.8400  0.9600  1.0000  0.9600  0.8400  0.6400  0.3600
0.10  ?.---- -?.----  ?.---- -?.----  ?.---- -?.----  ?.---- -?.----  ?.----
...
```

C'est encore pire ! Pourtant, on a raison de s'obstiner : avec $\Delta t = 5\ 10^{-3}$,

```
delta t=  0.0050
0.00  0.3600  0.6400  0.8400  0.9600  1.0000  0.9600  0.8400  0.6400  0.3600
0.10  0.1170  0.2222  0.3062  0.3595  0.3785  0.3595  0.3062  0.2222  0.1170
0.20  0.0429  0.0814  0.1122  0.1318  0.1387  0.1318  0.1122  0.0814  0.0429
0.30  0.0157  0.0299  0.0411  0.0483  0.0509  0.0483  0.0411  0.0299  0.0157
0.40  0.0058  0.0109  0.0151  0.0177  0.0186  0.0177  0.0151  0.0109  0.0058
0.50  0.0021  0.0040  0.0055  0.0065  0.0068  0.0065  0.0055  0.0040  0.0021
0.60  0.0008  0.0015  0.0020  0.0024  0.0025  0.0024  0.0020  0.0015  0.0008
0.70  0.0003  0.0005  0.0007  0.0009  0.0009  0.0009  0.0007  0.0005  0.0003
0.80  0.0001  0.0002  0.0003  0.0003  0.0003  0.0003  0.0003  0.0002  0.0001
0.90  0.0000  0.0001  0.0001  0.0001  0.0001  0.0001  0.0001  0.0001  0.0000
1.00  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
```

on commence a avoir des réponses raisonnables, confirmées avec $\Delta t = 2.5\ 10^{-3}$ (donc, 400 pas de temps pour aller de $t = 0$ à $t = 1$!) :

```
delta t=  0.0025
0.00  0.3600  0.6400  0.8400  0.9600  1.0000  0.9600  0.8400  0.6400  0.3600
0.10  0.1184  0.2252  0.3099  0.3643  0.3831  0.3643  0.3099  0.2252  0.1184
0.20  0.0439  0.0836  0.1150  0.1352  0.1422  0.1352  0.1150  0.0836  0.0439
0.30  0.0163  0.0310  0.0427  0.0502  0.0528  0.0502  0.0427  0.0310  0.0163
0.40  0.0061  0.0115  0.0158  0.0186  0.0196  0.0186  0.0158  0.0115  0.0061
0.50  0.0022  0.0043  0.0059  0.0069  0.0073  0.0069  0.0059  0.0043  0.0022
0.60  0.0008  0.0016  0.0022  0.0026  0.0027  0.0026  0.0022  0.0016  0.0008
0.70  0.0003  0.0006  0.0008  0.0010  0.0010  0.0010  0.0008  0.0006  0.0003
0.80  0.0001  0.0002  0.0003  0.0004  0.0004  0.0004  0.0003  0.0002  0.0001
0.90  0.0000  0.0001  0.0001  0.0001  0.0001  0.0001  0.0001  0.0001  0.0000
1.00  0.0000  0.0000  0.0000  0.0000  0.0001  0.0000  0.0000  0.0000  0.0000
```

Les valeurs correctes, calculées par $u(x,t) = \sum_{m \text{ impair}} \frac{32}{\pi^3 m^3} e^{-\pi^2 m^2 t} \sin(m\pi x)$, étant

```
0.00  0.3600  0.6400  0.8400  0.9600  1.0000  0.9600  0.8400  0.6400  0.3600
0.10  0.1189  0.2261  0.3112  0.3658  0.3846  0.3658  0.3112  0.2261  0.1189
```

```
0.20  0.0443  0.0843  0.1160  0.1363  0.1434  0.1363  0.1160  0.0843  0.0443
0.30  0.0165  0.0314  0.0432  0.0508  0.0534  0.0508  0.0432  0.0314  0.0165
0.40  0.0062  0.0117  0.0161  0.0189  0.0199  0.0189  0.0161  0.0117  0.0062
0.50  0.0023  0.0044  0.0060  0.0071  0.0074  0.0071  0.0060  0.0044  0.0023
0.60  0.0009  0.0016  0.0022  0.0026  0.0028  0.0026  0.0022  0.0016  0.0009
0.70  0.0003  0.0006  0.0008  0.0010  0.0010  0.0010  0.0008  0.0006  0.0003
0.80  0.0001  0.0002  0.0003  0.0004  0.0004  0.0004  0.0003  0.0002  0.0001
0.90  0.0000  0.0001  0.0001  0.0001  0.0001  0.0001  0.0001  0.0001  0.0000
1.00  0.0000  0.0000  0.0000  0.0001  0.0001  0.0001  0.0000  0.0000  0.0000
```

```
 100   for J=1 to 11:T=(J-1)/10
 110   print using(2,2),T;" ";
 120   for I=2 to 10:X=(I-1)/10
 130   U=0:for L=1 to 20 step 2:Pl=#pi*L:Pl32=32/Pl^3
 135     U=U+Pl32*exp(-Pl^2*T)*sin(Pl*X):next L
 140   print using(2,4),U;" ";:next I:print
 150   next J
```

Explication du phénomène : soit $\boldsymbol{u}(t) = [u_h(h,t), u_h(2h,t), \ldots, u_h(\ell - h, t)]^T$ le vecteur des valeurs spatiales de u_h au temps t. L'exploitation numérique de (59) réalise une liaison entre $\boldsymbol{u}(t)$ et $\boldsymbol{u}(t + \Delta t)$:

$$\frac{1}{\Delta t}(\boldsymbol{u}(t + \Delta t) - \boldsymbol{u}(t)) = -\sigma M_h \boldsymbol{u}(t),$$

où $-M_h$ est la matrice familière de discrétisation de u'' : $M_h = h^{-2} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & & -1 & 2 \end{bmatrix}$.

On a donc $\boldsymbol{u}(t + \Delta t) = (I - \sigma \Delta t M_h)\boldsymbol{u}(t)$, $\|\boldsymbol{u}(t + \Delta t)\| \leqslant \|I - \sigma \Delta t M_h\| \, \|\boldsymbol{u}(t)\|$, $\|I - \sigma \Delta t M_h\| = \max_{\lambda(M_h)} |1 - \sigma \Delta t \lambda(M_h)|$.

D'après (54), les valeurs propres de M_h sont positives et bornées par $4/h^2$. La norme de $I - \sigma \Delta t M_h$ sera donc bornée par l'unité si $|1 - 4\sigma \Delta t / h^2| \leqslant 1$: c'est la condition de stabilité numérique pour cette façon de résoudre le problème, assurant que les erreurs sur $\boldsymbol{u}$ ne seront pas amplifiées à chaque pas de temps.

Dans l'exemple illustré ci-dessus, on doit donc avoir effectivement

$$\Delta t \leqslant h^2/(2\sigma),$$

ce qui est bien $5\ 10^{-3}$!

2. Consistance et stabilité pour problèmes d'évolution $\partial u/\partial t + Mu = f$.

Par ***problème d'évolution*** (opposé à *problème stationnaire*), on entend la détermination d'une fonction u de plusieurs variables par étapes successives d'une des variables, toujours notée t : on a donc $u(x,t), x \in \Omega' \subset \mathbb{R}^{n-1}, t \in \mathbb{R}$.

2.1. Problèmes bien posés. Opérateur solution.

Soit $\partial u/\partial t + Mu = 0$, où M est défini sur une partie dense d'un espace de Banach U de fonctions de $n-1$ variables. Le problème de la détermination de $u(x,t)$ pour $t \in [0,T]$ est dit

bien posé si on arrive à montrer l'existence d'un opérateur $E(t)$ tels que $u(x,t) = E(t)u(x,0)$ vérifie

(1) $E(0) = I$,

(2) $\left\| \dfrac{u(x,t+\delta t) - u(x,t)}{\delta t} + Mu(x,t) \right\| \to 0$ quand $\delta t \to 0$ uniformément en $t \in [0,T]$. En particulier, $E(t) \to I$ quand $t \to 0$, $t > 0$.

(3) $\exists C < \infty : \|E(t)\| \leqslant C$, $t \in [0,T]$.

On a vu un exemple explicite d'opérateur $E(t)$ dans (61), p. 152. Dans L^∞, on a $\|E(t)\|$ = intégrale du noyau = 1. Dans L^2, le raisonnement qui suit montrera que l'on a également $\|E(t)\| \leqslant 1$. Par contre, dans L^1, $\|E(t)\| = 1/\sqrt{4\pi\sigma t} \to \infty$ quand $t \to 0$.

La vérification du caractère bien posé de problèmes linéaires à coefficients constants se fait par séries ou transformées de Fourier :

Trouver

$$u(x,t) = \sum_{\boldsymbol{k} \in K} c_{\boldsymbol{k}}(t) e^{i\boldsymbol{k}\cdot x}$$

vérifiant $\partial u/\partial t + Mu = 0$ à partir de

$$u(x,0) = \sum_{\boldsymbol{k} \in K} c_{\boldsymbol{k}}(0) e^{i\boldsymbol{k}\cdot x}$$

mène donc à

$$\frac{dc_{\boldsymbol{k}}}{dt} = -S_M(\boldsymbol{k}) c_{\boldsymbol{k}},$$

pour $\boldsymbol{k} \in K$, $0 \leqslant t \leqslant T$, où S_M est le symbole de l'opérateur M : $Me^{i\boldsymbol{k}\cdot x} = S_M(\boldsymbol{k}) e^{i\boldsymbol{k}\cdot x}$.

L'action de l'opérateur $E(t)$ est donc donnée par

$$E(t) \left(\sum_{\boldsymbol{k} \in K} c_{\boldsymbol{k}}(0) e^{i\boldsymbol{k}\cdot x} \right) = \sum_{\boldsymbol{k} \in K} c_{\boldsymbol{k}}(0) e^{-S_M(\boldsymbol{k})t} e^{i\boldsymbol{k}\cdot x}.$$

Si les exponentielles $e^{i\boldsymbol{k}\cdot x}$, $\boldsymbol{k} \in K$ forment une suite orthogonale complète de $U \subset L^2(\Omega')$, on a

$$\left\| E(t) \left(\sum_{\boldsymbol{k} \in K} c_{\boldsymbol{k}}(0) e^{i\boldsymbol{k}\cdot x} \right) \right\|^2 = \sum_{\boldsymbol{k} \in K} |c_{\boldsymbol{k}}(0)|^2 e^{-2S_M(\boldsymbol{k})t} \|e^{i\boldsymbol{k}\cdot x}\|^2,$$

(Parseval), d'où

$$\|E(t)\| = \max_{\boldsymbol{k} \in K} e^{-S_M(\boldsymbol{k})t}.$$

Le problème est donc certainement bien posé dès que M est semi-défini positif.

C'est le cas de l'équation de la chaleur (60), puisque $M = -\sigma\partial^2/\partial x^2 \Rightarrow S_M(k) = \sigma k^2$.

Vérification directe :

$$\frac{d}{dt}\|u(x,t)\|^2 = 2\int_0^\ell u(x,t) \frac{\partial u(x,t)}{\partial t}\, dx = 2\sigma \int_0^\ell u(x,t) \frac{\partial^2 u(x,t)}{\partial x^2}\, dx =$$
$$- 2\sigma \int_0^\ell \left(\frac{\partial u(x,t)}{\partial x} \right)^2 dx = -2a(u,u) \leqslant 0.$$

Vérifier le caractère bien posé d'un problème, c'est de l'analyse (semi-groupe : $E(t_1)E(t_2) = E(t_1 + t_2)$). La discrétisation de ce problème concerne l'analyse numérique.

2.2. Consistance et stabilité de discrétisations de problèmes d'évolution.

On discrétise $Lu = \partial u/\partial t + Mu = 0$ en rassemblant des contributions en $t + \Delta t$ et en t sous la forme

$$K_{1,h} u_h(., t + \Delta t) + K_{0,h} u_h(., t) = 0, \tag{68}$$

ou, en isolant tout ce qui dépend de $t + \Delta t$,

$$u_h(., t + \Delta t) = E_h u_h(., t).$$

La discrétisation est ***consistante*** si

$$\|(r_h E(\Delta t) - E_h r_h)u\| \leqslant \epsilon_h \Delta t,$$

pour toute fonction u suffisamment lisse dans une partie dense de U, et où $\epsilon_h \to 0$ quand $h \to 0$ (bien entendu, Δt est lié à h et $\to 0$ quand $h \to 0$).

(Pourquoi le produit $\epsilon_h \Delta t$? Quand $h \to 0$, $E(\Delta t)$ et E_h tendent vers l'identité. Une erreur en Δt ne représente aucune information pertinente).

La différence $(r_h E(\Delta t) - E_h r_h)u$ porte aussi le nom d'***erreur locale*** : c'est l'erreur consentie par le solveur numérique sur un seul pas de temps.

Ainsi, si M_h est une approximation consistante de M, $E_h = I - \Delta t M_h$ convient certainement, mais présente des inconvénients, comme on a vu.

Le schéma (68) est ***numériquement stable*** si $u_h(., t) = \boldsymbol{u}(t)$ reste borné par une même constante C (indépendante de h) pour $\forall t \in [0, T]$, donc si

$$\left\|E_h^N\right\| \leqslant C,$$

$N = 1, 2, \ldots, T/\Delta t$.

On a vu que (59) n'est numériquement stable que si $\Delta t \leqslant h^2/(2\sigma)$.

La recherche de schémas numériquement stables est une des principales difficultés en ces matières.

3. Théorème d'équivalence de Lax.

Il s'agit de relier

(1) ***Consistance :*** $\|E_h r_h u(., t) - r_h E(\Delta t) u(., t)\| \leqslant \epsilon_h \Delta t$,

(2) ***Stabilité numérique :*** $\|E_h^N\| \leqslant C,\ N \leqslant T/\Delta t$,

(3) et ***Convergence :*** $\|u_h(., t) - r_h u(., t)\| \to 0$ si $h \to 0$, où $u_h(., t) = E_h^{t/\Delta t} r_h u(., 0)$ et $u(., t) = E(t)u(., 0) = E(\Delta t)^{t/\Delta t} u(., 0)$.

Pour un problème bien posé, discrétisé de manière consistante, la stabilité numérique est une condition nécessaire et suffisante de convergence, $\forall u(x, 0) \in U$.

En effet,

(1) *Convergence ⇒ stabilité numérique.*
En effet, supposons $\{\|E_h^N\|\}$ non bornée, il existe alors $u_0 \in U$ telle que une suite extraite $\|(E_{h_i})^{N_i} u_0\| \to \infty$. Soit Δ_i le pas de temps correspondant à h_i. On peut encore extraire une suite $\{h_i, N_i\}$ telle que $N_i \Delta_i \to \tau \in [0, T]$.

Cependant, comme le schéma est convergent, les normes $\|(E_{h_i})^{N_i} u_0\|$ doivent tendre vers la norme de $u(x, \tau)$, la solution en $t = \tau$ correspondant à la condition initiale u_0.

La convergence impose donc $\|E_h^N u_0\| \leqslant$ une fonction de u_0, donc une constante, par le principe de la borne uniforme[3].

(2) *Stabilité numérique ⇒ convergence.*
Soit u une solution de $\partial u/\partial t + Mu = 0$, et $u_0(x) = u(x,0)$.

Par consistance, $\|(E_h r_h - r_h E(\Delta t))u(x,t)\| \leqslant \varepsilon_h \Delta t(h)$, avec $\varepsilon_h \to 0$ quand $h \to 0$.

Il faut montrer $\|[(E_{h_i})^{N_i} r_h u_0 - r_h E(\tau)]u_0\| \to 0$ si $N_i\Delta_i \to \tau \leqslant T$.

On a

$$(E_{h_i})^{N_i} r_h - r_h (E(\tau/N_i))^{N_i} = \sum_{p=0}^{N_i-1} (E_{h_i})^p [E_{h_i} r_h - r_h E(\tau/N_i)](E(\tau/N_i))^{N_i-1-p},$$

donc,

$$\begin{aligned}
&\|[(E_{h_i})^{N_i} r_h - r_h (E(\tau/N_i))^{N_i}]u_0\| \\
&\quad = \| \sum_{p=0}^{N_i-1} (E_{h_i})^p [E_{h_i} r_h - r_h E(\tau/N_i)](E(\tau/N_i))^{N_i-1-p} u_0\| \\
&\quad = \| \sum_{p=0}^{N_i-1} (E_{h_i})^p [E_{h_i} r_h - r_h E(\tau/N_i)] u(\ ,\tau_p)\|, \text{ avec } \tau_p = (N_i - 1 - p)\tau/N_i, \\
&\quad \leqslant \sum_{p=0}^{N_i-1} \|(E_{h_i})^p\| \varepsilon_h \Delta_i \quad \text{(consistance} \\
&\quad \leqslant C\varepsilon_h N_i \Delta_i \quad \text{stabilité num.} \\
&\quad \leqslant C\varepsilon_h \tau.
\end{aligned}$$

4. Classe des équations paraboliques.

Soit $\Omega = \Omega' \times [0,T]$, où Ω' est un domaine (**spatial**) de $\mathbb{R}^{n-1}$, U' un espace de fonctions $\Omega' \mapsto \mathbb{R}$ muni d'un produit scalaire $(\ ,\ )_{U'}$.

On appelle ici **parabolique** une équation

$$\frac{\partial u}{\partial t} = -Mu,$$

où M est un opérateur symétrique (= formellement autoadjoint) sur (une partie de) U', défini positif, c.-à-d., $\forall u, v \in \mathrm{dom}(M)$,

$$(Mu, v)_{U'} = (u, Mv)_{U'} = a(u,v),$$

avec a symétrique définie positive : $a(u,u) > 0$.

Exemples

(1) $Mu = -u''$ (**équation de la chaleur**) sur $\mathscr{C}^2[a,b] \bigcap \{u : u(a) = u(b) = 0\}$, avec le produit scalaire de L^2 :
$$-\int_a^b u''(x)v(x)\,dx = \int_a^b u'(x)v'(x)\,dx.$$

[3]Si une suite d'opérateurs continus d'un Banach dans un normé est bornée en chaque point, les normes de ces opérateurs sont bornées dans leur ensemble (cf. point 2a, p. 45).

(2) laplacien : $Mu = -\Delta u$ sur Ω', (**équation de la chaleur**) avec $u(x) = 0$ sur la frontière Γ' de Ω', toujours avec le produit scalaire de L^2 :
$-\int_{\Omega'} \Delta u(x)v(x)\,dx = \int_{\Omega'} \boldsymbol{grad}u(x) \cdot \boldsymbol{grad}v(x)\,dx.$

(3) **Caractère parabolique de l'équation de Black-Scholes** $\partial u/\partial t = \alpha x^2 \partial^2 u/\partial x^2 + rx\partial u/\partial x - ru$.
On a donc $Mu = -\alpha x^2 u'' - rxu' + ru$, α et $r > 0$. Essayons un produit scalaire pondéré $(u, v) = \int_0^\infty u(x)v(x)w(x)\,dx$. On a
$(Mu, v) = \alpha \left[u'(x)x^2 w(x)v(x)\right]_0^\infty + \int_0^\infty \left\{\alpha u'(x)(x^2 w(x)v(x))' - rxu'(x)w(x)v(x) + ru(x)v(x)w(x)\right\} dx$
qui est effectivement symétrique défini positif si $w(x) = x^{r/\alpha - 2}$ (et si les fonctions de U' tendent assez vite vers 0 quand $x \to \infty$).

cf. Advanced Finite Difference Methods for Financial Instrument

`http ://www.datasim.nl/education/coursedetails.asp?coursecategory=FE&coursecode=AFDM`

`http ://www.math.univ-montp2.fr/~mohamadi/tipengweb/black_fr_presentation.htm`

`http ://www.math.univ-montp2.fr/ mohamadi/tipengweb/black_fr_modelisation.htm`

4.1. Examen de quelques schémas.

Dans chaque cas, on examinera

(1) Le comportement de
$$\|(E_h r_h - r_h E(\Delta t))u(\ ,t)\| \leqslant \varepsilon_h \Delta t(h)$$
quand u est solution de $\partial u/\partial t + Mu = 0$ (consistance),

(2) L'existence d'une borne $\|(E_h)^N\| \leqslant C$ pour $N\Delta t \leqslant T$ (stabilité numérique),

pour constater et quantifier $\|u_h(\ ,t) - r_h u(\ ,t)\| \leqslant C\varepsilon_h t$.

4.2. Schémas à deux niveaux de temps.

(1) Euler explicite. Reprenons encore d'abord (59) :
$E_h r_h u(x,t) = u(x,t) + \sigma\Delta t(u(x-h,t) - 2u(x,t) + u(x+h,t))/h^2$.
Si u est quatre fois continûment dérivable en x,
$$u(x \pm h, t) = u(x,t) \pm h\frac{\partial u(x,t)}{\partial x} + \frac{h^2}{2}\frac{\partial^2 u(x,t)}{\partial x^2} \pm \frac{h^3}{6}\frac{\partial^3 u(x,t)}{\partial x^3} + \frac{h^4}{24}\frac{\partial^4 u(x,t)}{\partial x^4} + \cdots$$
Donc, $E_h r_h u(x,t) = u(x,t) + \sigma\Delta t\left(\frac{\partial^2 u(x,t)}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 u(x,t)}{\partial x^4} + O(h^4)\right)$,
et $E(\Delta t)u(x,t) = u(x,t+\Delta t) = u(x,t) + \Delta t\frac{\partial u(x,t)}{\partial t} + \frac{(\Delta t)^2}{2}\frac{\partial^2 u(x,t)}{\partial t^2} + O((\Delta t)^3)$,

$$E_h r_h u(x,t) - r_h E(\Delta t)u(x,t) = \sigma\Delta t\left(\frac{\partial^2 u(x,t)}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 u(x,t)}{\partial x^4} + O(h^4)\right)$$
$$- \Delta t\frac{\partial u(x,t)}{\partial t} - \frac{(\Delta t)^2}{2}\frac{\partial^2 u(x,t)}{\partial t^2} + O((\Delta t)^3)$$
$$= \sigma\Delta t\frac{h^2}{12}\frac{\partial^4 u(x,t)}{\partial x^4} + O(h^4\Delta t) - \frac{(\Delta t)^2}{2}\frac{\partial^2 u(x,t)}{\partial t^2} + O((\Delta t)^3)$$
$$= \sigma\Delta t\left(\frac{h^2}{12} - \sigma\frac{\Delta t}{2}\right)\frac{\partial^4 u(x,t)}{\partial x^4} + O((h^4 + \Delta t^2)\Delta t)$$

où a tenu compte de $\partial u/\partial t = \sigma\partial^2 u/\partial x^2$ d'abord, et de sa conséquence $\partial^2 u/\partial t^2 = \sigma^2\partial^4 u/\partial x^4$ ensuite.

On a donc $\varepsilon_h = O(h^2+\Delta t)$, qui tombe à $\varepsilon_h = O(h^4+\Delta t^2) = O(h^4)$, si $\Delta t = \dfrac{h^2}{6\sigma}$ (si $u \in \mathscr{C}^6$).

Quant à la stabilité numérique, on a vu (page 157) que $C = 1$ dès que $\Delta t \leqslant \dfrac{h^2}{2\sigma}$.

(2) **Schéma de Crank et Nicolson**[4].

On évite la nécessité de recourir à de très petits pas de temps $\Delta t = O(h^2)$ par le schéma ***implicite*** très élégant

$$\frac{u_h(x,t+\Delta t) - u_h(x,t)}{\Delta t} = \frac{\sigma}{2}\left\{\frac{u_h(x-h,t+\Delta t) - 2u_h(x,t+\Delta t) + u_h(x+h,t+\Delta t)}{h^2}\right.$$
$$\left. + \frac{u_h(x-h,t) - 2u_h(x,t) + u_h(x+h,t)}{h^2}\right\} \tag{69}$$

On n'obtient donc pas directement une valeur numérique de $u_h(x,t+\Delta t)$ à partir de valeurs de $u_h(\ ,t)$, mais un système d'équations pour $\boldsymbol{u}(t+\Delta t) = \begin{bmatrix} u_h(x_1,t+\Delta t) \\ \vdots \\ u_h(x_N,t+\Delta t)\end{bmatrix}$ à partir de $\boldsymbol{u}(t)$. Pour l'équation parabolique $\partial u/\partial t = -Mu$:

$\dfrac{\boldsymbol{u}(t+\Delta t) - \boldsymbol{u}(t)}{\Delta t} = -\dfrac{\sigma}{2}(M_h\boldsymbol{u}(t+\Delta t) + M_h\boldsymbol{u}(t))$, soit

$$\boldsymbol{u}(t+\Delta t) = E_h\boldsymbol{u}(t) = \left(I + \Delta t\frac{\sigma}{2}M_h\right)^{-1}\left(I - \Delta t\frac{\sigma}{2}M_h\right)\boldsymbol{u}(t).$$

L'inconvénient du caractère implicite de ce schéma est compensé par d'excellentes propriétés de stabilité et de précision.

Voyons d'abord la ***stabilité numérique*** : dès que M_h est symétrique de spectre positif, on a $\|E_h\| = \max|\lambda(E_h)| = \max\left|\dfrac{1-\Delta t\sigma\lambda(M_h)/2}{1+\Delta t\sigma\lambda(M_h)/2}\right| \leqslant 1$, donc $\|(E_h)^N\| \leqslant 1$, $\forall N$.

[4]Crank, J., Nicolson, P., "A Practical Method for Numerical Evaluation of Solutions of Partial Differential Equations of the Heat-Conduction Type", *Proceedings of the Cambridge Philosophical Society*, **43**, (1947), pp. 50-67.

Au passage, notons aussi le ***nombre de condition*** $\kappa(I + \Delta t\sigma M_h/2) = \lambda_{\max}/\lambda_{\min} \leqslant 1 + \Delta t\sigma\lambda_{\max}(M_h)/2 \leqslant 1+2\Delta t\sigma/h^2$, puisque $\lambda_{\max}(M_h) \leqslant 4/h^2$ (cf. (54), p. 126). Comme on n'a plus de contrainte de stabilité sur Δt, on s'attend à $\Delta t \approx h$, donc à $\kappa = O(h^{-1})$ au lieu de $O(h^{-2})$ pour des problèmes stationnaires du second ordre. Ceci peut justifier le recours à des schémas évolutifs de résolution de problèmes stationnaires...

Consistance et précision :

Soit $L = \dfrac{\partial}{\partial t} + \sigma M$.

Examinons $E_h r_h u - r_h E(\Delta t)u$.

$E(\Delta t)u$ est la fonction $v(x;\Delta t)$, solution en $t = \Delta t$ de

$$\frac{\partial v(x;t)}{\partial t} = -\sigma(Mv)(x;t)$$

avec la condition initiale $v(x;0) \equiv u(x)$.

Taylor (en t) :

$$\begin{aligned} E(\Delta t) &= I + \Delta t\frac{\partial}{\partial t} + \frac{1}{2}\Delta t^2\frac{\partial^2}{\partial t^2} + \frac{1}{6}\Delta t^3\frac{\partial^3}{\partial t^3} + \cdots \\ &= I - \Delta t\sigma M + \frac{1}{2}\Delta t^2\sigma^2 M^2 - \frac{1}{6}\Delta t^3\sigma^3 M^3 + \cdots \end{aligned}$$

D'autre part, $r_h u$ est le vecteur de composantes $u(x)$, pour $x =$ des multiples de h,

$E_h = \dfrac{2I_h - \sigma\Delta t M_h}{2I_h + \sigma\Delta t M_h}$où M_h est la matrice de discrétisation de M, par exemple, si $M = -\partial^2/\partial x^2$: $M_h = h^{-2}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & & -1 & 2 \end{bmatrix}$.

Taylor encore :

$$E_h = I_h - \Delta t\sigma M_h + \frac{1}{2}\Delta t^2\sigma^2 M_h^2 - \frac{1}{4}\Delta t^3\sigma^3 M_h^3 + \cdots$$

les mêmes coefficients jusqu'au terme en Δt^2 y compris !

$$\frac{E_h r_h - r_h E(\Delta t)}{\Delta t} = \sigma(r_h M - M_h r_h) + \frac{1}{2}\Delta t\sigma^2(M_h^2 r_h - r_h M^2) + \Delta t^2\sigma^3\left(r_h\frac{M^3}{6} - \frac{M_h^3}{4}r_h\right) + \cdots$$

borné par const. $(h^2 + \Delta t^2)$ si $\|(r_h M - M_h r_h)u\| = O(h^2)$.

(3) Généralisation des précédents :

$$\begin{aligned} \frac{u_h(x,t+\Delta t) - u_h(x,t)}{\Delta t} = \sigma\Bigg\{&\theta\frac{u_h(x-h,t+\Delta t) - 2u_h(x,t+\Delta t) + u_h(x+h,t+\Delta t)}{h^2} \\ &+(1-\theta)\frac{u_h(x-h,t) - 2u_h(x,t) + u_h(x+h,t)}{h^2}\Bigg\} \end{aligned}$$

avec $0 \leqslant \theta \leqslant 1$. $\theta = 0$: (59) ; $\theta = 1$: schéma implicite pur ; $\theta = 1/2$: (69).

On a maintenant

$$E_h = (I + \Delta t\theta\sigma M_h)^{-1}\,(I - \Delta t(1-\theta)\sigma M_h)\,.$$

Spectre de $E_h \in ((\theta-1)/\theta, 1)$: stable si $\theta \geqslant 1/2$.
Si $\theta < 1/2$, on a encore la stabilité numérique si $1 - \Delta t(1-\theta)\sigma\lambda_{\max} \geqslant -1 - \Delta t\theta\sigma\lambda_{\max} \Rightarrow \Delta t \leqslant \dfrac{h^2}{2(1-2\theta)\sigma}$ avec $\lambda_{\max} \approx 4/h^2$.

Consistance et précision :
On a maintenant

$$y = u + \sigma\Delta t\left(\theta\left[y'' + \frac{h^2}{12}y^{\mathrm{iv}}\right] + (1-\theta)\left[+u'' + \frac{h^2}{12}u^{\mathrm{iv}}\right] + O(h^4)\right),$$

$$\begin{aligned} y &= u + O(\Delta t) \\ &= u + \sigma\Delta t\left(u'' + O(h^2)\right) + O(\Delta t^2) \\ &= u + \sigma\Delta t\left(u'' + \theta\sigma\Delta t u^{\mathrm{iv}} + \frac{h^2}{12}u^{\mathrm{iv}} + O(h^4)\right) + O(\Delta t^3). \end{aligned}$$

On a donc

$$E_h u = u + \sigma\Delta t\frac{\partial^2 u}{\partial x^2} + \sigma\Delta t\left(\theta\sigma\Delta t + \frac{h^2}{12}\right)\frac{\partial^4 u}{\partial x^4} + O(h^4\Delta t) + O(\Delta t^3),$$

$$E(\Delta t)u = u + \Delta t\frac{\partial u}{\partial t} + \frac{\Delta t^2}{2}\frac{\partial^2 u}{\partial t^2} + O(\Delta t^3),$$

$$\begin{aligned} E_h r_h u - r_h E(\Delta t)u &= +\sigma\Delta t\left(\theta\sigma\Delta t + \frac{h^2}{12}\right)\frac{\partial^4 u}{\partial x^4} - \frac{\Delta t^2}{2}\frac{\partial^2 u}{\partial t^2} + O(h^4\Delta t) + O(\Delta t^3) \\ &= \sigma\Delta t\left((\theta - 1/2)\sigma\Delta t + \frac{h^2}{12}\right)\frac{\partial^4 u}{\partial x^4} + O(h^4\Delta t) + O(\Delta t^3), \end{aligned}$$

toujours grâce à $\partial^2 u/\partial t^2 = \sigma^2\partial^4 u/\partial x^4$.

Donc $\varepsilon_h = O(h^2) + O(\Delta t^2)\,;\, O(h^4)$ si $\theta = \dfrac{1}{2} - \dfrac{h^2}{12\sigma\Delta t}$, soit : $\Delta t = \dfrac{h^2}{6(1-2\theta)\sigma}$.

4.3. Schémas à plus de deux niveaux de temps.

On a maintenant

$$K_{p,h}u_h(.,t+p\Delta t) + K_{p-1,h}u_h(.,t+(p-1)\Delta t) + \cdots + K_{0,h}u_h(.,t) = 0 \qquad (70)$$

au lieu de (68). On espère améliorer la précision des résultats en gardant une certaine simplicité des calculs, puisque $u_h(\ ,t+(p-1)\Delta t)$,…, $u_h(\ ,t)$ sont disponibles au moment d'aborder le calcul de $u_h(\ ,t+p\Delta t)$. Il faut cependant amorcer la résolution par un calcul séparé de $u_h(\ ,\Delta t)$, $u_h(\ ,2\Delta t)$,…, $u_h(\ ,(p-1)\Delta t)$.

Les discussions de stabilité et de consistance se font maintenant à l'aide d'un opérateur E_h reprenant p niveaux de temps :

$$E_h\begin{bmatrix} u_h(\ ,t) \\ u_h(\ ,t+\Delta t) \\ \vdots \\ u_h(t+(p-1)\Delta t) \end{bmatrix} = \begin{bmatrix} u_h(\ ,t+\Delta t) \\ u_h(\ ,t+2\Delta t) \\ \vdots \\ u_h(t+p\Delta t) \end{bmatrix},$$

c'est-à-dire une matrice d'ordre Np :

$$E_h = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & & \cdots & 1 \\ -K_p^{-1}K_0 & -K_p^{-1}K_1 & \cdots & & -K_p^{-1}K_{p-1} \end{bmatrix}$$

On a toujours $\left\| \begin{bmatrix} u_h(\ ,t) \\ u_h(\ ,t+\Delta t) \\ \vdots \\ u_h(t+(p-1)\Delta t) \end{bmatrix} - r_h \begin{bmatrix} u(\ ,t) \\ u(\ ,t+\Delta t) \\ \vdots \\ u(t+(p-1)\Delta t) \end{bmatrix} \right\| \leqslant C\varepsilon_h t$ si

$\left\| E_h r_h \begin{bmatrix} u(\ ,t) \\ u(\ ,t+\Delta t) \\ \vdots \\ u(t+(p-1)\Delta t) \end{bmatrix} - r_h \begin{bmatrix} u(\ ,t+\Delta t) \\ u(\ ,t+2\Delta t) \\ \vdots \\ u(t+p\Delta t) \end{bmatrix} \right\| \leqslant \varepsilon_h \Delta t$ (***consistance***) quand u est solution de $\partial u/\partial t + Mu = 0$, et $\|(E_h)^n\| \leqslant C$ pour $n\Delta t \leqslant T$ (***stabilité numérique***).

Schéma de DuFort et Frankel.

Ces auteurs ont proposé

$$\frac{u_h(x,t+\Delta t) - u_h(x,t-\Delta t)}{2\Delta t} = \sigma \frac{u_h(x+h,t) - u_h(x,t+\Delta t) - u_h(x,t-\Delta t) + u_h(x-h,t)}{h^2} \tag{71}$$

pour résoudre (60).

On a donc $K_2 u_h(x,\ ) = \left(\dfrac{1}{2\Delta t} + \dfrac{\sigma}{h^2}\right) u_h(x,\ )$, $K_0 u_h(x,\ ) = \left(\dfrac{-1}{2\Delta t} + \dfrac{\sigma}{h^2}\right) u_h(x,\ )$ (matrices diagonales), et $K_1 u_h(x,\ ) = -\sigma \dfrac{u_h(x+h,\ ) + u_h(x-h,\ )}{h^2}$.

Donc, $-(K_2)^{-1}K_0 = -\dfrac{\dfrac{\sigma}{h^2} - \dfrac{1}{2\Delta t}}{\dfrac{\sigma}{h^2} + \dfrac{1}{2\Delta t}} = -\dfrac{1-\alpha}{1+\alpha}$, où $\alpha = h^2/(2\sigma\Delta t)$;

$-(K_2)^{-1}K_1 = -\left(\dfrac{\sigma}{h^2} - \dfrac{1}{2\Delta t}\right)^{-1} \sigma N_h = \dfrac{2I - h^2 M_h}{1+\alpha}$, où $N_h = h^{-2} \begin{bmatrix} 0 & -1 & & \\ -1 & 0 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 0 \end{bmatrix}$

$= M_h - 2I/h^2$, où $M_h = h^{-2} \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 \end{bmatrix}$ est la matrice de la discrétisation familière de $-\partial^2/\partial x^2$, donc

$$E_h = \begin{bmatrix} 0 & I \\ -\dfrac{1-\alpha}{1+\alpha} & \dfrac{2I - h^2 M_h}{1+\alpha} \end{bmatrix}$$

On a vu que le schéma explicite (59) n'est numériquement stable que si $\Delta t \leqslant h^2/(2\sigma)$, ce qui correspond à $\alpha > 1$. Tout en restant explicite, le schéma de DuFort & Frankel permet de diminuer α, tout en gardant cependant $\alpha \geqslant$ constante > 0.

Consistance : $E_h \begin{bmatrix} r_h u(\ ,t-\Delta t) \\ r_h u(\ ,t) \end{bmatrix}$ doit être proche, à moins de $\varepsilon_h \Delta t$, de $r_h \begin{bmatrix} u(\ ,t) \\ u(\ ,t+\Delta t) \end{bmatrix}$ quand $\partial u/\partial t = -\sigma M u$:

$$E_h r_h u = \begin{bmatrix} 0 & I \\ -\dfrac{1-\alpha}{1+\alpha} & \dfrac{2I-h^2M_h}{1+\alpha} \end{bmatrix} \begin{bmatrix} u(\ ,t-\Delta t) \\ u(\ ,t) \end{bmatrix} \quad , \quad E(\Delta t) \begin{bmatrix} u(\ ,t-\Delta t) \\ u(\ ,t) \end{bmatrix} = \begin{bmatrix} u(\ ,t) \\ u(\ ,t+\Delta t) \end{bmatrix}$$

Taylor deuxième composante :

$$\begin{aligned} & u(\ ,t) - \frac{h^2 M_h r_h u}{1+\alpha} + \frac{1-\alpha}{1+\alpha}\Delta t \frac{\partial u}{\partial t} - \frac{1-\alpha}{1+\alpha}\Delta t^2 \frac{\partial u^2}{2\partial t^2} + \cdots \\ &= u(\ ,t+\Delta t) - \frac{h^2 M_h r_h u}{1+\alpha} - \frac{2\alpha}{1+\alpha}\Delta t \frac{\partial u}{\partial t} - \frac{2}{1+\alpha}\Delta t^2 \frac{\partial u^2}{2\partial t^2} + \cdots \\ &= u(\ ,t+\Delta t) - \frac{h^2 M_h r_h u}{1+\alpha} + \sigma \frac{2\alpha}{1+\alpha}\Delta t r_h M u - \frac{1}{1+\alpha}\Delta t^2 \sigma^2 r_h M^2 u + \cdots \\ &= u(\ ,t+\Delta t) + \sigma \frac{2\alpha}{1+\alpha}\Delta t (r_h M u - M_h r_h u) - \frac{1}{1+\alpha}\Delta t^2 \sigma^2 r_h M^2 u + \cdots \end{aligned}$$

ce qui tend bien à être infiniment plus petit que Δt si $(r_h M - M_h r_h)u \to 0$ quand $h \to 0$ (on a utilisé $2\alpha\sigma\Delta t = h^2$).

Stabilité numérique :

On peut apprécier $\|\boldsymbol{A}^m\|$ pour une matrice non nécessairement symétrique à partir d'une similitude

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{J}\boldsymbol{V}^{-1} \Rightarrow \boldsymbol{A}^{\boldsymbol{m}} = \boldsymbol{V}\boldsymbol{J}^m\boldsymbol{V}^{-1},$$

provenant, par exemple, d'une forme de Jordan ou d'une forme de Schur.

Alors, $\|\boldsymbol{A}^m\| \leqslant \|\boldsymbol{V}\|\ \|\boldsymbol{J}^m\|\ \|\boldsymbol{V}^{-1}\|$.

Dans le cas le plus simple, si $\boldsymbol{A}$ admet une base de vecteurs propres, $\boldsymbol{J}$ est la matrice diagonale des valeurs propres, et $\boldsymbol{V}$ est la matrice dont les colonnes sont les vecteurs propres : $\boldsymbol{A}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{J}$. De plus, si $\boldsymbol{A}$ est symétrique, $\boldsymbol{V}$ est orthogonale, les normes de $\boldsymbol{V}$ et $\boldsymbol{V}^{-1}$ valent 1 et on retrouve $\|\boldsymbol{A}^m\| = (\max|\lambda(\boldsymbol{A})|)^m$.

Examinons les valeurs et vecteurs propres de E_h :

$$\begin{bmatrix} 0 & I \\ -\dfrac{1-\alpha}{1+\alpha} & \dfrac{2I-h^2M_h}{1+\alpha} \end{bmatrix} \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u} \\ -\dfrac{1-\alpha}{1+\alpha}\boldsymbol{v} + \dfrac{2\boldsymbol{u}-h^2M_h\boldsymbol{u}}{1+\alpha} \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{u} \end{bmatrix}.$$

Il faut donc $\boldsymbol{v} = \lambda^{-1}\boldsymbol{u}$, que $\boldsymbol{u}$ soit vecteur propre de M_h, et

$$\lambda = -\frac{1-\alpha}{1+\alpha}\lambda^{-1} + \frac{2-h^2\lambda(M_h)}{1+\alpha},$$

soit une équation du second degré pour chacune des $2N$ valeurs propres de E_h :

$$(1+\alpha)\lambda^2 - [2-h^2\lambda(M_h)]\lambda + 1 - \alpha = 0.$$

Comme $\mu := h^2\lambda(M_h) \in (0,4)$, les valeurs propres *réelles* sont entre -1 et 1 (le trinôme ci-dessus est positif en $\lambda = \pm 1$ et la demi-somme des racines est $(1-\mu/2)/(1+\alpha) \in (-1,1)$). Quant aux valeurs propres *complexes*, elles ont toutes la même valeur absolue (racine carrée du produit) $[(1-\alpha)/(1+\alpha)]^{1/2} < 1$.

Exercice. Pourquoi n'avoir pas pris

$$\frac{u_h(x,t+\Delta t) - u_h(x,t-\Delta t)}{2\Delta t} = \sigma \frac{u_h(x+h,t) - 2u_h(x,t) + u_h(x-h,t)}{h^2}$$

au lieu de (71) ? Précisément pour une question de stabilité numérique : on obtient le trinôme $\alpha\lambda^2 + \mu\lambda - \alpha$ qui vaut $-\mu < 0$ en $\lambda = -1 \Rightarrow$ une racine < -1.
Revenons à (71), E_h ***n'étant plus symétrique***, $|\lambda(E_h)| \leqslant 1$ ne suffit pas à assurer $\|(E_h)^N\| \leqslant C$:

Exercice. Montrez que, si $h = 0$, $\alpha = 0$, $(E_h)^N = \begin{bmatrix} 0 & 1 \\ -1 & 2 \end{bmatrix}^N = \begin{bmatrix} 1-N & N \\ -N & N+1 \end{bmatrix}$.

Soit $\boldsymbol{U}_h$ la matrice orthogonale des vecteurs propres de M_h. Comme les vecteurs propres de E_h sont de la forme $\begin{bmatrix} \lambda^{-1}\boldsymbol{u} \\ \boldsymbol{u} \end{bmatrix}$, où $\boldsymbol{u}$ est vecteur propre de M_h, la matrice des vecteurs propres de E_h est $\begin{bmatrix} \boldsymbol{U}_h\Lambda_1^{-1} & \boldsymbol{U}_h\Lambda_2^{-1} \\ \boldsymbol{U_h} & \boldsymbol{U_h} \end{bmatrix}$, ou Λ_1 et Λ_2 sont des matrices diagonales de valeurs propres de E_h. Cette matrice ne représente plus une base de $\mathbb{R}^{2N}$ si Λ_1 et Λ_2 ont des éléments communs (valeurs propres multiples). De toute façon, on ne peut contrôler la borne de l'inverse si des valeurs propres sont presques confondues.

On s'en sort en effectuant une similitude avec une matrice orthogonale mettant bien en évidence ce qui se passe. Ici, considérons $\mathcal{U}_h = \begin{bmatrix} \boldsymbol{U_h} & 0 \\ 0 & \boldsymbol{U_h} \end{bmatrix}$, d'inverse $\begin{bmatrix} \boldsymbol{U}'_h & 0 \\ 0 & \boldsymbol{U}'_h \end{bmatrix}$.

On a

$$\mathcal{U}_h^{-1} E_h \mathcal{U}_h = \begin{bmatrix} \boldsymbol{U}'_h & 0 \\ 0 & \boldsymbol{U}'_h \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -\dfrac{1-\alpha}{1+\alpha} & \dfrac{2-h^2M_h}{1+\alpha} \end{bmatrix} \begin{bmatrix} \boldsymbol{U}_h & 0 \\ 0 & \boldsymbol{U}_h \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\dfrac{1-\alpha}{1+\alpha} & \dfrac{2-h^2\Lambda(M_h)}{1+\alpha} \end{bmatrix},$$

où on a utilisé $\boldsymbol{U}'_h\boldsymbol{U}_h = I$ et $\boldsymbol{U}'_h M_h \boldsymbol{U}_h = \Lambda(M_h)$.

Soit μ_i la $i^{\text{ème}}$ valeur propre de h^2M_h. Il suffit de considérer la $N^{\text{ème}}$ puissance de la matrice d'ordre 2 formée des i et $i+N^{\text{èmes}}$ ligne et colonne de $\begin{bmatrix} 0 & 1 \\ -\dfrac{1-\alpha}{1+\alpha} & \dfrac{2-h^2\Lambda(M_h)}{1+\alpha} \end{bmatrix}$, soit

$$\begin{bmatrix} 0 & 1 \\ -\dfrac{1-\alpha}{1+\alpha} & \dfrac{2-\mu_i}{1+\alpha} \end{bmatrix}^N = \begin{bmatrix} 0 & 1 \\ -\lambda_1\lambda_2 & \lambda_1+\lambda_2 \end{bmatrix}^N = \begin{bmatrix} \dfrac{\lambda_2^{N-1} - \lambda_1^{N-1}}{\lambda_2^{-1} - \lambda_1^{-1}} & \dfrac{\lambda_2^N - \lambda_1^N}{\lambda_2 - \lambda_1} \\ \dfrac{\lambda_2^N - \lambda_1^N}{\lambda_2^{-1} - \lambda_1^{-1}} & \dfrac{\lambda_2^{N+1} - \lambda_1^{N+1}}{\lambda_2 - \lambda_1} \end{bmatrix}$$

On montre que cette matrice reste bornée pourvu que λ_1 et λ_2 ne puissent s'approcher *simultanément* du cercle unité, ce qui est le cas si α reste $\geqslant$ une constante > 0, puisque le produit des racines vaut $(1-\alpha)/(1+\alpha)$. On a

$$\left| \frac{\lambda_2^N - \lambda_1^N}{\lambda_2 - \lambda_1} \right| \leqslant 1 + |\lambda_1\lambda_2| + \cdots + |\lambda_1\lambda_2|^{N-1} \leqslant \frac{1+\alpha}{\min(\alpha, 1)}.$$

Autre démonstration, par théorème de Kreiss : voir plus loin.
On se risque parfois à prendre $\alpha \to$ lentement vers 0 quand $h \to 0$ [Gustafsson].

5. Equations hyperboliques.

Les solutions d'équations hyperboliques sont caractérisées par une vitesse de propagation finie : $u(x,t)$ ne dépend que des $u(y,0)$ avec $|y-x| \leqslant ct$.

> Je remarquai que d'une galaxie distante de cent millions d'années-lumière se détachait un carton. Dessus, il était écrit : JE T'AI VU. Je fis rapidement le calcul : la lumière de la galaxie avait mis cent millions d'années pour me joindre, et comme de là-bas ils voyaient ce qui se passait ici avec cent millions d'années de retard, le moment où ils m'avaient vu devait remonter à deux cents millions d'années.
>
> I. Calvino (*Cosmicomics*)

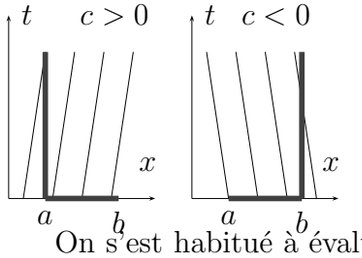
> Nothing travels faster than the speed of light with the possible exceptions of bad news,...
> The people of ...did try to build spaceships that were powered by bad news but etc.
>
> D. Adams (*Mostly harmless*)

5.1. Caractéristiques, domaine d'influence.

On a vu que les équations hyperboliques les plus simples tirent leur nom du genre de l'équation des directions caractéristiques.

Rappelons que les lieux caractéristiques sont tels qu'on ne peut y poser de condition de départ d'un problème de Cauchy.

Ainsi, $\partial u/\partial t + c\partial u/\partial x = 0$ détermine u à partir de ses valeurs sur un arc C à condition que C ne soit nulle part tangent à une droite $x - ct =$ constante : la dérivée tangentielle de u serait justement grad $u \cdot (c,1) = c\partial u/\partial x + \partial u/\partial t = 0$. Ceci montre bien sûr que u est constant sur ces caractéristiques : $u(x,t) = f(x-ct)$. Si la condition initiale n'est donnée que sur l'intervalle $a \leqslant x \leqslant b$, il faut aussi donner $u(a,t)$ si $c > 0$; $u(b,t)$ si $c < 0$:



On s'est habitué à évaluer u à partir de valeurs en $t = 0$. Ici, on a donc $f(x) = u(x,0)$ donnée $\Rightarrow u(x,t) = u(x-ct,0)$. La valeur de u en (x,t) ne dépend que de sa seule valeur initiale en $x - ct$.

Quelques discrétisations de $\partial u/\partial t = -c\partial u/\partial x$ sur $\{x_k = kh\}$.

(1) $\dfrac{u(x_k, t+\Delta t) - u(x_k,t)}{\Delta t} = -c\dfrac{u(x_k,t) - u(x_k - h,t)}{h}$. Upwind si $c > 0$.

(2) $\dfrac{u(x_k, t+\Delta t) - u(x_k,t)}{\Delta t} = -c\dfrac{u(x_k + h,t) - u(x_k,t)}{h}$. Upwind si $c < 0$.

(3) $\dfrac{u(x_k,t+\Delta t)-u(x_k,t)}{\Delta t} = -c\dfrac{u(x_k+h,t)-u(x_k-h,t)}{2h}$. Très mauvais !.

(4) $\dfrac{u(x_k,t+\Delta t)-\dfrac{u(x_k+h,t)+u(x_k-h,t)}{2}}{\Delta t} = -c\dfrac{u(x_k+h,t)-u(x_k-h,t)}{2h}$. Friedrichs.

Tous ces schémas sont consistants. Prenons le dernier, le plus intéressant : en $x = x_k$, $E(\Delta t)u(x)$ vaut $u(x-c\Delta t) = u - c\Delta t u' + ((c\Delta t)^2/2)u'' \cdots$ (Taylor). $E_h r_h u(x)$ est obtenu en isolant $u(x,t+\Delta t)$, donc vaut $[(1-(c\Delta t/h))u(x+h) + (1+(c\Delta t/h))u(x-h)]/2 = u - c\Delta t u' + (h^2/2)u'' + \cdots$, toujours par Taylor de $u(x \pm h) = u \pm hu' + (h^2/2)u'' + \cdots$. L'écart est donc $O((\Delta t)^2 + h^2) << \Delta t$ si $h = O(\Delta t)$.

Ce sont les propriétés de *stabilité numérique* qui justifient ou non ces choix. Les matrices E_h qui effectuent $[u_0,\ldots,u_N]^T(t+\Delta t) = E_h[u_0,\ldots,u_N]^T(t)$ sont :

1
$$\begin{bmatrix} 1-\alpha & & & \\ \alpha & 1-\alpha & & \\ & \ddots & \ddots & \\ & & \alpha & 1-\alpha \end{bmatrix}$$

2
$$\begin{bmatrix} 1+\alpha & -\alpha & & \\ & \ddots & \ddots & \\ & & 1+\alpha & -\alpha \\ & & & 1+\alpha \end{bmatrix}$$

3
$$\begin{bmatrix} 1 & -\alpha/2 & & \\ \alpha/2 & 1 & -\alpha/2 & \\ & \ddots & \ddots & \ddots \\ & & \alpha/2 & 1 \end{bmatrix}$$

4
$$\begin{bmatrix} 0 & (1-\alpha)/2 & & \\ (1+\alpha)/2 & 0 & (1-\alpha)/2 & \\ & \ddots & \ddots & \ddots \\ & & (1+\alpha)/2 & 0 \end{bmatrix}$$

où $\alpha = \dfrac{c\Delta t}{h}$. Ces matrices ne sont pas symétriques, il faut donc estimer leurs puissances plutôt que les valeurs propres de la seule matrice E_h.

Une condition suffisante simple de stabilité est que la somme des valeurs absolues sur chaque ligne et chaque colonne soit $\leqslant 1$ (voir chap. 7, p. 180).

(1) On tire que les puissances de la matrice E_h de la première méthode restent bornées si α et $1-\alpha \geqslant 0$, donc si $c > 0$ et $\Delta t \leqslant h/c$.

(2) La deuxième méthode est stable si $\alpha \leqslant 0$ et $1+\alpha \geqslant 0$ donc si $c < 0$ et $\Delta t \leqslant h/|c|$(Cf.[5]).

(3) La somme des valeurs absolues des éléments d'une ligne vaut maintenant $1+|\alpha| > 1$.
Par ailleurs, les valeurs propres de $\boldsymbol{I}+$ une matrice antisymmétrique (réelle) sont de la forme $1+i\beta$, avec $\beta = O(|\alpha|)$.
La condition nécessaire de von Neumann (voir chap. 7) demande donc $(1+\alpha^2)^{1/2} = 1 + O(\Delta t)$, qui aboutit à $\Delta t = O(h^2)$.

(4) La quatrième méthode, beaucoup plus élégante, retrouve la stabilité numérique dès que $|\alpha| \leqslant 1$, donc si $\Delta t \leqslant h/|c|$.

Le schéma "leap-frog" ("saute-mouton")

$$\frac{u(x,t+\Delta t)-u(x,t-\Delta t)}{2\Delta t} = -c\frac{u(x+h,t)-u(x-h,t)}{2h} \tag{72}$$

[5]Voir `http ://www.math.ucl.ac.be/~magnus/num2/upwind.txt`

est plus difficile à étudier. On a

$$E_h = \left[\begin{array}{cccc|c|cccc} 0 & 0 & \cdots & 0 & & 1 & 0 & \cdots & & 0 \\ 0 & 0 & \cdots & 0 & & 0 & 1 & \cdots & & 0 \\ \vdots & \vdots & \cdots & \vdots & & & & \ddots & & \\ 0 & 0 & \cdots & 0 & & 0 & 0 & \cdots & & 1 \\ -- & -- & -- & -- & + & -- & -- & -- & -- & -- \\ 1 & 0 & \cdots & 0 & & 0 & -\alpha & \cdots & & 0 \\ 0 & 1 & \cdots & 0 & & \alpha & & -\alpha & \cdots & \\ & & \ddots & & & & & \ddots & \ddots & \\ 0 & 0 & \cdots & 1 & & & & & \alpha & 0 \end{array}\right], \tag{73}$$

qui sera discuté plus loin.

Caractéristiques et domaine d'influence.

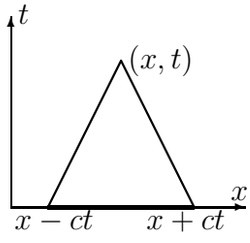
La solution u n'est pas toujours constante sur une caractéristique : prenons

$$\frac{\partial u}{\partial t} + a(x,t)\frac{\partial u}{\partial x} + b(x,t)u = F(x,t)$$

Si on donne u sur un arc de direction de tangente $\vec{\tau}$ passant par (x,t), on connaît $\operatorname{grad} u \cdot \vec{\tau} = \tau_x \partial u/\partial x + \tau_t \partial u/\partial t$ et on peut reconstruire le gradient de u sauf si $\tau_x/\tau_t = a(x,t)$. Les courbes caractéristiques sont donc $x = \phi(t)$ telles que $\phi'(t) = a(\phi(t),t)$. Sur une telle courbe, on a

$$\frac{du}{dt} = \frac{\partial u}{\partial x}\frac{dx}{dt} + \frac{\partial u}{\partial t} = \frac{\partial u}{\partial x}a(x,t) + \frac{\partial u}{\partial t} = -b(\phi(t),t)u + F(\phi(t),t).$$

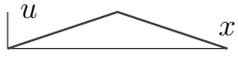
On voit bien que $u(x,t)$ dépend de la valeur $u(\phi(0),0)$ en $t = 0$ sur la caractéristique passant par (x,t).



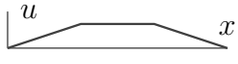
Équation des ondes $\partial^2 u/\partial t^2 = c^2 \partial^2 u/\partial x^2$. On a $u(x,t) = f(x-ct) + g(x+ct)$. On donne u et $\partial u/\partial t$ en $t = 0$: $f + g = u(x,0)$, $c(g' - f') = \partial u(x,0)/\partial t \Rightarrow f'$ et $g' = [\partial u(x,0)/\partial x \mp \partial u(x,0)/\partial(ct)]/2$, f et $g = [u(x,0) \mp \int^x \partial u(x,0)/\partial(ct)]/2$,

$$u(x,t) = \frac{u(x-ct,0)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} \partial u(\xi,0)/\partial t \; d\xi + \frac{u(x+ct,0)}{2}$$

montre bien que $u(x,t)$ ne dépend que des données initiales dans $[x-ct, x+ct]$: c'est un exemple de ***domaine d'influence***, déterminé par les deux caractéristiques passant par (x,t). N.B. : si $\partial u/\partial t = 0$ en $t = 0$, la solution est simplement $[u(x-ct,0) + u(x+ct,0)]/2$.



Exemple. Une corde de guitare est lâchée avec le déplacement $u(x,0) = u_0(1 - 2|x|/L)$ et $\partial u/\partial t = 0$ en $t = 0$. On a $\rho \partial^2 u/\partial t^2 = T \partial^2 u/\partial x^2$ [6]. La solution au temps t est donc $u_0(1 - |x-ct|/L - |x+ct|/L) = u_0(1 - 2ct/L)$ tant que $|x| \leqslant ct$, où $c = \sqrt{(T/\rho)}$. Que se passe-t-il quand $t > L/(2c)$?



[6] Exemple communiqué par J. Meinguet.

Jadis on nommait "influence" un liquide qui s'écoulait des astres sur la tête des hommes et les ondoyait dans la naissance.
P. Quignard, *Petits traités*, LVI[e] traité

Définition. *Soit un problème d'évolution bien posé dans $\mathbb{R}^{n-1} \times [0,T]$. Soit $t_2 > t_1$. Le* ***domaine d'influence*** *de (x,t_2) en $t = t_1$ est l'ensemble des (χ, t_1) où la donnée de u et des dérivées temporelles de u suffit à déterminer $u(x,t)$.*

Pour

$$\frac{\partial u}{\partial t} - K\frac{\partial^2 u}{\partial x^2} + \frac{K}{c^2}\frac{\partial^2 u}{\partial t^2} = 0, \tag{74}$$

(équation des télégraphistes, $K = 1/(\gamma r)$, $K/c^2 = \ell/r \Rightarrow c = 1/\sqrt{\gamma\ell}$), on pourrait encore donner une expression (compliquée, avec des fonctions de Bessel) montrant comment $u(x,t)$ dépend des données initiales, mais voici une autre façon de faire (Courant & Hilbert, vol. 2) :

Si u est une solution $\mathscr{C}^2$ de (74)*, et si u et $\partial u/\partial t$ sont nulles dans $[x-ct, x+ct]$ au temps $t = 0$, alors $u(x,t) = 0$.*

En effet, multiplions (74) par $2\partial u/\partial t$ et intégrons sur le triangle T de sommets $(x-ct,0)$, $(x+ct,0)$ et (x,t) dans le plan (χ,τ) :

$$\begin{aligned} 0 &= 2\frac{\partial u}{\partial \tau}\left[\frac{\partial u}{\partial \tau} - K\frac{\partial^2 u}{\partial \chi^2} + \frac{K}{c^2}\frac{\partial^2 u}{\partial \tau^2}\right] \\ &= 2\left(\frac{\partial u}{\partial \tau}\right)^2 - 2K\frac{\partial}{\partial \chi}\left(\frac{\partial u}{\partial \tau}\,\frac{\partial u}{\partial \chi}\right) + K\frac{\partial}{\partial \tau}\left(\left(\frac{\partial u}{\partial \chi}\right)^2\right) + \frac{K}{c^2}\frac{\partial}{\partial \tau}\left(\left(\frac{\partial u}{\partial \tau}\right)^2\right) \end{aligned}$$

et on intègre sur tout le triangle T. Les intégrales des trois derniers termes se résolvent immédiatement en intégrales sur la frontière de T :

$$0 = 2\int_T \left(\frac{\partial u}{\partial \tau}\right)^2 d\chi d\tau - 2K\int_0^t \left[\frac{\partial u}{\partial \tau}\frac{\partial u}{\partial \chi}\right]_{\chi_-}^{\chi_+} d\tau + K\int_{x-ct}^{x+ct}\left[\left(\frac{\partial u}{\partial \chi}\right)^2 + \frac{1}{c^2}\left(\frac{\partial u}{\partial \tau}\right)^2\right]_0^{\tau_+} d\chi,$$

où $\chi_\pm = x \pm c(t-\tau)$ sont les abscisses de ∂T en $t = \tau$, et $\tau_+ = t - |x-\chi|/c$ l'ordonnée correspondant à χ. Soit s l'abscisse curviligne le long de ∂T, on a $d\chi = c|d\tau| = ds/\sqrt{1+c^{-2}}$ (on ne s'occupe pas de la partie $\tau = 0$, puisque u et ses dérivées y sont nulles), on a enfin

$$0 = 2\int_T \left(\frac{\partial u}{\partial \tau}\right)^2 d\chi d\tau + \frac{K}{\sqrt{1+c^{-2}}}\int_{\partial T}\left[\frac{\partial u}{\partial \chi} \pm \frac{1}{c}\frac{\partial u}{\partial \tau}\right]^2 ds.$$

u est donc nulle dans tout T, donc en (x,t) : les valeurs initiales hors de $[x-ct, x+ct]$ n'ont aucune influence sur $u(x,t)$. □

Remarque : si $c \to \infty$, (74) tend vers l'équation de la chaleur, équation ***parabolique*** : les caractéristiques ne rencontrent plus l'axe des x (qui est en fait une courbe caractéristique!), la vitesse de propagation devient infinie (n'importe quel $u(\chi,0)$ a une influence sur $u(x,t)$, $\forall t > 0$).

Systèmes hyperboliques symétriques.

Équation pour $u(t, x_1, \ldots, x_{n-1})$:

$$\frac{\partial u}{\partial t} = \sum_1^{n-1} A_j \frac{\partial u}{\partial x_j}, \tag{75}$$

$u \in \mathbb{R}^m$, A_j est une matrice *symétrique* réelle (ou hermitienne complexe) d'ordre m.

Exemples

équation des ondes

$$n = 2 : \frac{\partial}{\partial t}\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 & c \\ c & 0 \end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix} u \\ v \end{bmatrix} \Rightarrow \frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}.$$

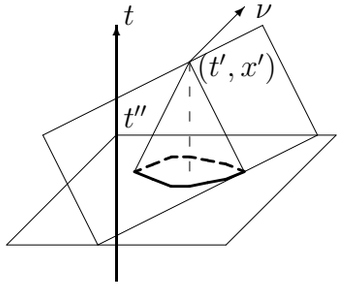
$$n=3: \frac{\partial}{\partial t}\begin{bmatrix}u\\v\end{bmatrix}=\begin{bmatrix}0&c\\c&0\end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix}u\\v\end{bmatrix}+\begin{bmatrix}c&0\\0&-c\end{bmatrix}\frac{\partial}{\partial y}\begin{bmatrix}u\\v\end{bmatrix}\Rightarrow \frac{\partial^2 u}{\partial t^2}=c^2\frac{\partial^2 u}{\partial x^2}+c^2\frac{\partial^2 u}{\partial y^2}.$$

équations de Maxwell $\mu\partial H/\partial t=-\mathbf{rot}\,E$, $\quad\varepsilon\partial E/\partial t=\mathbf{rot}\,H$:

$$\sqrt{\mu\varepsilon}\,\frac{\partial}{\partial t}\begin{bmatrix}u_1\\u_2\\u_3\\u_4\\u_5\\u_6\end{bmatrix}=\begin{bmatrix}&&&&&1\\&&&&-1&\\&&&&&\\&&-1&&&\\&1&&&&\end{bmatrix}\frac{\partial}{\partial x_1}\begin{bmatrix}u_1\\u_2\\u_3\\u_4\\u_5\\u_6\end{bmatrix}+\begin{bmatrix}&&&&&-1\\&&&&1&\\&&&1&&\\&&1&&&\\&-1&&&&\end{bmatrix}\frac{\partial}{\partial x_2}\begin{bmatrix}u_1\\u_2\\u_3\\u_4\\u_5\\u_6\end{bmatrix}+\begin{bmatrix}&&&&1&\\&&&-1&&\\&&&&&\\&&-1&&&\\1&&&&&\end{bmatrix}\frac{\partial}{\partial x_3}\begin{bmatrix}u_1\\u_2\\u_3\\u_4\\u_5\\u_6\end{bmatrix}$$

où $[u_1,u_2,u_3,u_4,u_5,u_6]=[\sqrt{\mu}H_1,\sqrt{\mu}H_2,\sqrt{\mu}H_3,\sqrt{\varepsilon}E_1,\sqrt{\varepsilon}E_2,\sqrt{\varepsilon}E_3]$,
(avec $\mathbf{div}\,E=\mathbf{div}\,H=0$), vitesse $c=1/\sqrt{\varepsilon\mu}$) : $n=4, m=6$.

Hyperplans caractéristiques.



Ce sont des hyperplans d'équation $\displaystyle\sum_{i=1}^{n-1}\nu_i x_i+\nu_0 t=\boldsymbol{\nu}_x\cdot\boldsymbol{x}+\nu_0 t=$ constante, tels que toute fonction de $\boldsymbol{\nu}_x\cdot\boldsymbol{x}+\nu_0 t$ (***onde plane***) soit solution de (75) homogène : $u(t,\boldsymbol{x})=\varphi(\boldsymbol{\nu}_x\cdot\boldsymbol{x}+\nu_0 t)\boldsymbol{u}\Rightarrow$

$$\det\left[\nu_0 I-\sum_{j=1}^{n-1}\nu_j A_j\right]=0.$$

Les hyperplans caractéristiques sont donc de normale $(\nu_0,\nu_1,\ldots,\nu_{n-1})$ *telle que*

$$\nu_0=\lambda\left(\sum_{j=1}^{n-1}\nu_j A_j\right).$$

À chaque direction $\boldsymbol{\nu}_x=(\nu_1,\ldots,\nu_{n-1})$ de $\mathbb{R}^{n-1}$ correspond donc m hyperplans d'équations $(x-x',\boldsymbol{\nu}_x)+(t-t')\nu_0=0$, où ν_0 est une des valeurs propres *supra*. Ne prenons que l'hyperplan correspondant à la plus grande valeur propre de $\sum_1^{n-1}\nu_j A_j$ et considérons toutes les directions $\boldsymbol{\nu}_x$. Les plans entourent un ***cône*** $\boldsymbol{C}$ de sommet (t',x').

5.2. Théorème. *Le domaine d'influence de* (t',x') *en* $t=t''<t'$ *est contenu dans le cône* $\boldsymbol{C}$.

En effet, résolvons (75) pour deux conditions initiales $\tilde{u}(x)$ et $\hat{u}(x)$ en $t=t''$ et montrons qu'on obtient la même valeur en (t',x') si $\tilde{u}$ et $\hat{u}$ coïncident dans $\boldsymbol{C}$. Par linéarité de (75), il suffit de montrer que $u(t',x')=0$ si $u(t'',x)=\tilde{u}(x)-\hat{u}(x)=0$ dans $\boldsymbol{C}\bigcap\{t=t''\}$. En fait, on va montrer que $u=0$ dans tout l'intérieur $\boldsymbol{D}$ de $\boldsymbol{C}$ pour $t''\leqslant t\leqslant t'$: multiplions

(75) par $2u^T$ et intégrons dans tout l'intérieur du cône,

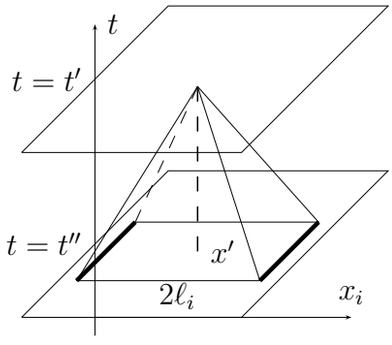
$$\begin{aligned}
0 &= \int_{\boldsymbol{D}} 2u^T \left[\frac{\partial u}{\partial t} - \sum_1^{n-1} A_j \frac{\partial u}{\partial x_j}\right] \, dtdx \\
&= \int_{\boldsymbol{D}} \left[\frac{\partial u^T u}{\partial t} - \sum_1^{n-1} \frac{\partial u^T A_j u}{\partial x_j}\right] \, dtdx \\
&= \int_{\boldsymbol{D}} \mathbf{div} \begin{bmatrix} u^T u \\ -u^T A_1 u \\ \vdots \\ -u^T A_{n-1} u \end{bmatrix} \, dtdx \\
&= \int_{\partial \boldsymbol{D}=\boldsymbol{C}} (u^T u \nu_0 - u^T A_1 u \nu_1 - \cdots - u^T A_{n-1} u \nu_{n-1}) \, dS \\
&= \int_{\boldsymbol{C}} u^T \left[\nu_0 I - \sum_1^{n-1} \nu_j A_j\right] u \, dS,
\end{aligned}$$

on voit apparaître une matrice semi définie positive, puisque $\nu_0 = \lambda_{\max}\left(\sum_{j=1}^{n-1} \nu_j A_j\right)$ sur chaque génératrice du cône. Prenons même un cône un peu plus grand, avec $\nu_0 >$ cette dernière valeur propre : on doit alors avoir $u = 0$ sur tout le cône. □

On voit d'ailleurs l'importance de la notion d'hyperplan caractéristique dans la représentation des solutions de (75) en ***superposition d'ondes planes*** :

$$u(x_1, \ldots, x_{n-1}, t) = \int_{\|\boldsymbol{\nu}\|=1} \left[\sum_{p=1}^{m} F_p(\boldsymbol{\nu} \cdot \boldsymbol{x} - t\lambda_p(\boldsymbol{\nu}) \boldsymbol{v}^{(p)}(\boldsymbol{\nu})\right] \, d\boldsymbol{\nu},$$

où $\lambda_p(\boldsymbol{\nu})$ et $\boldsymbol{v}^{(p)}(\boldsymbol{\nu})$, $p = 1, \ldots, m$ sont les valeurs et vecteurs propres de $\sum_{j=1}^{m} \nu_j A_j$, et où les F_p sont des fonctions scalaires arbitraires.



Nous n'aurons besoin que d'une version simplifiée du théorème ci-dessus :

Soit $t'' < t'$. La valeur $u(t', x')$ d'une solution de (75) *ne dépend que des valeurs $u(t'', y)$ dans l'hyperrectangle* $|y_i - x'_i| \leqslant \ell_i := (t' - t'')|\lambda(A_i)|_{\max}$, $i = 1, \ldots, n-1$.

En effet, considérons la pyramide construite sur la base indiquée (en $t = t''$, et on prend $\ell_i + \epsilon$), et de sommet (t', x'). L'intégrale sur la surface latérale de la pyramide est une somme de $2(n-1)$ integrales sur des faces d'équations $(\ell_i + \epsilon)(t - t') \pm (t' - t'')(x_i - x'_i) = 0$. La direction normale est donnée par le gradient, et nous devons donc évaluer l'intégrale de la forme quadratique $u^T[(\ell_i + \epsilon)I \mp (t' - t'')A_i]u$ sur ces faces, où les matrices $(\ell_i + \epsilon)I \mp (t' - t'')A_i$ ont des valeurs propres $\ell_i + \epsilon \mp (t' - t'')\lambda(A_i) > 0$ et sont donc définies positives.

5.3. Stabilité numérique.

On considère l'opérateur solution discret

$$(E_h u_h)(t,x) = \sum_{\boldsymbol{\beta}} e_{\boldsymbol{\beta}} u_h(t, x + \boldsymbol{\beta h}) \tag{76}$$

où $\boldsymbol{\beta} \in \mathbb{Z}^{n-1}$, $\boldsymbol{h} \in \mathbb{R}^{n-1}$.

Par ***exemple***, $u_h(t,x) + c\Delta t[u_h(t,x+h) - u_h(t,x)]/h$.

On s'attend bien sûr à trouver E_h dans un schéma ***consistant*** avec une EDP, par exemple, pour (75), E_h devra être proche de $E(\Delta t)$, lui-même proche (par Taylor) de

$$I + \Delta t \left(\sum_{j=1}^{n-1} A_j \frac{\partial}{\partial x_j} \right),$$

par exemple,

$$(E_h u_h)(t,x) = u_h(t,x) + \Delta t \left(\sum_{j=1}^{n-1} A_j [u_h(t, x+h_j) - u_h(t,x)]/h_j \right),$$

où $x + h_j$ désigne $(x_1, \ldots, x_{j-1}, x_j + h_j, x_{j+1}, \ldots, x_{n-1})$, ou encore

$$(E_h u_h)(t,x) = u_h(t,x) + \Delta t \left(\sum_{j=1}^{n-1} A_j [u_h(t, x+h_j) - u_h(t, x-h_j)]/(2h_j) \right).$$

Ces opérateurs ont de médiocres propriétés de stabilité numérique. Un premier exemple intéressant est ***l'opérateur de Lax-Friedrichs***

$$u_h(t+\Delta t, x) = (E_h u_h)(t,x) = \sum_{j=1}^{n-1} \left(\frac{u_h(t, x+h_j) + u_h(t, x-h_j)}{2(n-1)} + \Delta t A_j \frac{u_h(t, x+h_j) - u_h(t, x-h_j)}{2h_j} \right). \tag{77}$$

Théorème (Friedrichs).[7] *Si les matrices* $e_{\boldsymbol{\beta}}$ *de* (76) *sont hermitiennes, semi définies positives, et de somme* I*,*

$$\|E_h\| \leqslant 1,$$

et (76) *est donc* **numériquement stable**.

En effet, soit $\boldsymbol{u}$ un vecteur d'ordre Nm de valeurs $u(x)$ pour tous les x où sont définies les fonctions de U_h. $\|E_h\| = \max \|E_h \boldsymbol{u}\|$ sur $\boldsymbol{u} \in \mathbb{R}^{Nm}$ avec $\|\boldsymbol{u}\| \leqslant 1$, $\|E_h \boldsymbol{u}\|^2 = \boldsymbol{v}^T E_h \boldsymbol{u}$ avec $\boldsymbol{v} = E_h \boldsymbol{u}$. Ensuite,

$\boldsymbol{v}^T E_h \boldsymbol{u} = \sum_x \sum_\beta v^T(x) e_\beta u(x + \beta h)$,

$v^T(x) e_\beta u(x+\beta h) \leqslant \sqrt{v^T(x) e_\beta v(x)} \sqrt{u^T(x+\beta h) e_\beta u(x+\beta h)}$ [8],

donc $\boldsymbol{v}^T E_h \boldsymbol{u} \leqslant \sum_x \sum_\beta \sqrt{v^T(x) e_\beta v(x)} \sqrt{u^T(x+\beta h) e_\beta u(x+\beta h)}$

---

[7] K. Friedrichs, Symetric hyperbolic linear differential equations, *Comm. Pure Appl. Math.* **7** (1954) 345-392, =*Œuvres*. Cf. aussi V. Thomee, Stability theory for partial difference operators, *SIAM Rev.* **11** (1969) 152-195.

[8] Si e semi définie positive, $v^T e u = v^T e^{1/2} e^{1/2} u = (e^{1/2} v)^T e^{1/2} u \leqslant \|e^{1/2} u\| \, \|e^{1/2} v\|$ (Cauchy-Schwarz) $= \sqrt{u^T e u} \sqrt{v^T e v}$.

$\leqslant \sqrt{\sum_x \sum_\beta v^T(x) e_\beta v(x)} \sqrt{\sum_x \sum_\beta u^T(x+\beta h) e_\beta u(x+\beta h)}$ (Cauchy Schwarz),

$\leqslant \sqrt{\boldsymbol{v}^T \boldsymbol{v}}\ \sqrt{\boldsymbol{u}^T \boldsymbol{u}}$ (par $\sum_\beta e_\beta \leqslant I$ et par translation sur x pour les sommes en $u(x+\beta h)$),

enfin, $\|E_h \boldsymbol{u}\|^2 = \boldsymbol{v}^T E_h \boldsymbol{u} \leqslant \|\boldsymbol{v}\|\ \|\boldsymbol{u}\| = \|E_h \boldsymbol{u}\|\ \|\boldsymbol{u}\| \Rightarrow \|E_h\| \leqslant 1$. □

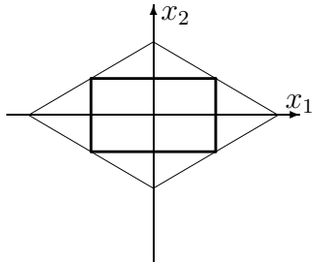
Reprenons maintenant le schéma de Friedrichs (77), on voit que $e_\beta = \dfrac{I}{2(n-1)} \pm \Delta t \dfrac{A_j}{2h_j}$, sur les $\beta = [0, \ldots, 0, \pm 1, 0, \ldots, 0]$. La condition de semi définition positive est $\lambda\left(e_\beta = \dfrac{I}{2(n-1)} \pm \Delta t \dfrac{A_j}{2h_j}\right) \geqslant 0$, donc

$$|\lambda(A_j)|\, \Delta t \leqslant \frac{h_j}{n-1} \tag{78}$$

pour toutes les valeurs propres de A_j, $j = 1, 2, \ldots, n-1$.

Domaine d'influence des équations aux différences.

Un schéma (explicite) aux différences montre immédiatement quelles valeurs $u_h(t, x + \beta h)$ entrent dans la détermination de $u_h(t + \Delta t, x)$.



Ainsi, dans (77), $u_h(t + \Delta t, x')$ dépend des $2(n-1)$ valeurs $u_h(t, x' \pm h_j[0, \ldots, 0, 1, 0, \ldots, 1])$ (sommets du losange dans la figure ci-contre).

D'autre part, par le théorème 5.2, le domaine d'influence au temps $t'' = t$ de $u(t', x')$, $t' = t + \Delta t$, est contenu dans les régions limitées par $\{x : \nu_0(t'' - t') + \sum_j \nu_j (x_j - x'_j) = -\Delta t \lambda(\sum_j \nu_j A_j) + \sum_j \nu_j (x_j - x'_j) = 0\}$, en particulier dans le

rectangle (pavé) délimité par les hyperplans $x_j - x'_j = \Delta t \lambda(A_j)$, $j = 1, \ldots, n-1$.

Ce lieu est bien dans le losange numérique

$$\sum_{j=1}^{n-1} \frac{|x_j - x'_j|}{h_j} \leqslant 1$$

si (78) est vérifié, puisque

$$\sum_{j=1}^{n-1} \frac{|x_j - x'_j|}{h_j} \leqslant \sum_{j=1}^{n-1} \frac{\Delta t |\lambda(A_j)|}{h_j} \leqslant 1.$$

On a donné ainsi un exemple de la ***condition de Courant-Friedrichs-Lewy***[9] : *Un schéma aux différences est numériquement stable si le domaine d'influence de l'équation*

[9] R. Courant, K.O. Friedrichs, H. Lewy, Über die partiellen Differenzengleichungen der mathematischen Physik, *Math. Ann.* **100** (1928) 32-74 = On the partial difference equations of mathematical physics, *IBM J. Research Dev.* **11** (1967) 215-234 = *Kurt Otto Friedrichs, Selecta*, vol.1, Birkäuser, 1986, 53-95 et 96-115.

P.D. Lax, Hyperbolic difference equations : a review of the Courant-Friedrichs-Lewy paper in the light of recent developments, *IBM J. Research Dev.* **11** (1967) 235-238 = *Kurt Otto Friedrichs, Selecta*, vol.1, Birkäuser, 1986, 49-52.

Cf. aussi J.M. Sanz-Serna, M.N. Spijker, Regions of stability, equivalence theorems and the Courant-Friedrichs-Lewy condition, *Numer. Math.* **49** (1986) 319-329.

différentielle au temps t d'une valeur au temps $t+\Delta t$ est contenu dans la fermeture convexe du domaine d'influence numérique.

L'un des domaines est contenu dans l'autre : comment ne pas se tromper ? Pour une discrétisation spatiale donnée, on améliore toujours la stabilité en diminuant Δt. Le cône de l'équadiff ne change pas d'inclinaison et découpe donc une région de plus en plus petite au temps t ; le cône numérique délimite toujours la même région dans l'ensemble des points de discrétisation spatiale : le domaine numérique ne varie pas.

Augmentation de la précision : Lax- Wendroff. On approche au second ordre

$$\begin{aligned} E(h) &= I + \Delta t \frac{\partial}{\partial t} + \frac{\Delta t^2}{2}\frac{\partial^2}{\partial t^2} + \cdots \\ &= I + \Delta t \left(\sum_{j=1}^{n-1} A_j \frac{\partial}{\partial x_j} \right) + \frac{\Delta t^2}{2} \left(\sum_{j=1}^{n-1}\sum_{i=1}^{n-1} A_i A_j \frac{\partial^2}{\partial x_i \partial x_j} \right) + \cdots \end{aligned}$$

(pour des matrices A_j constantes) par

$$E_h = I + \Delta t \left(\sum_{j=1}^{n-1} A_j \Delta_j \right) + \frac{\Delta t^2}{2} \left(\sum_{j=1}^{n-1}\sum_{i=1}^{n-1} A_i A_j \Delta_i \Delta_j \right)$$

où Δ_i est l'opérateur de différence partielle divisée

$$(\Delta_i u_h)(\boldsymbol{x}) = \frac{1}{2h_i} u_h \left(\boldsymbol{x} + h_i \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right) - \frac{1}{2h_i} u_h \left(\boldsymbol{x} - h_i \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right).$$

Schéma à 3 niveaux de temps : "leap-frog", ou "saute-mouton".

$$\frac{u_h(\boldsymbol{x}, t+\Delta t) - u_h(\boldsymbol{x}, t-\Delta t)}{2\Delta t} = \Delta t \left(\sum_{j=1}^{n-1} A_j \Delta_j u_h(\boldsymbol{x}, t) \right)$$

Stabilité de (73) (p. 170) à une variable spatiale : en procédant comme pour DuFort-Frankel, on arrive à des blocs $\begin{bmatrix} 0 & 1 \\ 1 & 2i\mu \end{bmatrix}$, avec $\mu \in (-|\alpha|, |\alpha|)$. Les valeurs propres $i\mu \pm \sqrt{1-\mu^2}$ ont des valeurs absolues > 1 si $|\alpha| > 1$. Même si $|\alpha| \leqslant 1$, on n'a pas la stabilité, sauf conditions particulières, voir [Iserles], aussi `http ://www.math.ucl.ac.be/~magnus/num2/leapfrog.txt`

Des méthodes ***adaptatives*** sont fort étudiées actuellement. Voici quelques abstracts d'une réunion d'analyse numérique tenue récemment :

Fonds National de la Recherche Scientifique

Groupe de Contact en Analyse Numérique

Numerical methods in ordinary and partial differential equations

Vendredi 5 décembre 1997, Friday December 5, 1997

Faculté Polytechnique de Mons

Laboratoire d'Automatique, boulevard Dolez 31, Mons

Space and Time Adaptivity in the Numerical Integration of Partial Differential Equations

W. E. Schiesser
Lehigh University
Bethlehem, PA 18015 USA
wes1@lehigh.edu

An essential aspect of the numerical integration of partial differential equations (PDEs) is the monitoring and control of numerical errors. In this paper, we review three basic approaches to adaptive error control (for discretization errors in space and time) :

(1) h refinement : Variation of the space and time integration intervals.

(2) p refinement : Variation in the order of the space and time approximations.

(3) r refinement : Adaptive movement of the grid points to regions of large variations in the solution.

These three strategies can be implemented manually (by observing the solution and making adjustments), or automatically (by having the computer code make the adjustments).

We illustrate these strategies within the method of lines (MOL), which is a flexible approach to the numerical integration of systems of ordinary differential equations (ODEs), differential-algebraic equations (DAEs) and PDEs.

Some representative numerical methods will be reviewed and illustrated through computer codes for : (a) a system of two initial-value ODEs of arbitrary stiffness and (b) a special case of the Einstein field equations of general relativity.

The talk will conclude with a brief discussion of the software available for MOL analysis, including a set of Matlab "m" files and the Fortran code for the Einstein field equations that will be provided to the attendees.

cf. `http ://www.lehigh.edu/~wes1/wes1.html` PDE : Method of lines.

An adaptive grid method based on smoothed equidistribution and its application to PDEs with higher order derivatives.

Dr. P. A. Zegeling
Dept. of Mathematics, University of Utrecht
P.O. Box 80.010, 3508 TA Utrecht, The Netherlands
e-mail : zegeling@math.ruu.nl
WWW : http ://www.math.ruu.nl/people/zegeling/

Abstract : Traditional numerical techniques to solve time-dependent PDEs integrate on a uniform spatial grid that is kept fixed on the entire time interval. If the solutions have regions of high spatial activity, a standard fixed-grid technique is computationally inefficient, since to afford an accurate numerical approximation, it should contain, in general, a very large number of grid points.

The grid on which the PDE is discretized then needs to be locally refined. Moreover, if the regions of high spatial activity are moving in time, like for steep moving fronts in reaction-diffusion or hyperbolic equations, then techniques are needed that also adapt (move) the grid in time : continuously deforming grid methods, which are also denoted by the term r-refinement.

In this talk I will describe an adaptive moving-grid method that is based on an equidistribution principle supplied with smoothing both in the time and space direction. In one space dimension this is rather straightforward and it can theoretically be shown that the ratios of adjacent spatial grid cells are forced to be bounded from below and above, due to the smoothing. In two dimensions, an adaptive moving grid can be defined by applying the 1D-principle along the two coordinate directions. Numerical results will be shown in one and two space dimensions for, among others, reaction-diffusion models and higher-order

PDEs in time and space, such as wave equations, KWdV-equation and the Extended Fisher-Kolmogorov PDE model.

Numerical Experiments with the Moving Finite Element Method

Alain Vande Wouwer
Faculté Polytechnique de Mons
Automatic Control Lab

In this talk, the moving finite element and the gradient weighted moving finite element methods for solving partial differential equations in one space dimension are briefly described. Several test-examples from science and engineering are used to illustrate the method features, including the selection of an initial node distribution, the parameter tuning and the use of matrix preconditioning.

Galerkin discontinu :

```
From: Roland Keunings <rk@mema.ucl.ac.be>
To: 'Cesame' <csam@csam.ucl.ac.be>; 'Winckelmans' <gsw@term.ucl.ac.be>;
'Cesame News' <hisette@auto.ucl.ac.be>
Cc: 'rk' <rk@mema.ucl.ac.be>
Sent: Monday, December 18, 2000 8:40 AM
Subject: annnonce seminaire

 Le Dr Jean-Francois Remacle (Scientific Computation Research Center
 Rensselaer Polytechnic Intsitute, USA) presentera un seminaire intitule "A
 parallel adaptive framework for solving complex fluid problems" le vendredi
 5 janvier a 11h a l'auditoire Euler.

 Le resume de son expose est repris ci-dessous.

 Cordialement

 Roland Keunings
 ========================
 Abstract

 We present a high-order formulation for solving hyperbolic
 conservation laws using the Discontinuous Galerkin Method (DGM).
 We introduce an orthogonal basis for the spatial discretization and use
 explicit Runge Kutta time discretization coupled with an original local
 time stepping scheme.

 Second part of the talk will focus on the parallel framework that
 is used in our DGM calculations.
 The Rensselaer Partitionning Model (RPM) is a paradigm for
 distributing data on parallel environnement. We apply RPM
 to distributed discretizations (meshes, octrees...) and
 discuss some important aspects of the problem like scalability,
 adaptivity, load balancing or network heterogeneity.
 The result is a parallel mesh database which specifications are presented.

 The numerical examples will focus on some challenging
 CFD problems (Richtmayer-Meshkov and Rayleigh-Taylor instabilities).
```

--
Prof. Roland Keunings
http://www.mema.ucl.ac.be/~rk

Centre for Systems Engineering and Applied Mechanics (CESAME)
Universite catholique de Louvain
Batiment Euler
Av. Georges Lemaitre, 4
B-1348 Louvain-la-Neuve
Belgium

Tel : +32-10-47 2350 (secr.) or 2087 (direct)
Fax : +32-10-47 2180
E-mail:rk@mema.ucl.ac.be
--

Voir aussi *Méthodes de Galerkin discontinues*, mémoire de C. Delfosse (MAP23, juin 2002)
http ://www.chez.com/cdelfoss/memoire.htm

Chapitre 7

Problèmes d'évolution : conditions de stabilité numérique.

Nous récapitulons et développons la condition de stabilité numérique

$$\|E_h^n\| \leqslant C, \qquad n = 1, 2, \ldots, \frac{T}{\Delta t},$$

qui doit être valable pour tout $\Delta t \to 0$, et où E_h est une matrice carrée réelle qui dépend de (et dont l'ordre dépend de) Δt. $T > 0$ est donné ; C peut dépendre de T (mais pas de Δt !).

Comme exemples, on a vu, pour la méthode d'Euler appliquée à l'équation de la chaleur, $\boldsymbol{u}(t + \Delta t) = (I - \sigma \Delta t M_h)\boldsymbol{u}(t)$, où $-M_h$ est la matrice familière de discrétisation de $-u''$:

$M_h = h^{-2} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & & -1 & 2 \end{bmatrix}$, donc

$E_h = I - \sigma \Delta t M_h = \begin{bmatrix} 1 - 2\sigma h^{-2}\Delta t & \sigma h^{-2}\Delta t & & & \\ \sigma h^{-2}\Delta t & 1 - 2\sigma h^{-2}\Delta t & \sigma h^{-2}\Delta t & & \\ & \ddots & \ddots & \ddots & \\ & & & \sigma h^{-2}\Delta t & 1 - 2\sigma h^{-2}\Delta t \end{bmatrix}$, matrice d'ordre inversement proportionnel à Δt !

Avec $\sigma = 1$, $h = 1/4$, $T = 1$ et quelques valeurs de Δt, on a

```
format long;h=0.25,
for nt=25:35;
 dt=1/nt; s=dt/h^2;
 Eh=s*diag(ones(nt-1,1),1)+(1-2*s)*diag(ones(nt,1))+s*diag(ones(nt-1,1),-1);
 no=[dt];for k=1:nt;no=[no,norm(Eh^k)]; end;  no,
end;
```

matlab

stabex

h = 0.25000000000000

$\Delta t = 1/25$, $\|E_h\| = 1.5506673588$, $\|E_h^2\| = 2.4045692578$, $\|E_h^3\| = 3.7286870601, \ldots,$ $\|E_h^{24}\| = 37363.4942340220$, $\|E_h^{25}\| = 57938.3509211103$.

$\Delta t = 1/30$, $\|E_h\| = 1.12786061161802$, $\|E_h^2\| = 1.27206955923938$, $\|E_h^3\| = 1.43471715110439, \ldots,$ $\|E_h^{30}\| = 36.95412303698682$.

$\Delta t = 1/32$, $\|E_h\| = 0.99547192257308$, $\|E_h^2\| = 0.99096434863135, \ldots, \|E_h^{32}\| = 0.86482549947922$.

exit

173111271 flops.

Les normes des puissances de E_h sont ici les mêmes puissances de la norme de E_h, ce qui simplifie la discussion. Cela se produit avec des matrices symétriques, ou plus généralement des matrices normales (voir plus loin).

Avec le schéma "leap-frog", on trouve un comportement moins simple :

```
format short;h=0.25,
for nt=5:10;
 dt=1/nt; s=dt/h;
 Eh=0*eye(2*nt); Eh(1:nt,nt+1:2*nt)=eye(nt);
                 Eh(nt+1:2*nt,1:nt)=eye(nt);
         Eh(nt+1:2*nt,nt+1:2*nt)= ...
                s*( diag(ones(nt-1,1),1)-diag(ones(nt-1,1),-1) );
 no=[dt];for k=1:nt;no=[no,norm(Eh^k)]; end;  no,
end;
```

stabex

h = 0.2500

$\Delta t = 1/5$, $\|E_h\| = 2.0157$, $\|E_h^2\| = 2.4800$, $\|E_h^3\| = 2.4468$, $\|E_h^4\| = 2.2044$, $\|E_h^5\| = 2.4029$.

$\Delta t = 1/6$, $\|E_h\| = 1.8309$, $\|E_h^2\| = 2.0315$, $\|E_h^3\| = 1.6243$, $\|E_h^4\| = 1.8127$, $\|E_h^5\| = 2.2129$, $\|E_h^6\| = 2.1102$.

...

$\Delta t = 1/8$, $\|E_h\| = 1.6025, \ldots$, $\|E_h^4\| = 1.6324, \ldots$, $\|E_h^8\| = 1.6365$.

$\Delta t = 1/10$, $\|E_h\| = 1.4693, \ldots$, $\|E_h^4\| = 1.5094, \ldots$, $\|E_h^9\| = 1.5097, \|E_h^{10}\| = 1.4224$.

exit 32063671 flops.

1. Norme matricielle.

La norme matricielle utilisée ici est toujours la norme matricielle subordonnée à la norme vectorielle euclidienne $\|\boldsymbol{x}\| = (\sum x_k^2)^{1/2} = (\boldsymbol{x} \cdot \boldsymbol{x})^{1/2}$.

On a donc $\|E_h\| = \sigma_{\max}(E_h) = [\lambda_{\max}(E_h^T E_h)]^{1/2}$.

Si E_h est ***symétrique***, on a $\|E_h\| = |\lambda(E_h)|_{\max}$.

Rappels :

(1) toute valeur absolue de valeur propre est inférieure à la norme. En effet, si $\boldsymbol{x} = \boldsymbol{v}$ est le vecteur propre correspondant à la valeur propre λ,

$$\|A\| = \max_{\boldsymbol{x}\neq 0} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \geqslant \frac{\|A\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = |\lambda|.$$

(2) Cette norme matricielle particulière est invariante par multiplication par une matrice unitaire :

$$\|UAV\| = \left[\lambda_{\max}(V^H A^H \underbrace{U^H U}_{=I} AV)\right]^{1/2} = \max_{\boldsymbol{x}\neq 0}\left[\frac{\boldsymbol{x}^H V^H A^H AV\boldsymbol{x}}{\boldsymbol{x}^H\boldsymbol{x}}\right]^{1/2} = \max_{\boldsymbol{y}=V\boldsymbol{x}\neq 0}\left[\frac{\boldsymbol{y}^H A^H A\boldsymbol{y}}{\boldsymbol{y}^H \underbrace{V^{-H}V^{-1}}_{=I}\boldsymbol{y}}\right]^{1/2} = \|A\|.$$

(N.B. : $\boldsymbol{x}^H := \overline{\boldsymbol{x}}^T$, $\boldsymbol{x}$ et U peuvent être complexes).

Inégalités. Par Gershgorin sur $A^H A$,

$$\|A\|^2 = \lambda_{\max}(A^H A) \leqslant \max_i \sum_{j=1}^m |(A^H A)_{i,j}| = \max_i \sum_{j=1}^m \left|\sum_{k=1}^m \overline{a_{k,i}} a_{k,j}\right|,$$

c'est-à-dire la plus grande somme des valeurs absolues des produits scalaires d'une colonne de A avec toutes les colonnes de A.

Borne supérieure par les valeurs absolues des éléments de A :

$$\|A\|^2 \leqslant \max_i \sum_{k=1}^{m} |a_{k,i}|\ell_k,$$

où $\ell_k := \sum_{j=1}^{m} |a_{k,j}|$, somme des valeurs absolues deséléments de la $k^{\text{ème}}$ ligne de A.

Enfin, $\|A\|^2 \leqslant (\max_i \gamma_i)(\max_k \ell_k)$, où $\gamma_i := \sum_{k=1}^{m} |a_{k,i}|$, somme des valeurs absolues deséléments de la $i^{\text{ème}}$ colonne de A. Démonstration directe : $\|A\|^2 = \lambda_{\max}(A^H A) \leqslant \|A^H A\|_\infty \leqslant \|A^H\|_\infty \|A\|_\infty = \|A\|_1 \|A\|_\infty$ (cas particulier d'inégalités de Marcel Riesz).

Borne inférieure. $\|A\| \geqslant$ la norme de toute sous-matrice (carrée ou rectangulaire) de A : $\|A\| = \max_{\boldsymbol{x}\neq 0} \dfrac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \geqslant$ la même expression où on ne retient que les vecteurs $\boldsymbol{x}$ avec des composantes non nulles aux indices des colonnes de la sous-matrice, et le numérateur est supérieur à la norme obtenue en ne retenant que les composantes dont les indices sont ceux des lignes de la sous-matrice.

2. Quelques conditions suffisantes.

2.1. Norme $\leqslant 1 +$ const.Δt.

Comme pour toute norme subordonnée à une norme vectorielle, on a $\|AB\| \leqslant \|A\|\ \|B\|$, donc tout va bien si $\|E_h\| \leqslant 1$, puisque $\|E_h^n\| \leqslant \|E_h\|^n$.

Par exemple, la condition (1.5) de la page 157 $\Delta t \leqslant h^2/(2\sigma)$ correspond à $\|E_h\| \leqslant 1$.

On peut se permettre une norme un peu plus grande que 1, si elle reste $\leqslant 1 + c\Delta t$, en effet

$$\|E_h\| \leqslant 1 + c\Delta t \implies \|E_h^n\| \leqslant (1 + c\Delta t)^n \leqslant (1 + c\Delta t)^{T/\Delta t} \leqslant e^{cT}.$$

2.2. Matrices symétriques, normales.

Une matrice symétrique, ou, plus généralement une ***matrice normale***[1], c'est-à-dire vérifiant $E_h \overline{E_h}^T = \overline{E}_h^T E_h$ admet toujours une base orthogonale de vecteurs propres, donc $E_h = U_h \Lambda_h U_h^{-1}$, avec U_h unitaire, donc de norme unité $\Rightarrow$ $E_h^n = U_h \Lambda_h^n U_h^{-1}$ de norme $(\max|\lambda(E_h)|)^n$.

La condition suffisante de stabilité numérique est donc alors $\max|\lambda(E_h)| \leqslant 1 + c\Delta t$.

2.3. Formes de Jordan et de Schur.

L'utilisation de la forme de Jordan $E_h = S_h J_h S_h^{-1}$ donne $E_h^n = S_h J_h^n S_h^{-1}$, très difficile à discuter, puisqu'il faudrait estimer $\|S_h\|\ \|S_h^{-1}\|$ = conditionnement de S_h, qui varie avec Δt !

[1] A normale $\iff A\overline{A}^T = \overline{A}^T A \iff A = B + iC$, avec B et C hermitiennes et $BC = CB$.

La forme de Schur $E_h = U_h T_h U_h^{-1}$, où U_h est unitaire, demande d'apprécier les normes des puissances de la matrice triangulaire T_h. On peut passer de la forme de Jordan à la forme de Schur en orthogonalisant par Gram-Schmidt la base des vecteurs propres et vecteurs principaux formant les colonnes de S_h : $S_h = U_h R_h$ (= QR de S_h), $T_h = R_h J_h R_h^{-1}$, ces trois dernières matrices étant triangulaires supérieures.

N.B. : les éléments diagonaux de J_h et T_h sont les valeurs propres de E_h.

3. Condition nécessaire de von Neumann.

La condition sur les valeurs propres de E_h, suffisante pour des matrices normales, est toujours nécessaire :

Théorème. $\|E_h^n\| \leqslant C,\ n\Delta t \leqslant T \Rightarrow |\lambda(E_h)| \leqslant 1 + c\Delta t$.

En effet $\|Ax\| \leqslant \|A\|\ \|x\|$ pour tout $x \Rightarrow \|A\| \geqslant |\lambda(A)\|$: il suffit de prendre $x =$ un vecteur propre de A. Donc $\|A^m\| \geqslant |\lambda(A^m)| = |\lambda(A)|^m$.

Ici, $|\lambda(E_h)|$ doit donc être inférieur à une constante$^{1/n} = e^{K/n} \leqslant e^{K\Delta t/T}$ quand n prend la plus grande valeur possible $T/\Delta t$ (on suppose évidemment $K > 0$). La condition sur le rayon spectral de E_h est donc d'autant plus sévère que Δt (qui $\to 0$ quand $h \to 0$) est petit. En tout cas, comme $\Delta t \leqslant T$, $e^{K\Delta t/T} \leqslant$ la valeur en $K\Delta t/T$ de l'interpolant linéaire de l'exponentielle en 0 et en K, interpolant qui vaut $1 + c\Delta t$ [2]. □

Si toutes les valeurs propres des différentes matrices E_h étaient confinées dans un disque $|\lambda| \leqslant \rho < 1$, il n'y aurait pas de problème (pas si évident, voir plus loin), mais ça n'arrive jamais ! Par **consistance**, les matrices E_h doivent avoir des valeurs propres proches de 1 !

En effet, $E_h r_h u(0)$ doit être proche de $r_h u(\Delta t)$, donc de $r_h u(0)$: $\|E_h r_h u(0) - r_h u(0)\| \to 0$ pour un ensemble suffisamment grand de fonctions $u(0)$. En tout cas, on voit que $\|(E_h - I)^{-1}\| \to \infty$: sinon, $\|r_h u(0)\| = \|(E_h - I)^{-1}[E_h r_h u(0) - r_h u(0)]\|$ tendrait vers 0 quand $h \to 0$.

On peut montrer [Richtmyer, p.69] que, si les valeurs propres de E_h sont toutes *sauf une* dans $|\lambda| \leqslant \rho < 1$, la condition de von Neumann suffit.

4. Théorème de Kreiss.

Il s'agit de préciser des conditions nécessaires et suffisantes de stabilité numérique en résumant et en étendant ce qui vient d'être fait.

4.1. Préparation.

On commence par effectuer des similitudes unitaires $U_h^{-1} E_h U_h$ de manière à faire apparaître des blocs diagonaux de taille ***bornée par une constante***. Par exemple, le schéma de DuFort-Frankel aboutissait à des blocs $\begin{bmatrix} 0 & 1 \\ -\dfrac{1-\alpha}{1+\alpha} & \dfrac{2-\mu_i}{1+\alpha} \end{bmatrix}$, où $\alpha = h^2/(2\sigma\Delta t)$ et $\mu_i =$ une valeur propre de $h^2 M_h$ est comprise entre 0 et 4.

[2] En fait, $1 + \dfrac{e^K - 1}{K} K \dfrac{\Delta t}{T}$.

Constatons ensuite d'abord que la connaissance d'une famille de matrices E_h numériquement stable équivaut à la connaissance d'une famille $\{A_\alpha\}$ de matrices dont ***toutes*** les puissances positives sont uniformément bornées :

$$\|A_\alpha^N\| \leqslant C, \qquad N = 0, 1, \ldots$$

(Remarquons que $N = 0 \Rightarrow C \geqslant 1$).

En effet,

(1) Cond. suffisante. Si $\left\|\left(C^{-\Delta t/(2T)} E_h\right)^n\right\| \leqslant C^{1/2}$, pour tout n entier $\geqslant 0$, alors $\|E_h^n\| \leqslant C$ pour $n = 0, 1, \ldots, T/\Delta t$.

(2) Cond. nécessaire. Si $\|E_h^n\| \leqslant C$ pour $n = 0, 1, \ldots, T/\Delta t$, évaluons une borne supérieure de $\|E_h^N\|$, N entier positif quelconque : soit $N = q(T/\Delta t) + r$, quotient et reste de la division de N par $T/\Delta t$.

$$\|E_h^N\| = \|E_h^{q(T/\Delta t)+r}\| \leqslant \|E_h^{T/\Delta t}\|^q \|E_h^r\| \leqslant C^{q+1} \leqslant C^{1+N\Delta t/T},$$

donc $\left\|\left(C^{-\Delta t/T} E_h\right)^N\right\| \leqslant C$, pour tout N entier $\geqslant 0$.

4.2. Le théorème de Kreiss.

Théorème[3]**.** *Soit une famille $\{A_\alpha\}$ de matrices complexes carrées d'ordre fixé m. Les quatre propositions suivantes sont équivalentes :*

(1) $\|A_\alpha^N\| \leqslant C_1$, *avec la même constante C_1 pour tout élément A_α de la famille de matrices, et pour tout entier $N \geqslant 0$.*

(2) *Pour tout z complexe avec $|z| > 1$, chaque matrice $zI - A_\alpha$ est inversible, et*

$$\|(zI - A_\alpha)^{-1}\| \leqslant \frac{C_2}{|z| - 1}. \tag{79}$$

(3) $\exists C_3, C_4$ *constantes pour toute la famille $\{A_\alpha\}$, et chaque matrice de la famille peut être mise sous la forme $A_\alpha = S_\alpha T_\alpha S_\alpha^{-1}$, avec $\|S_\alpha\|$ et $\|S_\alpha^{-1}\| \leqslant C_3$, T_α triangulaire d'éléments diagonaux (valeurs propres de A_α) de valeurs absolues $\leqslant 1$, et d'éléments non diagonaux vérifiant $|t_{i,j}| \leqslant C_4 \min(1 - |t_{i,i}|, 1 - |t_{j,j}|)$.*

(4) $\exists C_H \geqslant 1$ *constante pour toute la famille $\{A_\alpha\}$, et on peut associer à chaque matrice A_α une matrice hermitienne définie positive H_α vérifiant $\|H_\alpha\| \leqslant C_H$, $\|H_\alpha^{-1}\| \leqslant C_H$, et*
$H_\alpha - A_\alpha^H H_\alpha A_\alpha$ semi-définie positive.

4.3. Preuve. (d'après [Richtmyer & Morton, § 4.9])

– $1 \Rightarrow 2$. Comme les valeurs absolues des valeurs propres sont majorées par la norme, $|\lambda(A^n)| = |\lambda(A)|^n \leqslant C_1$, pour tout n entier positif, donc $|\lambda(A)| \leqslant 1$, et $zI - A$ est inversible si $|z| > 1$ (les valeurs propres de $zI - A$ sont $z - \lambda(A)$).

Dès lors, la série convergente $\displaystyle\sum_{k=0}^{\infty} z^{-k-1} A^k$ vaut bien $(zI - A)^{-1}$: la somme des p

[3] H.O. Kreiss, Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differenzialgleichungen approximieren, *BIT*, **2** (1962) 153-181.

premiers termes vaut $(zI-A)^{-1}(I-z^{-p}A^p) \to (zI-A)^{-1}$ quand $p \to \infty$ (et $|z|>1$). Et la norme de la série est bornée par $\sum_{k=0}^{\infty} C_1|z|^{-k-1} = C_1(|z|-1)^{-1}$. □

– $2 \Rightarrow 3$. On construit progressivement T en partant de la matrice triangulaire figurant dans la forme de Schur de A.
L'élément $(i,i+1)$ de $(zI-T)^{-1}$ est $-\dfrac{t_{i,i+1}}{(z-\lambda_i)(z-\lambda_{i+1})}$, borné par $C_2/(|z|-1)$, donc $|t_{i,i+1}| \leqslant C_2\dfrac{|z-\lambda_i|\,|z-\lambda_{i+1}|}{|z|-1}$ pour tout $|z|>1$ (on rappelle que $t_{i,i}=\lambda_i$). Soit $|\lambda_i| \geqslant |\lambda_{i+1}|$, prenons z aligné sur λ_i : $z=|z|\lambda_i/|\lambda_i|$, $|z-\lambda_i| = |z|-|\lambda_i|$, $|z-\lambda_{i+1}| \leqslant |z-\lambda_i|+|\lambda_i-\lambda_{i+1}|$. Avec $|z|=2-|\lambda_i|$, on voit que $|t_{i,i+1}| \leqslant 4C_2(1-|\lambda_i|)+2C_2|\lambda_i-\lambda_{i+1}|$.
Effectuons maintenant la similitude

$$T \longrightarrow \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & -\kappa & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix} T \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & \kappa & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix}$$

où κ se trouve en $(i,i+1)$. Le résultat est encore triangulaire supérieur, les éléments modifiés sont ceux de la $i^{\text{ème}}$ ligne $t_{i,j} \longrightarrow t_{i,j} - \kappa t_{i+1,j}$ pour $j > i+1$, et ceux de la $(i+1)^{\text{ème}}$ colonne $t_{j,i+1} \longrightarrow t_{j,i+1}+\kappa t_{j,i}$ pour $j<i$, et enfin $t_{i,i+1} \longrightarrow t_{i,i+1}+\kappa(\lambda_i-\lambda_{i+1})$. Comme on a vu que $t_{i,i+1} = 4C_2\alpha(1-|\lambda_i|) + 2C_2\beta(\lambda_i-\lambda_{i+1})$, avec $|\alpha|$ et $|\beta| \leqslant 1$, il suffit de prendre $\kappa = -2C_2\beta$ pour obtenir le nouveau $|t_{i,i+1}| \leqslant 4C_2(1-|\lambda_i|) = 4C_2\min(1-|\lambda_i|, 1-|\lambda_{i+1}|)$. On reprend pour tous les i. Les normes des matrices de similitude sont multipliées au plus par $(1+2C_2)$ à chaque fois.
On examine maintenant les $t_{i,i+k+1}$ avec $k>0$:
Pour $k = 1, 2, \ldots, m-1$:

– Si on est arrivé à une matrice triangulaire T dont tous les éléments $t_{i,j}$ vérifient 3. (avec une constante provisoire C_4') quand $j-i \leqslant k$, on examine les éléments $t_{i,i+k+1}$:
– Pour $i = 1, 2, \ldots, m-k-1$:
pour évaluer une borne de $|t_{i,i+k+1}|$ à partir de la borne $\|(zI-T)^{-1}\| \leqslant C_2'/(|z|-1)$, écrivons l'élément $(i,i+k+1)$ de $(zI-T)^{-1}(zI-T) = I$:

$$\begin{aligned} 0 &= \sum_{p=i}^{i+k+1} (zI-T)^{-1}_{i,p}(zI-T)_{p,i+k+1} \\ &= -\frac{t_{i,i+k+1}}{z-\lambda_i} - \sum_{p=i+1}^{i+k} (zI-T)^{-1}_{i,p}\, t_{p,i+k+1} + (zI-T)^{-1}_{i,i+k+1}(z-\lambda_{i+k+1}). \end{aligned}$$

On obtient donc $t_{i,i+k+1} = (z-\lambda_i)\times$ une somme de $k+1$ produits d'un élément de $(zI-T)^{-1}$ et de $z-\lambda_{i+k+1}$ ou d'un élément que l'on sait déjà être borné par $C_4'(1-|\lambda_{i+k+1}|) \leqslant C_4'(|z|-|\lambda_{i+k+1}|) \leqslant C_4'|z-\lambda_{i+k+1}|$, d'où

$$|t_{i,i+k+1}| \leqslant (1+kC_4')C_2'\frac{|(z-\lambda_i|\,|z-\lambda_{i+k+1}|}{|z|-1},$$

on se ramène au cas déjà discuté, avec une transformation de similitude $T \to (I-K)T(I+K) = T - KT + TK$, où K ne contient qu'un seul élément non nul κ en position $(i, i+k+1)$. On prendra $\kappa = -2(1+kC_4')C_2'\beta$ avec $|\beta| \leqslant 1$ pour obtenir $|t_{i,i+k+1}|_{\text{new}} = |t_{i,i+k+1,\text{old}} + \kappa(\lambda_i - \lambda_{i+k+1})| \leqslant 4(1+kC_4')C_2' \min(1-|\lambda_i|, 1-|\lambda_{i+k+1}|)$. Les autres $t_{p,q}$ modifiés sont

$$T \longrightarrow T - \kappa \underbrace{\begin{bmatrix} 0 & 0 & \cdots & 0 & t_{i+k+1,i+k+1} & t_{i+k+1,i+k+2} & \cdots & t_{i+k+1,m} \end{bmatrix}}_{\text{ligne } i} + \kappa \underbrace{\begin{bmatrix} t_{1,i} \\ t_{2,i} \\ \vdots \\ t_{i,i} \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\text{colonne } i+k+1}$$

et vérifient $j > i+k+1$, et seront donc traités lors des étapes ultérieures. Le nouveau C_4' vaut $\max(C_4', 4(1+kC_4')C_2'$, C_3' est multiplié par $1+2(1+kC_4')C_2'$, et le nouveau C_2' est multiplié par le carré de cette dernière valeur…

– end i.

end k □

– $3 \Rightarrow 4$. Si $A = STS^{-1}$ vérifie (3), il existe $K > 1$ (constant pour toute la famille des matrices A) tel que $\|DTD^{-1}\| \leqslant 1$, où D est la matrice diagonale d'éléments diagonaux $K, K^2, \ldots, K^m$. En effet, les éléments diagonaux de DTD^{-1} sont les $t_{i,i}$ inchangés ; et les éléments non diagonaux sont $K^{i-j}t_{i,j}$ avec $j > i$. La somme des valeurs absolues desléléments d'une ligne est donc bornée par $|t_{i,i}| + C_4(1-|t_{i,i}|)(K^{-1} + K^{-2} + \cdots) \leqslant 1$ si $K \geqslant 1 + C_4$. Il en va de même pour les colonnes, d'où $\|DTD^{-1}\| \leqslant 1$. Toutes les valeurs propres de $D^{-1}T^HD^2TD^{-1}$ sont donc $\leqslant 1$, $I - D^{-1}T^HD^2TD^{-1}$ est semi-définie positive, ainsi que $D^2 - T^HD^2T$, ce qui établit (4) avec $H = S^{-H}D^2S^{-1}$. On a $\|H\| \leqslant C_3^2K^{2m}$, $\|H^{-1}\| \leqslant C_3^2K^{-2}$.

– $4 \Rightarrow 1$. Pour tout vecteur de départ $\boldsymbol{w}_0$, considérons l'itération $\boldsymbol{w}_{k+1} = A\boldsymbol{w}_k$. La suite $\boldsymbol{w}_k^H H \boldsymbol{w}_k$ est décroissante :

$$\boldsymbol{w}_k^H H \boldsymbol{w}_k - \boldsymbol{w}_{k+1}^H H \boldsymbol{w}_{k+1} = \boldsymbol{w}_k^H (\underbrace{H - A^H H A}_{\text{déf. pos.}}) \boldsymbol{w}_k \geqslant 0,$$

$$C_H^{-1} \boldsymbol{w}_k^H \boldsymbol{w}_k \leqslant \boldsymbol{w}_k^H H \boldsymbol{w}_k \leqslant \boldsymbol{w}_0^H H \boldsymbol{w}_0 \leqslant C_H \boldsymbol{w}_0^H \boldsymbol{w}_0,$$

d'où $\|\boldsymbol{w}_k\|^2 = \|A^k \boldsymbol{w}_0\|^2 \leqslant C_H^2 \|\boldsymbol{w}_0\|^2$, et $\|A^k\| \leqslant C_H$. □

Exemple de Dufort-Frankel $E_\theta = \begin{bmatrix} 0 & 1 \\ -\dfrac{1-\alpha}{1+\alpha} & \dfrac{2\theta}{1+\alpha} \end{bmatrix}$, où α est fixé entre 0 et 1, et où θ peut parcourir tout l'intervalle $[-1, 1]$.

Les valeurs propres sont $\lambda_1(\theta)$ et $\lambda_2(\theta) = \dfrac{\theta \pm \sqrt{\theta^2 - 1 + \alpha^2}}{1+\alpha}$, complexes de même valeur absolue $[(1-\alpha)/(1+\alpha)]^{1/2}$ quand $|\theta| < \sqrt{1-\alpha^2}$, réelles dans les autres cas avec $\lambda_1 \to \pm 1$ quand $\theta \to \pm 1$. On commence par regarder la forme de Schur : la première colonne de U est le vecteur propre $[1, \lambda_1]^T / \sqrt{1 + |\lambda_1|^2}$, et on construit la seconde colonne orthogonale à la première :

```
%  DuFort Frankel
% Schur:
E=sym('[0,1 ; -L1*L2,L1+L2]'),
U=sym('[1,-conj(L1);L1,1]');U=symop(U,'/','sqrt(1+abs(L1)^2)' ),
EU=symop(E,'*',U);UH=inverse(U);T=symop(UH,'*',EU),
```

ce qui donne une matrice triangulaire $\begin{bmatrix}\lambda_1 & 1+\overline{\lambda_1}\lambda_2\\ 0 & \lambda_2\end{bmatrix}$

```
dufort

E=  [      0,   1 ]
    [ -L1*L2,L1+L2]

U = [   1/(1+abs(L1)^2)^(1/2), -1/(1+abs(L1)^2)^(1/2)*conj(L1)]
    [1/(1+abs(L1)^2)^(1/2)*L1,            1/(1+abs(L1)^2)^(1/2)]

T = [L1, L2*conj(L1)+1]
    [ 0,            L2]
```

Cette matrice convient tant que les valeurs propres sont complexes, avec $C_4 \leqslant \dfrac{1+(1-\alpha)/(1+\alpha)}{1-\sqrt{(1-\alpha)/(1+\alpha)}}$.

Quand les valeurs propres sont réelles, $t_{1,2}$ n'est plus uniformément bornée par un multiple de $1-|\lambda_1|$ quand λ_1 se rapproche de ± 1. On pratique alors une similitude par $\begin{bmatrix}1 & \kappa\\ 0 & 1\end{bmatrix}$:

```
%similitude
UK=sym('[1,k;0,1]');T2=symop(inverse(UK),'*',T,'*',UK),
T2 =

T2 = [L1, L1*k+L2*conj(L1)+1-k*L2]
     [ 0,                    L2]
```

où le nouvel élément $(1,2)$ vaut donc

$$1+\lambda_1\lambda_2+\kappa(\lambda_1-\lambda_2) = \frac{2+\kappa(1+\alpha)\lambda_1-\kappa(1-\alpha)/\lambda_1}{1+\alpha}$$

Avec $\kappa = -\operatorname{sign}(\lambda_1)/\alpha$, on obtient $\dfrac{[1-\alpha+(1+\alpha)|\lambda_1|][1-|\lambda_1|]}{\alpha(1+\alpha)|\lambda_1|} \leqslant \text{const.}\ (1-|\lambda_1|)$.

Développements récents : voir le remarquable E. Wegert and L. N. Trefethen, "From the Buffon needle problem to the Kreiss matrix theorem," *Amer. Math. Monthly* **101** (1994), 132-139, `http ://web.comlab.ox.ac.uk/oucl/work/nick.trefethen/buffon.ps.gz`

On s'est beaucoup occupé de quantifier la borne de $\|A^N\|$ à partir de (79). La réponse est $eC_2 \min(m, N-1)$ (Spijker).

Chapitre 8

Méthodes (pseudo) spectrales.

Extraits de

D. Gottlieb, S.A. Orszag, *Numerical Analysis of Spectral Methods : Theory and Applications*, SIAM Reg. Conf. Appl. Math., 1977,

C. Canuto, M.Y. Hussaini, A. Quarteroni, T.A. Zang, *Spectral Methods in Fluid Dynamics*, Springer, 1988.

B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Cambridge University Press, 1996, paperback : 1999.

1. Fonctions propres d'opérateurs autoadjoints.

Les auteurs classiques ont toujours cherché à résoudre des problèmes comme $Au = f$ ou $\partial u/\partial t = Au$, etc. par développements en séries de fonctions bien adaptées à la géométrie du problème et aux conditions aux limites.

Après les séries de Taylor, toujours précieuses en matière de fonctions analytiques, d'autres séries potentielles, les séries de ***Fourier*** ont fourni un paradigme. On a vu que rien n'est plus commode que de "voir" et discuter l'évolution de $\partial u/\partial u(x,t) = \sigma\partial^2 u/\partial x^2$ avec $u(0,t) = u(1,t) = 0$ par la série $u(x,t) = \sum_{m=1}^{\infty} A_m(0)\exp(-\sigma\pi^2 m^2 t\ell^2)\sin(\pi m x/\ell)$, que l'on peut parfaitement utiliser pour l'évaluation numérique de u.

Comme déjà vu auparavant, la conjonction d'un opérateur autoadjoint et d'une base $\{\varphi_j\}$ de fonctions propres connues autorise cette économie de la représentation : les φ_j sont orthogonales

$$\lambda_j(\varphi_j,\varphi_k) = (A\varphi_j,\varphi_k) = (\varphi_l, A\varphi_k) = \lambda_k(\varphi_j,\varphi_k) \Rightarrow (\varphi_j,\varphi_k) = 0, j \neq k,$$

ce qui donne une formule pour représenter toute fonction admissible dans la base des φ_j :

$$\forall f \in U,\quad f = \sum_j \frac{(f,\varphi_j)}{\|\varphi_j\|^2}\varphi_j,$$

et $Au = f$ se résout évidemment par

$$u = \sum_j \frac{(f,\varphi_j)}{\lambda_j\|\varphi_j\|^2}\varphi_j.$$

Toutes les configurations classiques ont -probablement- été épuisées depuis longtemps et on ne disposera probablement pas exactement de la base de fonctions propres pour le problème à résoudre. Aussi, on se sert dans des répertoires bien reconnus :

Un problème de **Sturm-Liouville**

$$Ly(x) := -\frac{d}{dx}\left[p(x)\frac{dy(x)}{dx}\right] + q(x)y(x) = \lambda y(x),$$

$p(a)y(a) = p(b)y(b) = 0$, est autoadjoint (au moins formellement) :

$$(u, Lv) = \int_a^b u\{-[pv']' + qv\}dx = [-puv']_a^b + \int_a^b \{pu'v' + quv\}dx = (Lu, v),$$

et admet donc des fonctions propres φ_n orthogonales.

Quelques exemples, outre les fonctions trigonométriques :

Polynômes de Legendre P_n : $-[(1-x^2)P_n']' = n(n+1)P_n$,

polynômes de Tchebycheff T_n : $-[(1-x^2)\varphi_n']' + \dfrac{2-x^2}{4(1-x^2)}\varphi_n = n^2\varphi_n, \varphi_n = (1-x^2)^{-1/4}T_n$,

polynômes de Jacobi $P_n^{(\alpha,\beta)}$: $\varphi_n = (1-x)^{\alpha/2}(1+x)^{\beta/2}P_n^{(\alpha,\beta)}$,

$$-[(1-x^2)\varphi_n']' + \frac{(\alpha+\beta)(\alpha+\beta+2)x^2 + 2(\alpha^2-\beta^2)x + (\alpha-\beta)^2 - 2(\alpha+\beta)}{4(1-x^2)}\varphi_n = n(n+\alpha+\beta+1)\varphi_n,$$

polynômes de Laguerre $L_n^{(\alpha)}$: $-(x\varphi_n')' + \dfrac{x^2+\alpha^2}{4x}\varphi_n = (n+\alpha+1/2)\varphi_n,\ \varphi_n = x^{\alpha/2}e^{-x/2}L_n^{(\alpha)}(x)$

polynômes d'Hermite $-\varphi_n'' + x^2\varphi_n = (2n+1)\varphi_n,\ \varphi_n = e^{-x^2/2}H_n(x)$.

D'après les fonctions propres du laplacien sur un disque ou une boule : système de Fourier-Bessel : $-\varphi_n'' + \dfrac{(d-1)(d-3)}{4x^2}\varphi_n = \lambda\varphi_n, \quad \varphi_n(x) = \sqrt{x}J_{(d-2)/2}(x\sqrt{\lambda_n})$, avec λ_n tel que $\varphi_n(R) = 0$.

2. Calcul en représentation spectrale.

Comme les φ_n ne seront probablement pas les vraies fonctions propres du problème que l'on a a traiter, on cherche à obtenir aussi exactement que possible Au_h à partir d'une représentation $u_h = \sum_n \alpha_n\varphi_n$, où les α_n sont les inconnues.

On cherchera ainsi à exprimer la multiplication d'une telle somme par des fonctions figurant dans l'opérateur A, et à représenter correctement des opérations de dérivation. Pour chaque choix plus ou moins judicieux de fonctions de base, on utilisera des formules de récurrence, identités différentielles, etc.

2.1. Ritz-Galerkin.

On résout $(Lu_h, v) = a(u_h, v) = \varphi(v)$, $\forall v \in U_h$, en utilisant une base de U_h :

Par **exemple**, résolvons

$$Lu = -u'' + u = 1, \qquad -1 \leqslant x \leqslant 1, \quad u(-1) = u(1) = 0.$$

La solution $1 - (\cosh x)/(\cosh 1)$, et les fonctions propres $-\psi'' + \psi = \lambda\psi$ sont des combinaisons d'exponentielles et de fonctions trigonométriques.

Soit U_h = polynômes trigonométriques nuls en ± 1 de base $\{\cos((k+1/2)\pi x),\ k = 0, 1, \ldots, n$. Donc,

$$u_h = \sum_{k=0}^{n} c_k \cos((k+1/2)\pi x, Lu_h = -u_h'' + u_h = \sum_{k=0}^{n}[(k+1/2)^2\pi^2 + 1]c_k \cos((k+1/2)\pi x,$$

$$0 = \int_{-1}^{1}(Lu_h - 1)\cos(p+1/2)\pi x\ dx = [(p+1/2)^2\pi^2 + 1]c_p - \frac{2(-1)^p}{(p+1/2)\pi},$$

$p = 0, 1, \ldots, n$ par orthogonalité des cosinus. On obtient une somme partielle de la série de Fourier de la solution u. La série est assez lentement convergente car l'extension périodique de u n'est que $\mathscr{C}^1_I \ldots$

On n'a donc pas trop bien choisi le sous-espace d'approximation.

Soit $U_h = \mathscr{P}_n$, polynômes de degré $\leqslant n$ s'annulant en $x = \pm 1$.

Par exemple, avec une base de polynômes de Tchebycheff $T_k - T_{k-2}$, $k = 2, 3, \ldots, n$,

$$0 = \int_{-1}^{1} (Lu_h - 1)(T_p(x) - T_{p-2}(x))dx = \int_{-1}^{1} \{\sum_{2}^{n} c_k[(T'_k(x) - T'_{k-2}(x))(T'_p(x) - T'_{p-2}(x)) + (T_k(x) - T_{k-2}(x))(T_p(x) -$$

$p = 2, 3, \ldots, n$. Probablement assez compliqué.

Autre base polynomiale : nous allons tenter une résolution approchée à l'aide de polynômes de Legendre, fonctions propres d'un tout autre opérateur. En fait, on prendra les ***primitives*** des polynômes de Legendre :
$\varphi_n(x) = \displaystyle\int_{-1}^{x} P_n(t)\, dt = \frac{P_{n+1}(x) - P_{n-1}(x)}{2n+1}$, $n = 1, 2, \ldots, N$. Ces polynômes sont effectivement nuls en -1 et en 1, et leurs dérivées ont une expression commode. Posons donc $u_h = \sum_1^N \alpha_n \varphi_n \in U_h = (1 - x^2)\mathscr{P}_{N-1}$ et résolvons par Ritz-Galerkin :

$$0 = \int_{-1}^{1} (-u'' + u - 1)v\, dx = \int_{-1}^{1} (u'v' + uv - v)\, dx \quad \forall v \in U_h,$$

$$\int_{-1}^{1} P_j \sum_{1}^{N} \alpha_n P_n \; + \frac{P_{j+1} - P_{j-1}}{2j+1} \sum_{1}^{N} \alpha_n \frac{P_{n+1} - P_{n-1}}{2n+1} \; - \frac{P_{j+1} - P_{j-1}}{2j+1} = 0, j = 1, \ldots, N$$

d'où, en utilisant les propriétés d'orthogonalité des polynômes de Legendre :

$$-\frac{\alpha_{j-2}}{(2j-3)(2j-1)} + \left(1 + \frac{2}{(2j+3)(2j-1)}\right)\alpha_j - \frac{\alpha_{j+2}}{(2j+3)(2j+5)} = -\delta_{j,1}, \quad j = 1, 2, \ldots, N$$

ce qui fait bien N équations pour N inconnues ($\alpha_{-1} = \alpha_{N+2} = 0$).

```
%spectr1.m
%  -u''+u=1  par prim. de Legendre
diary
N=input('entrez N  ');N=2*ceil(N/2)+1
% Solution du systeme tridiagonal
a(N+2)=0;a(N)=1;
for j=N:-2:3;
  tg=-1/((2*j-3)*(2*j-1));tc=1+2/((2*j+3)*(2*j-1));td=-1/((2*j+3)*(2*j+5));
  a(j-2)=-( tc*a(j)+td*a(j+2) )/tg;
end;
uu=-1.4*a(1)+a(3)/35;
for j=1:2:N;a(j)=a(j)/uu;end;
format short e;a(1:2:N)
% valeur en x=0
% phi_(2n+1)(0)=(-1)^n (2n-1)...(-1)/((2n+2)...2)
phi=-0.5;va=a(1)*phi;
for j=3:2:N;phi=-phi*(j-2)/(j+1);va=va+a(j)*phi;end;
format long;[1-1/cosh(1)  ;  va]
```

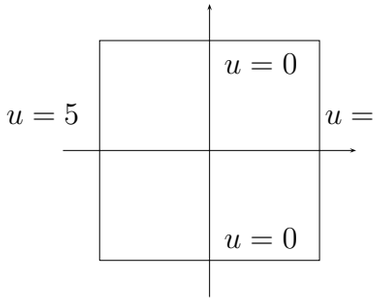
```
N=input('entrez N  ');N=2*ceil(N/2)+1
N =         3

format short e;a(1:2:N)
ans =
-7.1522E-001  -4.5652E-002

format long;[1-1/cosh(1)  ;  va]
ans =
       0.351945726336115
       0.35190

N=input('entrez N  ');N=2*ceil(N/2)+1
N =         7

format short e;a(1:2:N)
ans =
-7.1522E-001  -4.5659E-002  -7.1259E-004  -4.9385E-006

format long;[1-1/cosh(1)  ;  va]
ans =
       0.351945726336115
       0.3519457258

N=input('entrez N  ');N=2*ceil(N/2)+1
N =         11

format short e;a(1:2:N)
ans =
-7.1522E-001  -4.5659E-002  -7.1259E-004  -4.9385E-006  -1.9259E-008  -4.8085E-011

format long;[1-1/cosh(1)  ;  va]
ans =
       0.351945726336115
       0.351945726336113
```

On a donc trouvé une base permettant d'écrire économiquement des dérivées $(\sum \alpha_k \varphi_k)' = \sum \alpha_k P_k$ et des multiplications de u_h et u'_h par φ_j et φ'_j.

On constate aussi une décroissance rapide (exponentielle) des coefficients. C'est alors que les méthodes quasi-spectrales montrent leur efficacité !

2.2. Un problème de Trefethen.

http ://www.siam.org/siamnews/01-02/challenge.pdf ; Chastened Challenge Sponsor : "I Misjudged" (The $100, 100-digit `http ://www.siam.org/siamnews/07-02/challengeupdate.pdf` ; `http ://www.ma.tum.de/m3/bornemann/challengebook`



$u(x, y, t)$ vaut 0 sur 3 côtés du carré $[-1, 1] \times [-1, 1]$ et vaut 5 sur le $4^{\text{ème}}$ côté. Au temps $t = 0$, u est nul dans l'intérieur du carré. Enfin,

$$\frac{\partial u}{\partial t} = \Delta u. \tag{80}$$

Quand $u(0, 0, t)$ vaudra-t-elle 1 ? Donnez au moins dix décimales de cette valeur de t !

Mais $u(0, 0, t)$ atteint-elle seulement cette valeur ?

$\{\sin(k\pi(1 - y)/2)\}_1^\infty$ est une suite orthonormale totale de $L^2(-1, 1)$ vérifiant les conditions aux frontières en $y = \pm 1$, donc

$$u(x, y, t) = \sum_{k=1}^{\infty} c_k(x, t) \sin\left(k\frac{\pi(1 - y)}{2}\right).$$

(80) devient

$$\frac{\partial c_k(x, t)}{\partial t} = \frac{\partial^2 c_k(x, t)}{\partial x^2} - \frac{k^2\pi^2}{4} c_k(x, t), \qquad k = 1, 2, \ldots \tag{81}$$

En $t = \infty$, la dérivée en t est nulle, $c_k(x, \infty)$ est une combinaison de $\exp(\pm k\pi x/2)$ devant s'annuler en $x = 1$, soit $c_k(x, \infty) = c_k(-1, \infty)\dfrac{\sinh(k\pi(1 - x)/2)}{\sinh(k\pi)}$.

Enfin, $u = 5$ en $x = -1 \Rightarrow$

$$c_k(-1, \infty) = \int_{-1}^{1} 5 \ \sin\left(\frac{k\pi(1 - y)}{2}\right) \ dy = 5\frac{1 - \cos k\pi}{k\pi/2},$$

soit $20/(k\pi)$ si k est impair, 0 si k est pair.

On trouve alors

$$u(0, 0, \infty) = \sum_{k \text{ impair}} \frac{10(-1)^{(k-1)/2}}{k\pi \cosh(k\pi/2)}$$

qui se trouve valoir 1.25[1].

On reprend maintenant $t < \infty$: $c_k(x, t) - c_k(x, \infty)$ est une fonction s'annulant en $x = \pm 1$, soit

$$c_k(x, t) = c_k(x, \infty) - \sum_{m=1}^{\infty} d_{k,m}(t) \sin\left(m\frac{\pi(1 - x)}{2}\right)$$

avec $d_{k,m}(0)$ tels que $c_k(x, 0) = 0$, donc

$$\begin{aligned} d_{k,m}(0) = \int_{-1}^{1} c_k(x, \infty) \sin\left(m\frac{\pi(1 - x)}{2}\right) \ dx &= 2c_k(-1, \infty) \int_0^1 \frac{\sinh(k\pi t) \sin(m\pi t)}{\sinh(k\pi)} \ dt \\ &= \frac{2(-1)^{m-1} m \, c_k(-1, \infty)}{\pi(k^2 + m^2)}. \end{aligned}$$

[1] En effet, Green $\Rightarrow u(0, 0, \infty) = \dfrac{1}{2\pi} \displaystyle\int_{\partial D} u \frac{\partial G}{\partial n} \, ds$, où G est la fonction harmonique dans $D \setminus \{(0, 0)\}$, nulle au bord ∂D, et avec $G - \log r$ bornée. Alors, si D est un polygone régulier, l'intégrale de $\partial G/\partial n$ est la même sur chaque côté et donc, si u est constante sur chaque côté, la valeur de u au centre est la moyenne des valeurs sur les côtés. □

Enfin (81) $\Rightarrow dd_{k,m}(t)/dt = -\pi^2(k^2+m^2)d_{k,m}(t)/4$:

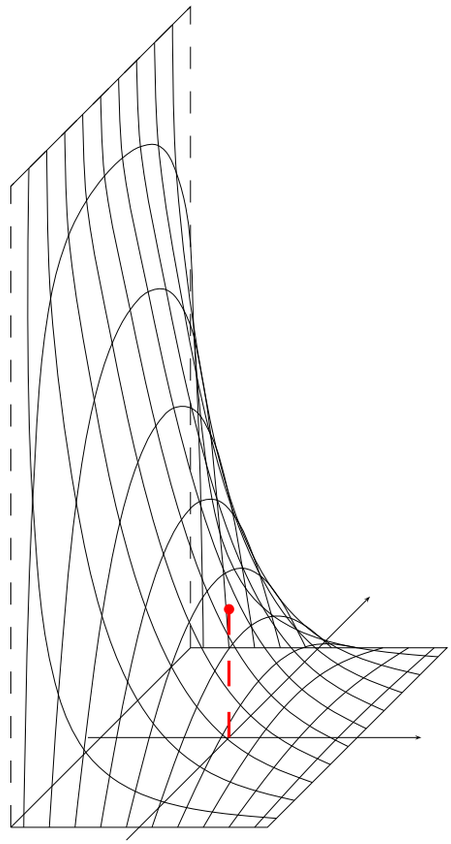
$$u(x,y,t) =$$

$$\frac{20}{\pi^2}\sum_{k \text{ impair}}\left[\frac{\sinh(k\pi(1-x)/2)}{\sinh(k\pi)} - \sum_{m=1}^{\infty}\frac{2m(-1)^{m-1}}{k^2+m^2}e^{-\pi^2(k^2+m^2)t/4}\sin\left(m\pi\frac{1-x}{2}\right)\right]\frac{\sin\left(k\pi\frac{1-y}{2}\right)}{k}.$$

En $(x,y) = (0,0)$:

t	$u(0,0,t)$
0.10000000000000000000	0.123524160473132733516
0.20000000000000000000	0.504418477389377249941
0.30000000000000000000	0.789736409260429938346
0.40000000000000000000	0.968577062427970117126
0.42401138703368836379 7	1.000000000000000000
0.50000000000000000000	1.078155121939073902 78
0.60000000000000000000	1.145085926180581002 12
0.70000000000000000000	1.185949900931502701 33
0.80000000000000000000	1.210897518596871706 49
0.90000000000000000000	1.226128010420449432 29
1.00000000000000000000	1.235426197341393633 52

Graphe de u en cette valeur de t où $u(0,0,t)$ vaut 1 :



2.3. Méthode des tau.

Si on introduit une somme finie $\sum_0^{N-1} \alpha_k \varphi_k$ dans une identité valable pour la série $\sum_0^{\infty} \alpha_k \varphi_k$, on observera probablement des termes résiduels de la forme

$$\tau_0 \varphi_N + \tau_1 \varphi_{N+1} + \cdots$$

On appelle ***méthode des tau*** (Lanczos) une façon de construire l'approximation $\sum_0^{N-1} \alpha_k \varphi_k$ tirant parti de cette forme du résidu $Au_h - f$.

On n'exige généralement pas en méthode des tau que chaque φ_k vérifie les conditions (essentielles) aux limites, on prend donc un développement souvent plus simple que pour Ritz-Galerkin.

Prenons le même exemple $-u'' + u = 1$ avec les polynômes de Tchebycheff $T_0 = 1, T_1(x) = x, T_2(x) = 2x^2 - 1, T_3(x) = 4x^3 - 3x, T_4(x) = 8x^4 - 8x^2 + 1, \ldots$, plus précisément, leurs primitives itérées deux fois :

$$\varphi_2(x) = \frac{x^2}{2} = \frac{T_2 + T_0}{4}, \varphi_3(x) = \frac{x^3}{6} = \frac{T_3 + 3T_1}{24}, \varphi_4(x) = \frac{x^4}{6} - \frac{x^2}{2} = \frac{T_4 - 8T_2 - 9}{48},$$

$$\varphi_{k+2} := \iint T_k = \int \left[\frac{T_{k+1}}{2(k+1)} - \frac{T_{k-1}}{2(k-1)} \right] = \frac{T_{k+2}}{4(k+1)(k+2)} - \frac{T_k}{2(k-1)(k+1)} + \frac{T_{k-2}}{4(k-1)(k-2)}, k > 2,$$

et $\varphi_0 = 1$, $\varphi_1(x) = x$. Remarquons que $\varphi_{k+2}(\pm 1) = \dfrac{3(-1)^k}{(k^2-4)(k^2-1)}$, $k > 2$.

On porte $-\left(\sum_0^{N-1} \alpha_k \varphi_k(x) \right)'' + \sum_0^{N-1} \alpha_k \varphi_k(x) = 1 + \tau_0 \varphi_{N-2} + \tau_1 \varphi_{N-1}$.

L'identification des coefficients de $\varphi_0, \ldots, \varphi_{N-1}$ (en fait, $T_0, \ldots, T_{N-1}$) représente N équations pour les $N+2$ inconnues $\alpha_0, \ldots, \alpha_{N-1}$ et τ_0 et τ_1. Les équations manquantes sont fournies par les conditions aux limites.

Ici :

$T_0 : -\alpha_2 + \alpha_0 + \alpha_2/4 - 3\alpha_4/16 = 1,$

$T_1 : -\alpha_3 + \alpha_1 + 3\alpha_3/4 + \alpha_5/8 = 0,$

$T_k : -\alpha_{k+2} + \dfrac{\alpha_k}{4(k-1)k} - \dfrac{\alpha_{k+2}}{2(k-1)(k+1)} + \dfrac{\alpha_{k+4}}{4k(k+1)} = 0,$

pour $k = 2, 3, \ldots, N-3$ (où $\alpha_{N+1} = 0$). Enfin, les valeurs en ± 1 sont

$$\alpha_0 + \frac{\alpha_2}{2} - \alpha_4 + \sum_{\substack{k=4 \\ k \text{ pair}}}^{N-3} \frac{3\alpha_{k+2}}{(k^2-4)(k^2-1)} \pm \left[\alpha_1 + \frac{\alpha_3}{6} + \sum_{\substack{k=3 \\ k \text{ impair}}}^{N-3} \frac{3\alpha_{k+2}}{(k^2-4)(k^2-1)} \right].$$

3. Calcul en représentation ponctuelle.

On décide de travailler avec le vecteur $[u_h(x_1), \ldots, u_h(x_N)]$ des valeurs de u_h en N points (***points sélectionnés***) de $\overline{\Omega}$. Toute opération sur u_h devra se faire en accord avec l'espace des approximations :

$$u_h = \text{ interpolant de } [u_h(x_1), \ldots, u_h(x_N)] \text{ dans } U_h.$$

Il est souvent précieux de passer de la représentation spectrale à la représentation ponctuelle et, surtout, *vice-versa*, c'est-à-dire pouvoir réaliser l'interpolant. Le cas le plus intéressant est celui où les fonctions φ_k, déjà orthogonales selon un premier produit scalaire, sont aussi orthogonales selon un produit scalaire discret

$$(f, g)_N = \sum_1^N w_m f(x_m) g(x_m).$$

En effet, si $(\varphi_i, \varphi_j)_N = \delta_{i,j}$, $0 \leqslant i, j \leqslant N-1$, on a immédiatement le coefficient $\alpha_k = (u_h, \varphi_k)_N$, donc

$$u_h(x) = \sum_{k=0}^{N-1} \alpha_k \varphi_k(x) = \sum_{k=0}^{N-1} \left(\sum_{m=1}^{N} u_h(x_m) w_m \varphi_k(x_m) \right) \varphi_k(x).$$

Ainsi, avec les ***polynômes trigonométriques***, la base $\cos(2k\pi x/\ell)$, $k = 0, 1, \ldots, N/2$, $\sin(2k\pi x/\ell)$, $k = 1, \ldots, N/2-1$, est encore orthogonale pour $(f, g)_N = \sum_{m=0}^{N-1} f(\ell m/N) g(\ell m/N)$. Le passage d'une représentation à l'autre est d'ailleurs fortement accéléré par la *transformée de Fourier rapide* (Fast Fourier Transform, **FFT**).

Les ***polynômes orthogonaux*** $p_0, \ldots, p_{N-1}$ sont encore orthogonaux pour le produit scalaire discret issu de la ***formule d'intégration de Gauss*** à N points (les zéros de p_N)[2].

3.1. Méthode de collocation.

Si on peut parfaitement résoudre un problème de Ritz-Galerkin dans la représentation ponctuelle, on choisira souvent cette représentation pour approcher la solution de $Au = f$ par

$$\text{interpolation de } (Au_h - f) = \text{interpolation de } \{A(\text{ interpolant de } [u_h(x_1), \ldots, u_h(x_N)] \text{ dans } U_h) - f\} = 0$$

(***méthode de collocation***, ou méthode de points sélectionnés.

On peut encore considérer la méthode de collocation comme une méthode de Galerkin (projection) si on dispose d'un produit scalaire discret approprié :

$$Au_h - f = 0 \text{ aux zéros de } \varphi_N \iff (Au_h - f, v)_N = 0, \ \forall v \in U_h,$$

puisque le produit scalaire $(\ ,\ )_N$ ne fait appel qu'aux valeurs où $\varphi_N = 0$.

3.2. Représentation matricielle des opérateurs différentiels.

Si $\{\ell_1, \ldots, \ell_N\}$ est la base de Lagrange appropriée au schéma d'approximation retenu, on applique des opérateurs différentiels à l'interpolation de u

$$u_h(x) = \sum_{1}^{N} u_h(x_j)\ \ell_j(x).$$

Ainsi, la dérivée première

$$\frac{d}{dx} u_h(x) = \sum_{1}^{N} u_h(x_j)\ \ell_j'(x)$$

que nous évaluons en x_i, $i = 1, \ldots, N$, ce qui met en évidence une matrice

$$D = \begin{bmatrix} \ell_1'(x_1) & \ell_2'(x_1) & \cdots & \ell_N'(x_1) \\ \ell_1'(x_2) & \ell_2'(x_2) & \cdots & \ell_N'(x_2) \\ \vdots & \vdots & \cdots & \vdots \\ \ell_1'(x_N) & \ell_2'(x_N) & \cdots & \ell_N'(x_N) \end{bmatrix}.$$

L'application de l'opérateur $L(d/dx)$ à u_h se représente par le produit de $L(D)$ et du vecteur des $u_h(x_j)$.

[2] On utilise également des formules de ***Radau*** et de ***Lobatto***.

Exemple des polynômes de degré $\leqslant N$ interpolés en $x_0, \ldots, x_N$:

$$\ell_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{N} \frac{x - x_i}{x_j - x_i}.$$

On effectue progressivement

$$\ell_{j,\text{new}}(x) = \ell_{j,\text{old}}(x) \frac{x - x_i}{x_j - x_i}\,, \qquad \ell'_{j,\text{new}}(x) = \ell'_{j,\text{old}}(x) \frac{x - x_i}{x_j - x_i} + \ell_{j,\text{old}}(x) \frac{1}{x_j - x_i}$$

etc.

```
c-----------------------------------------------------------------------
      subroutine weights1(xi,x,n,m,c,ndim)
c
c  Cf. B. Fornberg, A Practical Guide to Pseudospectral Methods,
c   Cambridge University Press, 1996,  paperback: 1999.
c
c  B. Fornberg, Calculation of weights in finite difference
c     formulas, SIAM Rev. vol. 40 no 3 , 685-691, Sept. 1998.
c
c+---------------------------------------------------------------------+
c|  INPUT PARAMETERS:                                                  |
c|    XI  POINT AT WHICH THE APPROXIMATIONS ARE TO BE ACCURATE         |
c|    X   X-COORDINATES FOR GRID POINTS, ARRAY DIMENSIONED X(0:N)      |
c|    N   THE GRID POINTS ARE AT X(0),X(1),...,X(N) (I.E. N+1 IN ALL)  |
c|    M   HIGHEST ORDER OF DERIVATIVE TO BE APPROXIMATED               |
c|    ndim    see dimensions of c                                      |
c|                                                                     |
c| OUTPUT PARAMETER:                                                   |
c|    C   WEIGHTS, ARRAY DIMENSIONED C(0:Ndim,0:M).                      |
c|        ON RETURN, THE ELEMENT C(J,K) CONTAINS THE WEIGHT TO BE      |
c|        APPLIED AT X(J) WHEN THE K:TH DERIVATIVE IS APPROXIMATED     |
c|        BY A STENCIL EXTENDING OVER X(0),X(1),...,X(N).              |
c+---------------------------------------------------------------------+
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION X(0:N),C(0:Ndim,0:M)
      C1=1
      C4=X(0)-XI
      DO 10 K=0,M
         DO 10 J=0,N
   10       C(J,K)=0
      C(0,0)=1
      DO 50 I=1,N
         MN=MIN(I,M)
         C2=1
         C5=C4
         C4=X(I)-XI
         DO 40 J=0,I-1
            C3=X(I)-X(J)
            C2=C2*C3
         DO 20 K=MN,1,-1
   20       C(I,K)=C1*(K*C(I-1,K-1)-C5*C(I-1,K))/C2
         C(I,0)   =-C1*C5*C(I-1,0)/C2
            DO 30 K=MN,1,-1
```

```
   30          C(J,K)= (C4*C(J,K)-K*C(J,K-1))/C3
   40       C(J,0)=C4*C(J,0)/C3
   50    C1=C2
      RETURN
      END
c-------------------------------------------------------------
```

4. Exemples de conditions de stabilité numérique :

Coercivité *selon le produit scalaire discret* : $(Au,u)_N \geqslant c\|u_h\|^2$, $\forall u_h \in U_h \Rightarrow \|u_h\| \leqslant \|f\|/c$. (Prendre $v = u_h$).

Condition inf-sup (Babuška). Dans des cas où on ne peut établir la coercivité discrète, on peut parfois arriver à

$$\inf_{u\in U_h} \sup_{v\in U_h} \frac{(Au,v)_N}{\|u\|\,\|v\|} \geqslant c > 0.$$

On a alors $v \in U_h$ tel que $(Au_h, v)_N : \|v\| \geqslant c\|u_h\|$, donc $\|f\| \geqslant (f,v)_N/\|v\| \geqslant c\|u_h\|$, d'où encore $\|u_h\| \leqslant \|f\|/c$.

Fin de MATH2180

$\mathscr{T}$hat's all, $\mathscr{F}$olks

Index

bilinéaire, 38
Black-Scholes, 161
Burgers, 20

caractéristique, 20, 22
CFL, 175
chaleur, éq. de la, 24, 151
coercive, 39, 100
condition inf-sup, 54, 197
conditionnement, 110
consistance, 159
crimes variationnels, 88, 104

définie positive, 38, 43
divergence, théorème de la, 31, 34

EDP, 20
Elément fini, définition, 58
Elément triangulaire linéaire, 76, 82
Eléments unidimensionnels, 61
elliptique, 23
elliptique, prob., 39
essentielle, condition, 40
euros, 18, 152
evolution, problème d', 157

FAQ, 9
flambage, 41
fonction de forme, 61
Fourier, 155
frontière lipschitzienne, 97

Galerkin, 10, 37, 42, 53, 54
Gauss-Green, 31, 34
Gershgorin, 111–113, 181
Green, fonction de, 33, 144

Hadamard, 111
Hermite cubique 1-D, 63
Hilbert, espace de, 46
hyperbolique, 23, 168

inf-sup, 54, 197
inf-sup, conditions, 54
intégrales, équations, 143

Kreiss, théorème de, 183

Laplace, éq. de, 24, 25, 124
Lax-Milgram, th. de, 48
leapfrog, 169

masse, matrice de, 112
matrice stochastique, 141
membrane, 21
morceaux, f. dérivables par, 32
multigrilles, 128
multipolaire, 144

naturelle, condition, 40

parabolique, 23, 160
Petrov, 10, 54
Poisson, éq. de, 25
problème de Cauchy, 20, 21
programmes, 9

rigidité, matrice de, 43
Ritz, 37, 42
Robin, 52, 126

Sobolev, espace de, 90
Sobolev, lemme de, 93
Sobolev, produit scalaire de, 89
stabilité numérique, 159, 180
stochastique, matrice, 141

telégraphistes, éq. des, 23
trace, 97

unisolvant, ensemble, 59

variationnelle, formulation, 33, 34, 40
variations, calcul des, 29
von Neumann, cond. de, 183