# Analyse numérique de base

Sophie et Rémi Abgrall

18 août 2005

# Table des matières

## Avant propos

Ce cours a été réalisé à l'Université de Nouvelle Calédonie, Nouméa, en août 2005. Nous souhaiterions remercier le Professeur H. Bonnel pour son invitation, mais aussi les étudiants, motivés et agréables.

Ce cours doit beaucoup aux notes que Pierre Charrier et Denise Aregba nous ont donnés. Ils sont respectivement Professeur et Maître de Conférence à l'Université Bordeaux I. Ce cours a beaucoup emprunté aussi à divers ouvrages

- Algèbre linéaire : les références [1], [2], [3] et [4]
- Interpolation et quadratures numériques : [5], [6], [7].

Faure de temps, nous n'avons pas pu traiter l'approximation d'équations différentielles ordinaires. On peut trouver les informations nécessaires dans [5], [6].

# Chapitre 1

# Introduction

Le but du cours est de donner quelques éléments de base d'analyse numérique. Qu'est ce que l'analyse numérique ? C'est la science, l'art aussi, qui permet, pour un problème «continu» donné de construire et analyser un algorithme, c'est à dire une méthode constructive, réalisable en temps fini, qui fournit une solution approchée du problème qu'on s'est donné. Il est donc nécessaire de connaître les propriétés de la méthode : coût (nombre d'opérations nécessaires pour une machine donnée), stabilité (sensibilité aux erreurs d'arrondis par exemple, mais aussi respect ou non respect des bornes *a priori* du problème continu), convergence (quelle est la qualité de l'approximation), etc.

Ce document n'est pas exempt de coquilles, il y en a certainement beaucoup. Nous nous en excusons par avance en espérant qu'elles ne perturberont pas trop le lecteur.

**Exemple de la météorologie.** Le physicien/mécanicien écrit les modèles permettant de décrire les phénomènes physiques «intéressants» : force et direction du vent, humidité, nuages, intéraction atmosphère-océan, rôle de la topographie, des zones vertes, etc). On écrit ainsi un système d'équations aux dérivées partielles qui relie les différentes variables du modèle (température, pression, vitesse, salinité, etc). Ces modèles n'ont en général pas de solution analytique calculable en pratique.

Le mathématicien s'occupe de l'analyse des équations (problème bien posé, existence, unicité, comportement des solutions, etc) et de l'analyse numérique du problème (définition d'un algorithme, analyse de cet algorithme (existence de solution, stabilité, ...), implémentation en machine, expolitation des résultats (visualisation, ...).

Il s'agit d'un art dans le sens où il est très rare que les modèles et les algorithmes, en raison de leur complexité, soient complètement analysable de manière rigoureuse. On est donc conduit à raisonner par analogie, utiliser une certaine intuition des algorithmes et des modèles afin d'être capable de donner suffisament d'éléments permettant de conduire à un jugement raisonné de la méthode.

**Un exemple très simple.** Considérons le modèle suivant qui consiste à trouver une fonction $t \mapsto y(t)$ telle que

$$\begin{aligned} y' &= y \qquad t \in [0,T] \\ y(0) &= 1. \end{aligned}$$

On sait que la solution est $y(t) = e^t$. Plutôt que de raisonner sur $[0, T]$ tout entier (un continuum de nombres réels), on va considérer un ensemble discret d'instants, $0 = t_0 < t_1 < \ldots < t_k < \ldots < t_N = T$ où $t_n = n\Delta t$ et $\Delta t = \frac{T}{N}$. Ainsi, plutôt que de considérer la fonction $y$ (un continuum de valeurs), on va considérer un nombre fini de leurs valeurs, ou plutôt une approximation de ceux–ci, à savoir $y_n \simeq y(t_n)$, $n = 0, \ldots, N$.

L'algorithme est

$$y_{n+1} = y_n + \Delta t y_n \qquad \left( = (1 + \Delta t) y_n \right).$$

Le but de l'analyse numérique est donc d'analyser cet algorithme (quelle est sa pertinence, comment approche–t'on la solution, quelle est la qualité de l'approximation, quel est le coût, est–ce stable, ect).

Il ne fait pas croire que c'est simple. On peut très facilement construire des algorithmes insensés, même en travaillant dur. Même si l'algorithme est parfait sur le papier, on peut avoir de graves problèmes une fois qu'on l'a implémenté en machine. Un exemple est donné par la choses suivante.

On sait bien que $1 = (10^n + 1) - 10^n$, quelque soit $n \in \mathbb{N}$. On veut le vérifier à l'aide du programme suivant

```
      program test
      implicit none
!
! but montrer le role des erreurs d'arrondi sur un exemple tres simple
!
       real temp, x, y
       integer i
        temp=1
      do i=1,35
      temp=10*temp
      x=temp+1
      y=temp
      print*, "n= ",i,x-y
      enddo
      end program test
```

Si on le compile avec le compilateur `g77`, on trouve

```
 n=  1  1.
 n=  2  1.
 n=  3  1.
 n=  4  1.
 n=  5  1.
 n=  6  1.
 n=  7  1.
 n=  8  0.
 n=  9  0.
```

```
n=  10  0.
n=  11  0.
n=  12  0.
n=  13  0.
n=  14  0.
n=  15  0.
n=  16  0.
```

Si on le compile avec le compilateur `ifort` d'INTEL, on a le même résultat. Si maintenant on le compile avec l'option `- r8`, on obtient

```
n=           1   1.00000000000000
n=           2   1.00000000000000
n=           3   1.00000000000000
n=           4   1.00000000000000
n=           5   1.00000000000000
n=           6   1.00000000000000
n=           7   1.00000000000000
n=           8   1.00000000000000
n=           9   1.00000000000000
n=          10   1.00000000000000
n=          11   1.00000000000000
n=          12   1.00000000000000
n=          13   1.00000000000000
n=          14   1.00000000000000
n=          15   1.00000000000000
n=          16  0.000000000000000E+000
n=          17  0.000000000000000E+000
n=          18  0.000000000000000E+000
n=          19  0.000000000000000E+000
n=          20  0.000000000000000E+000
n=          21  0.000000000000000E+000
n=          22  0.000000000000000E+000
n=          23  0.000000000000000E+000
```

On voit donc que si $n$ est assez grand, le calcul numérique est faux. On voit aussi que cela dépend de la précision demandée (simple, double précision). Par exemple, le même programme compilé avec l'option `-r16` (quadruple précision) donne le résultat déroutant

```
n=           1   1.0000000000000000000000000000000000
n=           2   1.0000000000000000000000000000000000
n=           3   1.0000000000000000000000000000000000
n=           4   1.0000000000000000000000000000000000
n=           5   1.0000000000000000000000000000000000
n=           6   1.0000000000000000000000000000000000
n=           7   1.0000000000000000000000000000000000
n=           8   1.0000000000000000000000000000000000
```

```
 n=            9   1.000000000000000000000000000000000
 n=           10   1.000000000000000000000000000000000
 n=           11   1.000000000000000000000000000000000
 n=           12   1.000000000000000000000000000000000
 n=           13   1.000000000000000000000000000000000
 n=           14   1.000000000000000000000000000000000
 n=           15   1.000000000000000000000000000000000
 n=           16   1.000000000000000000000000000000000
 n=           17   1.000000000000000000000000000000000
 n=           18   1.000000000000000000000000000000000
 n=           19   1.000000000000000000000000000000000
 n=           20   1.000000000000000000000000000000000
 n=           21   1.000000000000000000000000000000000
 n=           22   1.000000000000000000000000000000000
 n=           23   1.000000000000000000000000000000000
 n=           24   1.000000000000000000000000000000000
 n=           25  0.500000000000000000000000000000000
 n=           26  0.500000000000000000000000000000000
 n=           27  0.500000000000000000000000000000000
 n=           28  0.500000000000000000000000000000000
 n=           29   1.000000000000000000000000000000000
 n=           30   1.000000000000000000000000000000000
 n=           31   1.000000000000000000000000000000000
 n=           32   1.000000000000000000000000000000000
 n=           33   1.000000000000000000000000000000000
 n=           34   1.000000000000000000000000000000000
 n=           35  0.000000000000000000000000000000000E+0000
```

On a là affaire à un exemple typique d'erreur d'arrondi : si $n$ est trop grand, $10^n$ dépasse la valeur maximale du plus grand réel (cf déclarations du programme) représentable dans le mode choisi (simple, double, quadruple précision).

Si maintenant on considère le programme

```
      program test
      implicit none
!
! but montrer le role des erreurs d'arrondi sur un exemple tr\'es simple
!
       integer temp, x, y
       integer i
        temp=1
      do i=1,35
      temp=10*temp
      x=temp+1
      y=temp
      print*, "n= ",i,x-y
```

```
      enddo
      end program test
```

(une seule ligne de difference), on obtient en simple précision

```
n=           1           1
 n=          2           1
 n=          3           1
 n=          4           1
 n=          5           1
 n=          6           1
 n=          7           1
 n=          8           1
 n=          9           1
 n=         10           1
 n=         11           1
 n=         12           1
 n=         13           1
 n=         14           1
 n=         15           1
 n=         16           1
 n=         17           1
 n=         18           1
 n=         19           1
 n=         20           1
 n=         21           1
 n=         22           1
 n=         23           1
 n=         24           1
 n=         25           1
 n=         26           1
 n=         27           1
 n=         28           1
 n=         29           1
 n=         30           1
 n=         31           1
 n=         32           1
 n=         33           1
 n=         34           1
 n=         35           1
```

Mais il ne faut pas croire que le résultat est juste pour autant : si on imprime la valeur de temp, on a (quatrième colonne)

```
 n=          1           1          10
 n=          2           1         100
 n=          3           1        1000
```

```
n=           4           1        10000
n=           5           1       100000
n=           6           1      1000000
n=           7           1     10000000
n=           8           1    100000000
n=           9           1   1000000000
n=          10           1   1410065408
n=          11           1   1215752192
n=          12           1   -727379968
n=          13           1   1316134912
n=          14           1    276447232
n=          15           1 -1530494976
n=          16           1   1874919424
n=          17           1   1569325056
n=          18           1 -1486618624
n=          19           1 -1981284352
n=          20           1   1661992960
n=          21           1   -559939584
n=          22           1 -1304428544
n=          23           1   -159383552
n=          24           1 -1593835520
n=          25           1   1241513984
n=          26           1   -469762048
n=          27           1   -402653184
n=          28           1    268435456
n=          29           1 -1610612736
n=          30           1   1073741824
n=          31           1 -2147483648
n=          32           1            0
n=          33           1            0
n=          34           1            0
n=          35           1            0
```

en simple précision, avec l'option `-i8`, on a

```
n=                1                   1                  10
n=                2                   1                 100
n=                3                   1                1000
n=                4                   1               10000
n=                5                   1              100000
n=                6                   1             1000000
n=                7                   1            10000000
n=                8                   1           100000000
n=                9                   1          1000000000
n=               10                   1         10000000000
n=               11                   1        100000000000
```

| | | | |
|---|---|---|---|
| n= | 12 | 1 | 1000000000000 |
| n= | 13 | 1 | 10000000000000 |
| n= | 14 | 1 | 100000000000000 |
| n= | 15 | 1 | 1000000000000000 |
| n= | 16 | 1 | 10000000000000000 |
| n= | 17 | 1 | 100000000000000000 |
| n= | 18 | 1 | 1000000000000000000 |
| n= | 19 | 1 | -8446744073709551616 |
| n= | 20 | 1 | 7766279631452241920 |
| n= | 21 | 1 | 3875820019684212736 |
| n= | 22 | 1 | 1864712049423024128 |
| n= | 23 | 1 | 200376420520689664 |
| n= | 24 | 1 | 2003764205206896640 |
| n= | 25 | 1 | 1590897978359414784 |
| n= | 26 | 1 | -2537764290115403776 |
| n= | 27 | 1 | -6930898827444486144 |
| n= | 28 | 1 | 4477988020393345024 |
| n= | 29 | 1 | 7886392056514347008 |
| n= | 30 | 1 | 5076944270305263616 |
| n= | 31 | 1 | -4570789518076018688 |
| n= | 32 | 1 | -8814407033341083648 |
| n= | 33 | 1 | 4089650035136921600 |
| n= | 34 | 1 | 4003012203950112768 |
| n= | 35 | 1 | 3136633892082024448 |
| n= | 36 | 1 | -5527149226598858752 |
| n= | 37 | 1 | 68739955140067328 |
| n= | 38 | 1 | 687399551400673280 |
| n= | 39 | 1 | 6873995514006732800 |
| n= | 40 | 1 | -5047021154770878464 |
| n= | 41 | 1 | 4870020673419870208 |
| n= | 42 | 1 | -6640025486929952768 |
| n= | 43 | 1 | 7386721425538678784 |
| n= | 44 | 1 | 80237960548581376 |
| n= | 45 | 1 | 802379605485813760 |
| n= | 46 | 1 | 8023796054858137600 |
| n= | 47 | 1 | 6450984253743169536 |
| n= | 48 | 1 | 9169610316303040512 |
| n= | 49 | 1 | -537617205517352960 |
| n= | 50 | 1 | -5376172055173529600 |
| n= | 51 | 1 | 1578511669393358848 |
| n= | 52 | 1 | -2661627379775963136 |
| n= | 53 | 1 | -8169529724050079744 |
| n= | 54 | 1 | -7908320945662590976 |
| n= | 55 | 1 | -5296233161787703296 |
| n= | 56 | 1 | 2377900603251621888 |

```
n=                    57                    1    5332261958806667264
n=                    58                    1   -2017612633061982208
n=                    59                    1   -1729382256910270464
n=                    60                    1    1152921504606846976
n=                    61                    1   -6917529027641081856
n=                    62                    1    4611686018427387904
n=                    63                    1   -9223372036854775808
n=                    64                    1                      0
n=                    65                    1                      0
n=                    66                    1                      0
n=                    67                    1                      0
n=                    68                    1                      0
n=                    69                    1                      0
```

Tout ce qu'on fait est de reculer les échéances.

Cet exemple illuster le problème de l'overflow. En TD, on verra des exemples d'underflow.

# Chapitre 2

# Rappels et compléments d'algèbre linéaire

Dans ce chapitre, on donne quelques rappels d'algèbre linéaire (espaces vectoriel, matrices, normes) et quelques compléments (conditionnement).

## 2.1 Algèbre linéaire de base

Dans tout le cours, $\mathbb{K}$ désigne un corps commutatif qui sera $\mathbb{R}$ ou $\mathbb{C}$, et dans la pratique sera toujours $\mathbb{R}$.

**Définition 2.1.1 (Espace vectoriel sur $\mathbb{K}$).** *Un ensemble $E$ muni d'une loi interne $+$ et d'une loi externe $.$ est un espace vectoriel sur $\mathbb{K}$ si*

1. *quelques soient $x, y \in E$, $x + y \in E$ : loi interne,*
2. *$+$ est associatif et commutatif,*
3. *$+$ admet un élément neutre noté $0$ : quelque soit $x \in R$, $x + 0 = x$.*
4. *Tout élément admet un opposé pour la loi $+$ : quelque soit $x \in E$, il existe $y \in E$ tel que $x + y = 0$. On note l'opposé par $-x$,*
5. *Quels que soient $x, y \in R$, $\lambda, \mu \in \mathbb{K}$, on a*
   *(a)* $(\lambda + \mu) \,.\, x = \lambda \,.\, x + \mu \,.\, y$,
   *(b)* $\lambda(x + y) = \lambda \,.\, x + \mu \,.\, y$,
   *(c)* $\lambda \,.\, (\mu \,.\, x) = (\lambda\mu) \,.\, x$

On a la notion de sous–espace vectoriel.

**Définition 2.1.2 (Famille génératrice–Famille libre).** *On dit que $F = \{x_1, \ldots, x_n\}$*
- *est une famille génératrice de $E$ si le sous–espace engendré par $F$ est égal à $E$.*
- *est une famille libre si toute combinaison linéaire nulle formée d'éléments de $F$ est nécessairement à cœfficients nuls*

$$\sum_{\xi \in G \subset F} \alpha_\xi \xi = 0 \qquad \textit{implique } \alpha_\xi = 0 \textit{ quelque soit } \xi \in G.$$

Ensuite,

**Définition 2.1.3 (Base).** *On dit que $B$ est une base de $E$ si c'est une famille génératrice libre.*

Remarquons que toutes les sommes sont *finies*.

On dit qu'un espace vectoriel est de dimension finie s'il admet une famille génératrice finie. On a le résultat suivant

**Théorème 2.1.1.** *Soit $E$ un espace vectoriel différent de $\{0\}$. Il admet une base. SiE est de dimension finie, toute base admet le même cardinal qu'on appelle la* dimension *de $E$.*

On définit ensuite la somme de deux sous–espaces vectoriels $F$ et $G$ de $E$ comme étant l'espace engendré par $F \cup G$. On dit que $F$ et $G$ sont en somme directe si $F + G = E$ et $F \cap G = \{0\}$. Il est alors clair que $\dim E = \dim F + \dim G$. Enfin, on montre que tout sous–espace $F$ de $E$ admet un supplémentaire $H$.

**Exemple 2.1.4.** *L'espace $\mathbb{R}^N$ admet pour base la base canonique. Il est donc de dimension $N$. Un supplémentaire de $\mathbb{R}^p \times \{0\}^{N-p}$ est par exemple $\{0\}^n \times \mathbb{R}^{N-p}$.*

## 2.2 Applications linéaires.

**Définition 2.2.1.** *Soit $f : E \to F$ une application entre deux espaces vectoriels. On dit que c'est une application linéaire si quelques soit $x$, $y \in E$ et $\lambda \in \mathbb{K}$, $f(x+y) = f(x)+f(y)$ et $f(\lambda x) = \lambda f(x)$.*

**Exemple 2.2.2.** *Des exemples sont l'application identité, les applications $p$ et $s$ définies sur $\mathbb{R}^2$ par $p(x, y) = x$ et $s(x, y) = (y, x)$.*

On munit les applications linéaires de l'addition et du produit par un scalaire ce qui fait de $\mathcal{L}(E, F)$ un espace vectoriel de dimension $\dim E \times \dim F$. Une bijection entre deux espaces vectoriels est un isomorphisme. Enfin, on a la composition des applications qui est stable.

On définit le noyau et l'image d'une application linéaire $f : E \to G$

$$\mathrm{Ker} f = \{x \in E, f(x) = 0\}$$

et

$$\mathrm{Im} f = \{y \in F, \text{ il existe } x \in E, y = f(x)\}.$$

Clairement un endomorphisme est injectif si $\mathrm{Ker} f = \{0\}$, il est surjectif si $\mathrm{Im} f = F$. C'est un isomorphisme s'il est injectif et surjectif.

Si $B = \{e_1, \ldots, e_n\}$ est une base de $E$, Im $f$ est engendré par $\{f(e_1), \ldots, f(e_n)\}$. Enfin, on a le théorème du rang :

**Théorème 2.2.1.** *Si $f : E \to F$ est une application linéaire et si $E$ est de dimension finie, on a*

$$dim\ E = dim\ Ker\ f + dim\ Im\ f.$$

## 2.3 Matrices

Une façon de les définir est de le faire en utilisant la notion d'application linéaire. Soit $f : E \to F$ une application linéaire, $B_E = \{e_1, \ldots, e_n\}$ une base de $E$ et $B_F = \{f_1, \ldots, f_p\}$ une base de $F$. On sait que

$$f(e_j) = \sum_{l=1}^{p} a_{lj} f_l.$$

On définit la matrice de $f$ relativement aux bases $B_E$ et $B_F$ par le tableau

$$M(f, B_E, B_F) = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \ldots & a_{pn} \end{pmatrix}.$$

La première colonne donne les composantes de $f(e_1)$, etc. Si

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$$

, on a

$$\begin{aligned} f(x) &= \sum_{l=1}^{n} x_l f(e_l) = \sum_{l=1}^{n} x_l \Big\{ \sum_{m=1}^{p} a_{ml} f_m \Big\} \\ &= \sum_{m=1}^{p} \Big\{ \sum_{l=1}^{n} a_{ml} x_l \Big\} f_m \end{aligned}$$

et donc matriciellememnt

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \ldots & a_{pn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \sum_{l=1}^{n} a_{1l} x_l \\ \vdots \\ \sum_{l=1}^{n} a_{kl} x_l \end{pmatrix}$$

En transposant les opérations $+$ et $.$ des opérateurs linéaires, on définit une structure d'espace vectoriel. Enfin, le produit est défini comme la transposition matricielle de la composition des applications : soient $f : E \to F$ et $g : F \to G$ deux endomorphismes. $g \circ f : E \to G$ et donc (en utilisant des bases)

$$M(g \circ f, B_E, B_G) = M(g, B_F, B_E) M(f, B_E, B_G).$$

Posons $M(f, B_E, B_F) = (a_{ij})_{1 \le n, 1 \le m}$ et $M(g, B_E, B_F) = (b_{ij})_{1 \le i \le p, 1 \le j \le m}$ En faisant les calculs, on a si $M(g \circ f, B_E, B_G) = (c_{ij})_{1 \le i \le n, 1 \le j \le p}$ avec

$$c_{ij} = \sum_{l=1}^{p} b_{il} a_{lj}.$$

Pour voir cela, il suffit d'évaluer $g \circ f$ sur les éléments de la base $B_E$.

On montre que le produit est associatif (car la composition l'est).

**Remarque 2.3.1.** *En faisant ce calcul, on n'utilise que la structure algebrique de $\mathbb{K}$, et en fait une partie de celle–ci : on n'emploie que la structure d'algèbre de $\mathbb{K}$. Ainsi, si les éléments d'une matrice sont des matrices (matrice par bloc), les calculs précédents s'adaptent immédiatement. Par exemple, considérons les matrices blocs*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \qquad et \qquad B = \begin{pmatrix} B_{12} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \end{pmatrix}.$$

*Dans $A$, les matrices $A_{11}$ et $A_{12}$ d'une part et $A_{21}$ et $A_{22}$ d'autre part, ont le même nombre de lignes. De même, les matrices $A_{11}$ et $A_{21}$ d'une part et $A_{12}$ et $A_{22}$ d'autre part ont le même nombre de colonnes. Le produit de $A$ et $B$ est donné par*

$$C = \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} & A_{11}B_{13} + A_{12}B_{23} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} & A_{21}B_{13} + A_{22}B_{23} \end{pmatrix}.$$

*Là encore, le produit est associatif.*

**Exercice 2.3.1.** *On définit $E_{ij}$ la matrice $(n, n)$ dont tous les coefficients sont nuls sauf celui qui est à la croisée de la $i^e$ligne et de la $j$ $^e$colonne. Celui–là vaut 1. Montrer que $E_{ij}E_{kl} = \delta_{jk}E_{il}$ où $\delta_{jk}$ est le symbole de Krönecker* [1]

Enfin, on définit la matrice identité $\mathrm{Id}_n$, la matrice associée à l'endomorphisme Id de $E$, c'est à dire le tableau dont toutes les entrées sont nulles sauf les entrées situées sur la diagonale et qui valent 1.

Si $A \in M_n(\mathbb{K})$, $A$ est inversible s'il existe $B \in M_n(\mathbb{K})$ telle que $AB = \mathrm{Id}_n$. Si on se donne une base de $E$, $A$ est associée à l'endomorphisme $f$ et $B$, son inverse, est associé à $f^{-1}$. On note $B = A^{-1}$. On sait que $A$ est inversible si et seulement si son déterminant est nul.

## 2.4 Valeurs propres

Soit $\lambda \in \mathbb{K}$ et $A \in M_n(\mathbb{K})$ une matrice $n \times n$. Considérons la matrice $A - \lambda\ \mathrm{Id}_n$. Deux cas peuvent se produire : ou bien $A - \lambda\ \mathrm{Id}_n$ est inversible, ou bien elle ne l'est pas. Ceci est le cas si

$$\det(A - \lambda\ \mathrm{Id}_n) = 0.$$

Dans ce cas, (en identifiant endomorphisme et matrice une fois pour toute), il existe $x \in \mathbb{K}^n$ non nul dans le noyau de $A\lambda\ \mathrm{Id}_n$, i.e.

$$Ax = \lambda x.$$

On dit que $\lambda$ est une valeur propre de $A$. On a les résultats suivants

**Définition 2.4.1.** *On dit que $A \in M_n(\mathbb{K})$ est diagonalisable dans $\mathbb{K}$ s'il existe une matrice inversible $P \in M_n(\mathbb{K})$ et une matrice diagonale $D = diag\ (\lambda_i)$ telle que*

$$A = P^{-1}DP.$$

*On dit aussi que $A$ est semblable à une matrice diagonale.*

[1] $\delta_{jk} = 1$ si $j = k$ et 0 sinon.

On définit les matrices de Jordan par

$$J_k(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \ddots & 1 & \ddots & \\ \vdots & \vdots & & \ddots & \\ 0 & \dots & & & 1 \end{pmatrix}.$$

On a le

**Théorème 2.4.1 (Forme de Jordan).** *Soit $A \in M_n(\mathbb{C})$. $A$ est semblable à une matrice diagonale bloc de la forme*

$$D = \begin{pmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ \vdots & \ddots & \\ \dots & 0 & A_k \end{pmatrix}$$

*où les matrices $A_l$ sont soit de la forme $A_l = \lambda_l \ Id_{n_l}$ ou est un bloc de Jordan $J_{n_l}(\lambda_l)$ avec $\lambda_l \in \mathbb{C}$.*

On définit le rayon spectral de $A$, noté $\rho(A)$ comme étant le max des valeurs propres (éventuellement complexes) de $A$.

## 2.5 Normes

Ici encore $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$.

**Définition 2.5.1.** *Soit $E$ un espace vectoriel sur $\mathbb{K}$. Une norme est une application $|||| \ : E \to R^+$ telle que*

1. *quelques soient $x$ et $y \in E$, $||x + y|| \leq ||x|| + ||y||$,*
2. *quelques soient $x \in E$ et $\lambda \in K$, $||\lambda x|| = |\lambda| \ ||x||$,*
3. *$||x|| = 0$ si et seulement si $x = 0$.*

**Exemple 2.5.2.** *On a les exemples suivants pour $E = \mathbb{K}^n$ et $x = (x_1, \dots, x_n)$.*

- $||x||_\infty = \max_j |x_j|$
- *pour $p > 1$,* $||x||_p = \left( \sum_{j=1}^{n} |x_j|^p \right)^{1/p}$

*sont des normes.*

*On a de plus l'inégalité de Hölder–Minkowski : si $q$ est défini par $1/p + 1/q = 1$,*

$$\sum_{j=1}^{n} |x_j y_j| \leq ||x||_p \ ||y||_q.$$

*On regarde plus en détail le cas $p = q = 2$, pour lequel cette inégalité est celle de Cauchy–Schwarz.*

**Cas des espaces euclidiens.** Soit $E$ un espace vectoriel sur $\mathbb{R}$. On le munit d'une forme bilinéaire symétrique définie positive $< \ . \ , \ . > : E \times E \to \mathbb{R}$, c'est à dire d'une forme vérifiant

forme bilinéaire Pour $x$ fixé, $E \to \mathbb{R}, y \mapsto < x, y >$ est linéaire. De même, $y \mapsto < x, y >$ est linéaire quelque soit $x \in E$.

forme symétrique Quelques soient $x, y \in E$, $< x, y. =, y, x.$,

forme positive Quelque soit $x \in E$, $, x, x. \geq 0$,

forme définie . Soit $x \in E$ tel que quelque soit $y \in E$, $, x, y. = 0$ alors $x = 0$. Remarquons qu'il suffit de vérifier *ici* que si $< x, x >= 0$ alors $x = 0$.

Si on considère $E = \mathbb{R}^n$ muni de sa base canonique (qui est orthonormée pour le produit scalaire canonique), si $A \in M_n(\mathbb{R})$, en identifiant $A$ et l'endomorphisme dont $A$ est la matrice, on définit la transposition (notée $A^*$ ou ${}^tA$ ou encore $A^T$) par

$$\text{quelques soient } x, y \in E, < Ax, y >=, x, A^T y > .$$

Du point de vue matriciel, la matrice $A^T$ est définie par

$$A^T = (b_{ij})_{1 \leq i,j \leq n} \text{avec } b_{ij} = a_{ji}.$$

On peut bien sûr définir la transposée d'une matrice $(n, m)$ avec $n \neq m$.

Une matrice est symétrique si elle est égale à sa transposée. Dans ce cas, on sait qu'elle est diagonalisable à valeurs propres réelles, et on a l'exercice suivant

**Exercice 2.5.1.** *Soit $A \in M_n(\mathbb{R})$. Il existe une matrice diagonale $D$ à coefficients réels et deux matrices unitaires[2] $U$ et $V$ telles que*

$$A = U^T D V.$$

A partir de là, on peut définir une notion de norme induite sur $M_n(\mathbb{R})$. On a la

**Définition 2.5.3.** *Soit $A \in M_n(\mathbb{R})$. On définit*

$$||A|| = \sup_{x \neq 0} \frac{||Ax||}{||x||} \left( = \sup_{||x||=1} ||Ax|| \right).$$

Montrons pour commencer que si on est en dimension finie, alor ce nombre ($||A||$) est fini. Soit $x = (x_1, \ldots, x_n)$, on note encore par $\{e_1, \ldots, e_n\}$ la base canonique de $\mathbb{K}^n$. On a alors $x = \sum_i x_i e_i$ et donc

$$||Ax|| \leq \sum_i |x_i| \ ||Ae_i|| \leq C \sum_i |x_i| = C \ ||x||_1$$

où $C = \max_i ||Ae_i||$. Etant en dimension finie, toutes les normes sont équivalentes. Il existe $C_1 > 0$ et $C_2 > 0$ tels que

$$C_1 ||x||_1 \leq ||x|| \leq C_2 ||x||_2,$$

donc

$$||Ax|| \leq \frac{C}{C_1} \ ||x||$$

d'où le résultat.

[2] $R$ est unitaire si $RR^T = Id$.

**Remarque 2.5.4.** *On a les résultats suivants*
- $||\ Id_n|| = 1$,
- *Quelque soit la norme* $||\ .\ ||$, *on a* $\rho(A) \leq ||A||$.

**Exercice 2.5.2.** *Déterminer les normes induites pour les normes* $L^1$, $L^\infty$, $L^2$.

On a aussi l'inégalité suivante : quelques soient $A, B \in M_n(\mathbb{K})$,

$$||AB|| \leq ||A||\ ||B||. \tag{2.1}$$

On peut aussi montrer que si $||\ .\ ||$ est une norme sur $M_n(\mathbb{K})$ qui vérifie l'inégalité (2.1), alors il existe une norme sur $\mathbb{K}^n$ telle quelque soit $x \in \mathbb{K}^n$, $||Ax|| \leq ||A||\ ||x||$.[3] Ceci montre que la norme de Shur définie par $||A|| = \mathrm{tr}(AA^T)$ qui n'est pas une norme matricielle (car $||Id|| = n$) vérifie néanmoins $||Ax|| \leq ||A||\ ||x||$ pour une norme définie sur $\mathbb{K}^n$ bien choisie.

En employant la forme de Jordan des matrices on peut montrer le résultat suivant

**Théorème 2.5.3.** *Soit* $A \in M_n(\mathbb{K})$. *La suite* $\{A^n\}_{n\in\mathbb{N}}$ *converge vers la matrice nulle si et seulement si* $\rho(A) < 1$.

dont on déduit le corrolaire

**Corollaire 2.5.5.** *La suite* $\sum_{l=0}^{n} A^l$ *et convergente si et seulement si* $\rho(A) < 1$. *Si elle converge, elle onverge vers* $(\ Id - A)^{-1}$.

On a alors l'exercice

**Exercice 2.5.4.** *Si* $A \in M_n(\mathbb{K})$, *quelque soit la norme induite,*

$$\rho(A) = \lim_{k\to+\infty} ||A^k||^{1/k}.$$

## 2.6 Conditionnement d'un système linéaire

Si $A \in M_n(\mathbb{K})$, on peut rarement résoudre *exactement* le problème $Ax = b$. Ici on ne parle pas de la résolution effective du problème, ceci sera abordé au chapitre suivant, mais du fait qu'en pratique la matrice $A$ et/ou le vecteur $b$ sont connus imparfaitement (erreur de mesure, erreurs de calculs, d'arrondis, etc). Quelle est l'influence de cela sur la solution $x$, supposant qu'on a une méthode parfaite de résolution.

On va regarder le problème où au lieu de résoudre $Ax = b$, on résoud

$$(A + \delta A)y = b + \delta b.$$

On commence par un exemple. On prend

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \qquad b = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}.$$

[3] Indication. si $x \in \mathbb{K}^n$, on considère la matrice $[x]$ dont les vecteurs colonnes sont $x$ et 0 $n-1$ fois. On pose alors $||x|| = |||[x]|||$.

La solution est

$$x = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Si on perturbe les données par

$$A + \delta A = \begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.99 & 9.89 & 9 \\ 6.99 & 4.99 & 9 & 9.98 \end{pmatrix} \qquad b = \begin{pmatrix} 32.01 \\ 22.99 \\ 33.01 \\ 30.99 \end{pmatrix}.$$

La solution est alors

$$y = \begin{pmatrix} 1.82 \\ -1.36 \\ 0.35 \\ -0.21 \end{pmatrix}.$$

La matrice $\delta A$ est

$$\delta A = \begin{pmatrix} 0 & 0 & 0.1 & 0.2 \\ 0.08 & 0.04 & 0 & 0 \\ 0 & -0.02 & -0.11 & 0 \\ -0.01 & -0.01 & 0 & -0.02 \end{pmatrix}$$

et $y - x = \begin{pmatrix} 0.82 \\ -1.36 \\ 0.35 \\ -0.21 \end{pmatrix}$ On voit que de petites causes peuvent avoir de grands effets... De même on voit que l'équation

$$(A + \delta A)z = b$$

produit

$$z = \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}$$

soit

$$z - x = \begin{pmatrix} -82 \\ 136 \\ -35 \\ 21 \end{pmatrix}.$$

On voit donc que sans perturber le second membre, les résultats sont pires.

Comment quantifier de phénomène ? On a les résultats suivants

**Théorème 2.6.1.** *Soit $A$ une matrice inversible. Soient $x$ et $x + \delta x$ les solutions de*

$$Ax = b \qquad A(x + \delta x) = b + \delta b.$$

*Alors*

$$\frac{||\delta x||}{||x||} \leq ||A||\,||A^{-1}||\,\frac{||\delta b||}{||b||}.$$

*Démonstration.* On a $A\delta x = \delta b$ soit $\delta x = A^{-1}\delta b$, d'où

$$||\delta x|| \leq ||A^{-1}||\ ||\delta b||.$$

De même, puisque $Ax = b$, on a

$$||b|| \leq ||A||\ ||x||$$

donc

$$\frac{||\delta x||}{||x||} \leq ||A^{-1}||\frac{||\delta b||}{||b||} \leq ||A||\ ||A^{-1}||\ \frac{||\delta b||}{||b||}.$$

□

**Théorème 2.6.2.** *Soit $A$ une matrice inversible, $x$ et $x + \delta x$ les solutions de*

$$Ax = b, \qquad (A + \delta A)(x + \delta x) = b.$$

*Alors*

$$\frac{||x||}{||x + \delta x||} \leq ||A||\ ||A^{-1}||\ \frac{||\delta A||}{||A||}.$$

*Démonstration.* Puisque

$$b = (A + \delta A)(x + \delta x) = Ax,$$

on a

$$A\delta x + \delta A(x + \delta x) = 0$$

soit $\delta x = -A^{-1}\delta A(x + \delta x)$ et donc $||\delta x|| \leq ||A^{-1}||\ ||\delta A||\ ||x + \delta x||$ d'où le résultat. □

On définit alors le conditionnement d'une matrice par

**Définition 2.6.1.** *On appelle conditionnement d'une matrice inversible $A \in M_n(\mathbb{R})$ le nombre*

$$cond\ A = ||A||\ ||A^{-1}||.$$

On pose cond $_p(A) = ||A||_p\ ||A^{-1}||_p$.
On alors

**Proposition 2.6.2.** *Soit $A \in M_n(\mathbb{R})$. On a les propriétés suivantes*

- *cond $(\alpha A) =$ cond $(A)$ quelque soit $\alpha \neq 0$,*
- *cond $A \geq 1$.*
- *cond $_2(A) = \frac{\mu_m ax}{\mu_{min}}$ où $\mu_{max}$ (resp. $\mu_{min}$) est la plus grande (resp. plus petite) valeur propre de $AA^T$.*
- *cond $_2(A) = 1$ si et seulement si $A = \alpha Q$ où $Q$ est unitaire,*
- *Si on change de norme, il existe des constantes $K_1$ et $K_2$ positives telles que*

$$K_1\ cond\ _{||||}A \leq\ cond\ _{||\ ||'}A \leq K_2\ cond\ _{||||}A.$$

Les démonstrations sont simples et laissées en exercice. La dernière propriété montre que si le conditionnement dans une norme est mauvais, on n'a que très peu de chance de l'améliorer en changeant de norme. Il faut recourir à d'autres méthodes pour améliorer le conditionnement d'une matrice, par exemple pré–multiplier $A$ par une «bonne» matrice $P$ dite de préconditionnement afin que cond $(PA)$ soit sensiblement plus petit que cond $(A)$. La recherche de telles matrices de préconditionnement ne peut se faire qu'au cas par cas et est de toute manière très difficile. Deux choix classiques sont $P =$ la matrice diagonale formée des termes diagonaux de $A$, ou si $A = Id - B$ avec $||B|| < 1$, la matrice $P = Id + P$ qui est le premier terme du dévelopement en série de $A^{-1} = (Id - B)^{-1}$.

**Remarque 2.6.3.** *Il n'y a aucun lien entre le conditionnement d'une matrice et son déterminant. Considérons l'exemple suivant*

$$A = \begin{pmatrix} 1 & 2 & & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & \dots & & 1 \end{pmatrix}$$

*On a* $\det A = 1$. *L'inverse de* $A$ *est*

$$A^{-1} = \begin{pmatrix} 1 & -2 & 4 & \dots & (-2)^{n-1} \\ & \ddots & \ddots & \ddots & \\ 0 & & \dots & & 1 \end{pmatrix}$$

*On a* $||A||_\infty = ||A^T||_1 = 3$ *et* $A^{-1}||_\infty = ||A^{-1}||_1 = 1 + 2 + 4 + \ldots + 2^{N-1} = 2^N - 1$. *Ainsi,*

$$cond\ _1A = 3(2^N - 1).$$

En TD, on regardera les matrices suivantes

– Matrice de Hilbert. $H = (h_{ij})_{1 \le i,j \le n}$ avec

$$h_{ij} = \frac{1}{i + j - 1}.$$

– Soit $A$ une matrice diagonale d'ordre 100. On suppose $A_{11} = 1$ et $A_{ii} = 1/10$ dès que $i > 1$. On a $\det A = 10^{-99}$ mais cond $_2(A) = 10$.

On finit sur le résultat suivant qui relie le comportement de l'inverse d'une matrice au voisinage d'une matrice donnée au conditionnement de cette dernière

**Théorème 2.6.3.** *Soient* $A$ *et* $A + \delta A$ *deux matrices inversibles. Alors*

$$\frac{||(A + \delta A)^{-1} - A^{-1}||}{||(A + \delta A)^{-1}||} \le \ cond\ A \frac{||\Delta A||}{||A||}.$$

*Démonstration.* Puisque $(A + \delta A)^{-1}(A + \delta A) =$ Id,

$$(A + \delta A)^{-1}(Id + \delta A\ A^{-1}) = A^{-1}.$$

Donc

$$\begin{aligned} (A + \delta A)^{-1} - A^{-1} &= (A + \delta A)^{-1} - (A + \delta A)^{-1}(Id + \delta A\ A^{-1}) \\ &= (A + \delta A)^{-1} \Big( \text{Id} - \ \text{Id} - \delta A\ A^{-1} \Big) \\ &= -(A + \delta A)^{-1} \delta A\ A^{-1} \end{aligned}$$

d'où

$$||(A+\delta A)^{-1} - A^{-1}|| \leq ||(A+\delta A)^{-1}||\ ||\delta A||\ ||A^{-1}||$$

ce qui achève la preuve. □

# Chapitre 3

# Résolution de systèmes linéaires

## 3.1 Motivation

On veut connaître la répartition de chaleur dans une barre $[0, 1]$ chauffée aux deux bouts. Si $\sigma$ désigne la conductivité thermique du matériau, on sait que la température $T$ vérifie

$$\frac{\partial}{\partial x}\left(\sigma \frac{\partial T}{\partial x}\right) = f(x)$$

dans le matériau (ici $f$ désigne la source éventuelle de chaleur), et on a les deux conditions aux limites $T(0) = a$, $T(1) = b$. Si la conductivité thermique $\sigma$ dépend de $x$ (par exemple si la barre n'est pas homogène), il est très difficile voire impossible de résoudre analytiquement le problème. Bien sûr, il existe une formule analytique

$$T(x) = \int_0^x \left( \int_0^s f(t)dt \right) ds + Cx + C'$$

où $C$ et $C'$ sont deux constantes qu'on détermine en évaluant $T(0) = a$ et $T(1) = b$, mais il est très rare qu'on puisse evaluer explicitement l'intégrale double. Dans ce cas, il faut recourir à une formule de quadrature, donc une approximation. Qui plus est, la méthode employée ici (calcul successif de primitive) ne fonctionne qu'en une dimension d'espace.

Supposons pour simplifier que $\sigma$ soit constant. Pour évaluer numériquement la solution, on va approcher la dérivée seconde. Dans ce cas, on se donne un maillage $x_i = i\Delta x$, $\Delta x = 1/N$ et $i = 0, \dots N$. En employant des développements de Taylor on voit que

$$\frac{\partial^2 T}{\partial x^2} = \frac{T_{i+1} - 2T_i + T_{i-1}}{\Delta x^2} + \mathcal{O}(\Delta x^4).$$

On va décider d'approcher l'opérateur d'ordre deux par la différence divisée, et on est conduit au problème

$$\begin{aligned}
&b - 2T_1 + T2 = \Delta x^2 f_1 \\
&T_1 - 2T_2 + T_3 = \Delta x^2 f_2 \\
&\vdots \\
&T_{N-2} - 2T_{N-1} + a = \Delta x^2 f_N
\end{aligned}$$

On est donc conduit à résoudre le système $Ax = b$ avec

$$A = -\begin{pmatrix} 2 & -1 & \dots & 0 \\ -1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \\ 0 & \dots & -1 & 2 \end{pmatrix} \tag{3.1}$$

et $x = \begin{pmatrix} T_1 & \dots & T_N \end{pmatrix}$ et $b = \begin{pmatrix} \Delta x^2 f_1 - b \\ \Delta x^2 f_2 \\ \vdots \\ \Delta x^2 f_N - a \end{pmatrix}$

On va donc s'intéresser au problème suivant. Soit $A \in M_n(\mathbb{K})$ une matrice carré d'ordre $n$ et $b \in \mathbb{K}^n$. On cherche à déterminer $x \in \mathbb{K}^n$ tel que

$$Ax = b.$$

On sait qu'une condition nécessaire et suffisante pour qu'il existe une unique solution est $\det A \neq 0$.

On note $a_i$ le i$^{\text{e}}$vecteur colonne de $A$. Si $x = (x_1, \dots, x_n)$, la solution est donnée par la formule de Cramer

$$x_i = \frac{\det(a_1, a_2, \dots, \hat{a}_i, \dots)}{\det(a_1, \dots, a_n)}$$

où $\hat{a}_i = b$.

Evaluons le nombre d'opérations nécessaires au calcul dans le cas d'une matrice pleine (i.e. aucun ou très peu de coefficients $a_{ij}$ sont nul, ce qui n'est pas le cas de l'exemple plus haut). Néanmoins, même dans le cas d'une matrice creuse, les formules de Cramer sont généralement peu efficaces en terme de précision.

On a donc $n + 1$ déterminants à évaluer. Dans le cas du déterminant de $A$ (par exemple), on a

$$\det A = \sum_{\sigma \in \mathcal{S}_n} (-1)^{s(\sigma)} a_{1\sigma(1)} \dots a_{n\sigma(n)}.$$

Ici, $\mathcal{S}_n$ désigne l'ensemble de toutes les bijections de $\{1, \dots, n\}$, $s(\sigma)$ est la signature de $\sigma$. Le calcul de $\det A$ nécessite donc $n!$ additions et $nn!$ multipplications. En négligeant les additions (dont leur coup n'est cependant pas négligeable ...), on a donc $n(n+1)!$ opérations à effectuer.

Si on a un ordinateur qui effectiue $10^9$ opérations par seconde (un ordinateur puissant), pour résoudre un système $50 \times 50$, il faudrait de l'ordre de

$$\frac{50 \times 51!}{10^9}$$

secondes soit

$$\frac{50 \times 51!}{10^9 \times 3600 \times 24 \times 365} \simeq 2.45\ 10^{51} \text{ années } \dots$$

C'est donc sans espoir.

Dans la suite de ce chapitre, on va étudier deux grandes classes de méthodes : les méthodes directes où on cherche à déterminer la solution du problème avec un coût de calcul de l'ordre $n^k$ où $A$ la matrice est $(n, n)$ et $k$ est un entier de l'ordre 2 ou 3 ; et les méthodes itératives où on se contente de construire une suite de vecteurs $x_n$ qui converge vers la solution du problème. En effet, dans bien des cas, il n'est pas utile de connaître exactement la solution, mais seulement une (très) bonne approximation.

## 3.2 La méthodes de Gauss et quelques variantes

### 3.2.1 Introduction

La résolution de $Ax = b$ est très simple dans plusieurs cas :
- si $A$ est diagonale,
- si $A$ est unitaire : si $AA^T = Id$ alors $A^{-1} = A^T$,
- si $A$ est triangulaire supérieure ou inférieure.

Remarquons cependant que ce n'est pas parce que la résolution est simple conceptualement que c'est efficace. Un bon contre exemple est le calcul des coefficents d'un polynôme trigonométrique de degré donné fixé. Ce problème se résoud en écrivant un système linéaire dont la matrice est propostionnelle à une matrice unitaire. Ceci conduit à un algorithme de coût $\mathcal{O}(n^2)$. Dans bien des cas, on peut construire des algorithmes en $\mathcal{O}(\log n)$ stable, ce qui est bien meilleurs.

**Définition 3.2.1.** *On dit que $A = (a_{ij})_{1\leq i,j\leq n}$ est*
- *triangulaire supérieure si $a_{ij} = 0$ quand $i < j$,*
- *triangulaire inférieure si $a_{ij} = 0$ quand $i > j$.*

Par exemple,

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix}$$

est triangulaire supérieure et

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 5 & 6 \end{pmatrix}$$

est triangulaire inférieure.

La remarque essentielle est que quand $L$ est triangulaire inférieure ou supérieure, le système $Lx = b$ est très simple à résoudre si $L$ est inversible. Ecrivons le

$$\left\{ \begin{array}{ccccccccc} l_{11}x_1 & + & l_{12}x_2 & + & \dots & + & l_{1n}x_n & = & b_1 \\ & & l_{22}x_2 & + & \dots & + & l_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & \vdots & & \vdots & & \\ & & & & & & l_{nn}x_n & = & b_n \end{array} \right. \tag{3.2}$$

Ce système est inversible si et seulement si $l_{ii} \neq 0$ quelque soit $i$. On résoud (3.2) par «remontée». On a donc

$$x_n = \frac{b_n}{l_{nn}},$$

puis

$$x_{n-1} = \frac{1}{l_{n-1n-1}} \left( b_{n-1} - l_{n-1n} x_{n-1} \right),$$

etc. Ces formules sont explicites puisqu'on travaille de proche en proche.

La méthode de Gauss consiste (du point de vue matriciel) à transformer $A$ en la multipliant à droite ou à gauche par une matrice «bien choisie» explicite afin de se ramener à un système triangulaire.

On commence par un exemple. Considérons le système

$$\begin{cases} 2x & + & 4y & - & 4z & + & t & = & 0 \\ 3x & + & 6y & + & z & - & 2t & = & -7 \\ -x & + & y & + & 2z & + & 3t & = & 4 \\ x & + & y & - & 4z & + & t & = & 2 \end{cases}$$

qui correspond à

$$A = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 3 & 6 & 1 & -2 \\ -1 & 1 & 2 & 3 \\ 1 & 1 & -4 & 1 \end{pmatrix} \qquad b = \begin{pmatrix} 0 \\ -7 \\ 4 \\ 2 \end{pmatrix}.$$

On additionne la $2^{\text{e}}$ligne à $-3/2$ de la première, puis la $3^{\text{e}}$à $-1/2$ de la première, et la $4^{\text{e}}$à $-1/2$ de la seconde et on a

$$\begin{cases} 2x & + & 4y & - & 4z & + & t & = & 0 \\ & & & & 7z & - & \frac{7}{2}t & = & -7 \\ & & 3y & & & + & \frac{7}{2}t & = & 4 \\ & - & y & - & 2z & + & \frac{t}{2} & = & 4 \end{cases}$$

En permuttant la deuxième et la troisième ligne, on a

$$\begin{cases} 2x & + & 4y & - & 4z & + & t & = & 0 \\ & & 3y & & & + & \frac{7}{2}t & = & 4 \\ & & & & 7z & - & \frac{7}{2}t & = & -7 \\ & - & y & - & 2z & + & \frac{t}{2} & = & 4 \end{cases}$$

Pour finir, on additionne à la $4^{\text{e}}$ligne à 3 fois la seconde pour avoir

$$\begin{cases} 2x & + & 4y & - & 4z & + & t & = & 0 \\ & & 3y & & & + & \frac{7}{2}t & = & 4 \\ & & & & 7z & - & \frac{7}{2}t & = & -7 \\ & & & & & & \frac{5}{3}t & = & \frac{4}{3} \end{cases}$$

On peut obtenir la matrice finale

$$\begin{pmatrix} 2 & 4 & -4 & 1 \\ 0 & 3 & 0 & 7 \\ 0 & 0 & 7 & -\frac{7}{2} \\ 0 & 0 & 0 & \frac{2}{3} \end{pmatrix}$$

par multiplication à gauche de matrices triangulaires inférieures.

En effet

$$AL_1 = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 3 & 6 & 1 & -2 \\ -1 & 1 & 2 & 3 \\ 1 & 1 & -4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{3}{2} & 1 & 0 & 0 \\ \frac{1}{2} & 0 & 1 & 0 \\ -\frac{1}{2} & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 0 & 0 & 7 & -\frac{7}{2} \\ 0 & 7 & 0 & \frac{7}{2} \\ 0 & -1 & -2 & \frac{1}{2} \end{pmatrix} = A_1$$

Ceci correspond à la première étape du calcul. En multipliant à gauche par la matrice de permutation

$$P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

on a

$$A_2P_2 = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 0 & 3 & 0 & \frac{7}{2} \\ 0 & 0 & 7 & -\frac{7}{2} \\ 0 & -1 & -2 & \frac{1}{2} \end{pmatrix} = A_3$$

Puis en effectuant

$$A_3L_3 = A_2 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 3 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 0 & 3 & 0 & \frac{7}{2} \\ 0 & 0 & 7 & -\frac{7}{2} \\ 0 & 0 & 0 & \frac{2}{3} \end{pmatrix} = A_3.$$

On a donc $AL_1PL_2 = A_3$ qui est triangulaire supérieure.

### 3.2.2 Présentation de la méthode

On montre dans le cas général comment passe d'un système $AX = b$ à un système triangulaire $\widetilde{A}x = \widetilde{b}$ avec $\tilde{a}_{ij} = 0$ si $i > j$. Il suffit ensuite de «remonter» $x_n = \frac{\tilde{b}_n}{\tilde{a}_{nn}}$ etc.

**Etape 1** On pose $a_{ij}^{(1)} = a_{ij}$ et $b_i = b_i^{(1)}$. Pour $i = 2, \ldots, n$, on pose

$$\begin{array}{ll} m_1^{(1)} = \dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}} & \\ a_{ij}^{(2)} = a_{ij}^{(1)} - m_i^{(1)} a_{1j}^{(1)} & j = 2, \ldots, n \\ b_i^{(2)} = b_i^{(1)} - m_i^{(1)} b_1^{(1)}. & \end{array}$$

On obtient donc $A^{(2)}x = b^{(2)}$ avec

$$A^{(2)} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \ldots & a_{nn}^{(2)} \end{pmatrix}$$

**Etape** $k$ On procède par induction. Pour $i = k+1, \ldots, n$ on pose

$$\begin{array}{ll} m_1^{(k)} = \dfrac{a_{i1}^{(k)}}{a_{11}^{(k)}} & \\ a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_i^{(k)} a_{1j}^{(k)} & j = k+1, \ldots, n \\ b_i^{(k+1)} = b_i^{(k)} - m_i^{(k)} b_1^{(k)}. & \end{array}$$

qui conduit au système $A^{(k+1)}x = b^{(k+1)}$ avec

$$A^{(k+1)} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ 0 & a_{22}^{(2)} & \ldots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{kk^{(k)}} & a_{kk+1}^{(k)} & a_{kn}^{(k)} \\ \vdots & a_{nl}^{(k)} & \vdots & a_{nn}^{(n)} \end{pmatrix}.$$

La dernière étape est l'étape $n-1$.

On finit le calcul en résolvant un système triangulaire comme effectué plus haut.

Le nombre d'opération est de l'ordre $\frac{n^3}{3}$, voir TD.

**Remarque 3.2.2 (Calcul du déterminant).** *Ayant décomposé la matrice en un produit de matrices élémentaires (matrices triangulaire inférieures de déterminant 1 et matrices de permutation dont le déterminant est $(-1)^{s(\sigma)}$), on voit que le déterminant de $A$ vaut*

$$\det(A) = (-1)^s \Pi_{k=1}^n a_{kk}^{(k)}$$

*où $s$ est égal à la somme des signatures des permutations nécessaires au calcul.*

### 3.2.3 Cas d'un pivot nul, pivot maximal

Il se peut que $a_{kk}^{(k)} = 0$. C'est par exemple le cas de l'exemple numérique après la première étape. Dans ce cas, il faut permuter les équations, i.e. chercher $i > k$ tel que $a_{ik}^{(k)} \neq 0$ et permuter les lignes $i$ et $k$ dans $A^{(k)}$ et $b^{(k)}$.

Remarquons que puisque $A$ est inversible, $A^{(k)}$ est inversible. Son déterminant est

$$\det A^{(k)} = a_{11}^{(1)} a_{22}^{(2)} \dots a_{k-1k-1}^{(k-1)} \det A'$$

où $A'$ est le bloc restant de taille $(n-k+1, n-k+1)$. Ceci montre que $\det A' \neq 0$ donc nécessairement l'un des termes $a_{ik}^{(k)}$ est non nul.

En fait, on a intérêt à choisir le pivot $a_{ik}^{(k)}$ le plus grand possible en module. Pour voir cela, on considère l'exemple simple suivant

$$\begin{aligned} \varepsilon x_1 \quad + \quad x_2 &= \frac{1}{2} \\ x_1 + x_2 &= 1 \end{aligned}$$

Si on choisit $\varepsilon$ comme pivot, on a

$$\begin{aligned} \varepsilon x_1 \quad + \qquad x_2 &= \frac{1}{2} \\ \left(1 - \tfrac{1}{\varepsilon}\right) &= 1 - \frac{1}{2\varepsilon} \end{aligned}$$

et donc

$$x_2 = \frac{1 - \frac{1}{2\varepsilon}}{1 - \frac{1}{\varepsilon}} = \frac{2\varepsilon - 1}{2(\varepsilon - 1)} \simeq \frac{1}{2}$$

et

$$x_1 = \frac{1}{\varepsilon}\left(\frac{1}{2} - x_2\right) = \frac{1}{2(1-\varepsilon)} \simeq \frac{1}{2}$$

Si $\varepsilon$ est trop petit, $\frac{1}{\varepsilon}$ est très grand, et donc (voir introduction et TDs), on aura

$$x_2 = \frac{1}{2} \text{ et } x_1 = 0.$$

L'ordinateur ne fait pas de simplifications astucieuses, il se contente de faire les calculs qu'on lui demande.

Ceci ne se produit pas si on permute les deux premières lignes :

$$\begin{aligned} x_1 + x_2 &= 1 \\ \varepsilon x_1 \quad + \quad x_2 &= \frac{1}{2} \end{aligned}$$

donne

$$\begin{aligned} x_1 + x_2 &= 1 \\ + \quad (1-\varepsilon)x_2 &= \frac{1}{2} - \varepsilon \end{aligned}$$

d'où

$$\begin{aligned} x_2 &= \frac{\frac{1}{2} - \varepsilon}{1-\varepsilon} \\ x_1 &= 1 - \frac{\frac{1}{2} - \varepsilon}{1-\varepsilon} \end{aligned}$$

Quand $\varepsilon$ est très petit, la première relation donnera numériquement $x_2 \equiv \frac{1}{2}$ et la seconde donnera $x_1 \equiv \frac{1}{2}$, ce qui est le bon résultat. La différence essentielle entre les deux versions est que dans un cas, il faut introduite $1/\varepsilon$ qui est très grand, peut–être au–delà de l'infini machine, alors que la seconde n'introduit que $\varepsilon$. On a vu en TD des exemples où $x + \varepsilon = x$.

### 3.2.4 Propriétés supplémentaires.

**Sur les matrices triangulaires inférieures et supérieures**

On a le résultat suivant

**Proposition 3.2.3.** *On a*

1. *le produit de deux matrices triangulaires inférieures (resp. supérieures) est encore une matrice triangulaire inférieure (resp. supérieure),*
2. *L'inverse d'une matrice triangulaire inférieure (resp. supérieure) est du même type,*

*Démonstration.* Il suffit de faire la démonstration pour des matrices triangulaires supérieures. On a les résultats pour les matrices triangulaires supérieures en passant à la transposition.

1. Soit $L = (l_{ij})_{i,j}$ une matrice triangulaire inférieure inversible. Dans le paragraphe 3.2.1, on a résolu un système linéaire général triangulaire inférieur. Les résultats donnent

   $$x_n = l'_{nn} b_n$$

   avec $l'_{nn} = 1/l_{nn}$, puis on reporte dans l'avant dernière équation et on voit que $x_{n-1}$ dépend linéairement de $b_n$ et $b_{n-1}$ seulement,

   $$x_{n-1} = l'_{n-1,n-1} b_{n-1} + l_{n-1,n} b_n.$$

   de proche en proche, on voit que

   $$x_k = \sum_{j=k}^{n} l'_{kj} b_j.$$

   En réinterprétant cela de manière matricielle, on voit que

   $$L^{-1} = (l'_{ij})_{1 \leq i,j \leq n}$$

   où $l'_{ij} = 0$ si $i > j$ et $l'_{ij}$ est défini plus haut quand $i \leq j$ : la matrice $L^{-1}$ est triangulaire supérieure.

2. Soient $L$ et $M$ deux matrices triangulaires supérieures. Leur produit est la matrice de coefficient général

$$\sum_{k=1}^{n} l_{ik} m_{kj}.$$

Supposons que $i > j$. Alors,

$$\begin{aligned}\sum_{k=1}^{n} l_{ik} m_{kj} &= \sum_{k=1}^{i} l_{ik} l_{kj} + \sum_{k=i+1}^{n} l_{ik} l_{kj} \\ &= 0 + \sum_{k=i+1}^{n} l_{ik} l_{kj} \text{ car } i > k \\ &= 0 + 0 \text{ car } k > i > j\end{aligned}$$

□

**Remarque 3.2.4.** *L'inverse de la matrice*

$$L = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 & & \\ 0 & 1 & \dots & \dots & 00 & \ddots & m_{i+1}^{i} & 0 \\ \vdots & \vdots & m_{n}^{i} & \vdots & & & \end{pmatrix}$$

*est*

$$L^{-1} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 & & \\ 0 & 1 & \dots & \dots & 00 & \ddots & -m_{i+1}^{i} & 0 \\ \vdots & \vdots & -m_{n}^{i} & \vdots & & & \end{pmatrix}$$

*On peut vérifier cela en exercice.*

**Exercice 3.2.1.** *Montrer que le nombre d'opération de la méthode de Gauss est de l'ordre* $2n^3/3$.

### 3.2.5 Propriétés de la méthodes de Gauss

**Proposition 3.2.5.** *Soit $A$ une matrice réelle $(n, n)$ symétrique. Si les pivots de Gauss $a_{kk}^{(k)}$ sont non nuls, alors à chaque étape de la méthode de Gauss,*

$$a_{ij}^{(k+1)} = a_{ji}^{(k+1)}$$

*pour $i = k+1, \dots n$ et $j = k+1, \dots, n$ et $k = 1, \dots, n-1$.*

En conséquence, on ne calculera les $a_{ij}^{(k+1)}$ que pour $j \leq i$ ce qui divise par 2 le nombre d'opération à chaque étape.

*Démonstration.* On voit que c'est vrai à l'étape zéro : la matrice est symétrique. Supposons que ce soit vrai jusqu'à l'étape $k$. On applique l'algorithme

$$\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} \\ &= a_{ji}^{(k)} - \frac{a_{ki}^{(k)}}{a_{kk}^{(k)}} a_{jk}^{(k)} \text{récurence} \\ &= a_{ji}^{(k+1)} \end{aligned}$$

□

La matrice

$$-\begin{pmatrix} 2 & -1 & 0 & \dots & \\ 0 & \ddots & \ddots & & \\ \vdots & \vdots & \vdots & & \\ 0 & \dots & & -1 & 2 \end{pmatrix}$$

a presque tous ses termes nuls. Plus précisément, si $|i-j| > 1$ alors $a_{ij} = 0$. On définit la demi–largeur de bande

**Définition 3.2.6 (Demi–largeur de bande).** *Soit $A = (a_{ij})$ une matrice. Le plus petit entier $K$ tel que si $|i-j| > K$ alors $a_{ij} = 0$ est appelé demi–largeur de bande.*

Dans l'exemple précédent, $K = 1$.
On a alors la proposition suivante

**Proposition 3.2.7.** *Si $A \in M_n(\mathbb{R})$ est une matrice de demi–largeur de bande $K < n-2$ et si les pivots $a_{kk}^{(k)}$ sont tous non nuls, alors à chaque étape de la méthode de Gauss, $A^{(k)}$ est une matrice de demi–largeur de bande inférieure ou égale à $K$.*

L'intérêt de ce résultat est le suivant. On dit qu'une matrice est creuse si une grande majorité de ses coefficients sont nuls. Si le coefficient $a_{ij}$ de $A$ est nul, il n'y a aucune raison que le coefficient $a_{ij}^{(1)}$ soit aussi nul. Par contre le résultat précédent dit que si $|i-j| > K$ (donc $a_{ij} = 1$), alors $a_{ij}^{(1)} = 0$. Ainsi, il suffit de réserver en mémoire tous les coefficients tels que $|i-j| \leq K$ puisqu'on sait à l'avance que ce sont les seuls qui seront *a priori* non nuls dans les décompositions. Ceci permet de sauver beaucoup de mémoire dans de nombreux cas.

*Démonstration.* On procède par récurence. Supposons que la demi–largeur de bande de $A^{(k)}$ soit $K$.

Soit $j \geq i + K$ et $i, j \geq k+1$. Alors

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_i^{(k)} a_{kj}^{(k)}$$

est nul car d'une part $a_{ij}^{(k)} = 0$ par hypothèse de récurence et d'autre part $j \geq i + K$ et $i \geq k+1$ donc $j \geq k + K + 1$. Ainsi $a_{kj}^{(k)} = 0$.

Si $j \geq i+K$ et $i, j \leq k$, les coefficients sont nuls par construction. Enfin, si $j \leq i-K$, on procède par symétrie. □

**Remarque 3.2.8.** *La largeur de bande d'une matrice dépend fortement de la numérotation des coordonnées du vecteur inconnu $x = (x_1, \ldots, x_n)$. Considérons par exemple la matrice (3.1) pour $n = 5$ dont la demi largeur de bande est 1. Si on permute la seconde composante et la quatrième composante de $x$, c'est à dire qu'on multiplie à gauche par*

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

*on trouve*

$$\begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & -1 & 2 & -1 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

*La largeur de bande est devenue 2, c'est à dire le maximum possible pour cette matrice. Ainsi la numérotation qu'on choisi joue un rôle ....*

*La recherche d'une numérotation «optimale» permettant de minimiser la largeur de bande d'une matrice est un problème important mais délicat qui dépasse de loin le contenu de ce cours.*

### Décomposition LU

**Définition 3.2.9.** *Soit $A = (a_{ij})_{1 \leq i,j \leq n}$ une matrice de $M_n(\mathbb{K})$. On appelle mineurs principaux de $A$ les déterminants des matrices*

$$\alpha^{(k)} = \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & \vdots & \vdots \\ a_{k1} & \ldots & a_{kk} \end{pmatrix}.$$

La proposition suivante donne une condition suffisante pour que les pivots soient non nuls

**Proposition 3.2.10.** *Soit $A$ une matrice dont les mineurs principaux sont non nuls. Alors*

1. *les pivots $a_{kk}^{(k)}$ sont non nuls,*
2. *Il existe une matrice triangulaire inférieure $L$ et une matrice triangulaire supérieure $U$ telle que*

$$A = LU.$$

*Démonstration.* On fait la démonstration par induction. On pose $A^{(1)} = A$.

**Etape 1** Le premier mineur principal est $a_{11}$ qui est donc non nul par hypothèse. On a $a_{11} = a_{11}^{(1)}$. On calcule $A^{(2)}$ On a

$$\begin{array}{l} a_{ij}^{(2)} = a_{ij}^{(1)} - m_i^{(1)} a_{1j}^{(1)} \quad \text{pour } i,j = 2,\ldots,n \\ a_{1j}^{(2)} = a_{1j}^{(1)} \end{array} \tag{3.3}$$

A partir des formules (3.3), on voit qu'on peut écrire $A^{(2)} = M^{(1)} A^{(1)}$ avec

$$M^{(1)} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ -m_2^{(1)} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \\ -m_n^{(1)} & 0 & & 1 \end{pmatrix}$$

On sait que

$$(M^{(1)})^{-1} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ m_2^{(1)} & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \\ m_n^{(1)} & 0 & & 1 \end{pmatrix}$$

et $A^{(2)} = (M^{(1)})^{-1} A^{(2)}$.

On introduit la notation suivante. Si $l \geq 1$, on découpe une matrice $A$ en 4 blocs tels que le bloc supérieur gauche est de taille $(l,l)$,

$$A = \begin{pmatrix} A_{l,11} & A_{l,12} \\ A_{l,21} & A_{l,22} \end{pmatrix}$$

On découpe alors $A^{(1)}$, $(M^{(1)})^{-1}$ et $A^{(2)}$ en 4 blocs tels que le bloc supérieur gauche est de taille $(l,l)$. Un calcul facile donne (avec des notations évidentes)

$$A_{l,11}^{(1)} = (M^{(1)})_{l,11}^{-1} A_{1,11}^{(2)}$$

car le bloc supérieur droit de $(M^{(1)})_{l,11}^{-1}$ est nul, d'où

$$\det A_{l,11}^{(1)} = \det(M^{(1)})_{l,11}^{-1} \ \det A_{,11}^{(2)}.$$

Cela montre que le mineur principal de $A^{(1)}$ d'ordre $l$ est égal à celui de $A_{,11}^{(2)}$ car $\det(M^{(1)})_{l,11}^{-1} = 1$.

On note encore $\alpha_1$, $\alpha_2$, $\ldots$, les mineurs principax de

$$A^{(2)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \ldots \\ 0 & a_{22}^{(2)} & \ldots \\ \vdots & \vdots & \vdots \end{pmatrix}$$

donc $a_{22}^2 \neq 0$ car $\alpha^2 = a_{11}^{(1)} a_{22}^{(2)}$.

**Etape $k$.** Hypothèse de récurence : $\alpha_k = a_{11}^{(1)} \dots a_{kk}^{(k)} \neq 0$ et les mineurs principaux de $A$ sont ceux de $A^{(k)}$. De plus

$$A^{(k)} = M^{(k)} \dots M^{(1)} A^{(1)}.$$

Le pivot $a_{kk}^{(k)}$ étant non nul, on a

$$\begin{array}{l} a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_i^{(k)} a_{1j}^{(k)} \quad \text{pour } i,j = k+1, \dots, n \\ a_{1j}^{(k+1)} = a_{1j}^{(k)} \end{array} \tag{3.4}$$

et

$$A^{(k+1)} = \begin{pmatrix} 1 & & & & & \\ 0 & 1 & & & & \\ \vdots & \vdots & \ddots & & & \\ 0 & 0 & 1 & & & \\ \vdots & \vdots & -m_{k+1}^{(k)} & 1 & & \\ \vdots & \vdots & \vdots & & \ddots & \\ \vdots & \vdots & -m_n^{(k)} & & & 1 \end{pmatrix} \begin{pmatrix} \star & \star & \dots & \star & \dots & \dots & \star \\ 0 & \star & \dots & \dots & \dots & \dots & \star \\ \vdots & \vdots & & & & & \\ 0 & 0 & & 0 & a_{kk}^{(k)} & \dots & \star \\ \vdots & \vdots & & \vdots & \vdots & \ddots & \star \\ 0 & 0 & & 0 & \star & \dots & \star \end{pmatrix}$$

donc $A^{(k+1)} = M^{(k)} A^{(k)} = M^{(k)} \dots M^{(1)} A^{(1)}$ et

$$(M^{(k)})^{-1} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ 0 & & m_i^{(k)} & & & \\ \vdots & & \vdots & \vdots & \ddots & \\ 0 & & & & & 1 \end{pmatrix}$$

Pour tout $l \geq 0$, on découpe à nouveau en 4 blocs comme précédemment et avec la même technique qu'à l'étape 1, on voit que le mineur principal de $A^{(K+1)}$ d'ordre $l$ est égal à celui de $A_{l,11}^{(k)}$ donc à celui de $A$. En procédant comme à l'étape 1, on obtient donc le point 2 de la proposition, avec

$$L = \left( M^{(n-1)} M^{(n)} \dots M^{(1)} \right)^{-1}.$$

Un calcul montre que

$$L = \begin{pmatrix} 1 & & & & \\ m_2^{(1)} & 1 & 0 & & \\ \vdots & \vdots & \ddots & & \\ & & & 1 & \\ m_n^{(1)} & m_n^{(2)} & \dots & m_n^{(n-1)} & 1 \end{pmatrix}$$

□

**Remarque 3.2.11.** *Dans ce cas, on voit que*

$$\det A = \Pi_{k=1}^{n} a_{kk}^{(k)}$$

*car il n'y a pas de permutation de lignes.*

On a de plus la

**Proposition 3.2.12.** *Soit $A \in M_n(\mathbb{R})$ une matrice inversible. Si tous les mineurs principaux sont non nuls, il existe une unique décomposition $LU$ de $A$ avec $l_{ii} = 1$.*

*Démonstration.* **Existence :** La démonstration du théorème précedent fournit la réponse si on montre que le produit de deux matrices triangulaires inférieures dont les éléments diagonaux valent 1 est encore une matrice triangulaire inférieure (déjà vu) dont les éléments diagonaux valent 1. Pour ce derneir point, il suffit de voir que

$$(l^1_{ij})_{1\leq i,j\leq n}(l^2_{ij})_{1\leq i,j\leq n} = (l^3_{ij})_{1\leq i,j\leq n}$$

avec (quand les matrices son triangulaires), $l^3_{ii} = l^1_{ii}l^2_{ii}$ ce qui prouve le résultat.
**Unicité :** Si on a deux décompositions

$$L_1U_1 = L_2U_2$$

et $\det A = \det L_i \det U_i \neq 0$, donc les deux matrices sont inversibles,

$$L_2^{-1}L_1 = U_2U_2^{-1}$$

. La matrice $L_2^{-1}L_1$ est triangulaire inférieure et $U_2U_2^{-1}$ est triangulaire supérieure, donc elles sont diagonales. Puisque $l_{ii} = 1$, cela fait l'identité. □

Une fois que la décomposition est faite, la résolution de $Ax = b$ se fait en résolvant d'abord $Ly = b$ puis $Ux = Y$ par descente/remonté. Ceci permet de traiter facilement les problèmes où plusieurs seconds membres successifs avec la même matrice $A$ apparaissent. Il faut alors calculer et stocker $L$ et $U$. Remarquons que le stockage de $U$ nécessite $\frac{n(n+1)}{2}$ tandis que celui de $L$ demande $\frac{n(n-1)}{2}$ : il est inutile de stocker les éléments diagonaux.

Nous admettrons le théorème suivant

**Théorème 3.2.2 (Décomposition de Choleski).** *Une matrice $A$ est symétrique définie positive si et seulement si il existe une matrice triangulaire inférieure $L$ telle que*

$$A = L\,L^T.$$

Plutôt que de faire cette démonstration, on va détailler l'algorithme de Choleski.

**Algorithme pour obtenir la décomposition de Cholweski.** Partant de $A = LL^T$, on a

$$a_{ij} = \sum_{k=1}^{j} l_{ik}l_{jk}, \qquad j \leq i$$

**Première colonne** . On a $a_{11} = l^2_{11} > 0$ donc on prend $l_{11} = \sqrt{a_{11}} > 0$.

**Pour** $i = 2, \ldots, n$ On a $a_{i1} = l_{i1}l_{11}$ donc $l_{i1} = \frac{a_{i1}}{l_{11}}$.

**Colonnes suivantes.** Supposons avoir calculé les colonnes $1, \ldots, j$. On a

$$a_{jj} = \sum_{k=1}^{j} l_{jk}^2 = l_{j1}^2 + \ldots + l_{j,j-1}^2 + l_{jj}^2$$

donc on prend

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}$$

ce qui est bien défini grâce au théorème.

Pour $i = j+1, \ldots, n$,

$$a_{ij} = l_{i1} l_{j1} + \ldots + l_{ij} l_{jj}$$

d'où

$$l_{ij} = \frac{1}{l_{jj}} \left( a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk} \right)$$

On remarque que $\det A = \Pi_{j=1}^{n} l_{jj}^2 \neq 0$ donc $l_{jj} \neq 0$.

**Exercice 3.2.3.** *Montrer que le nombre d'opération de la méthode de Choleski est de l'ordre $n^3/3$.*

Pour finir,

**Proposition 3.2.13.** *Si $A$ est une matrice symétrique de demi–largeur de bande $K$, alors $L$ est aussi de demi–largeur de bande $K$.*

*Démonstration.* $a_{i1} = 0$ si $i \geq K+1$ donc $l_{i1} = \frac{a_{i1}}{l_{11}}$ aussi.

Supposons que $l_{ij'} = 0$ pour $j' = 1, \ldots, j-1$ et $i \geq j' + K$. Pour $i \geq j + K$ ($\geq j' + K$, $j' = 1, \ldots, j-1$),

$$l_{ij} = \frac{1}{l_{jj}} \left( 0 - l_{i1} l_{j1} - \ldots - l_{i,j-1} l_{j,j-1} \right) = 0$$

car $l_{i1} = \ldots = l_{i,j-1} = 0$. □

**Exercice 3.2.4.** *Montrer que la matrice définie par*

$$A_n = \begin{pmatrix} 2 & -1 & 0 & \ldots & \\ 0 & \ddots & \ddots & & \\ \vdots & \vdots & \vdots & & \\ 0 & \ldots & & -1 & 2 \end{pmatrix}$$

*de taille $(n, n)$ est symetrique définie positive,* $\det A_n = n+1$ *et regarder Choleski pour cette matrice.*

## 3.3 Un aperçu des méthodes itératives.

Une alternative aux méthodes directes de type Gauss ou Choleski est l'utilisation de méthodes itératives. Ces méthodes sont utilisées en particulier quand on a un stockage creux qui n'est pas du type bande. Elles ont aussi des version par blocs.

### 3.3.1 Méthodes de décomposition.

On veut résoudre $Ax = b$. On décompose $A$ sous la forme $A = M - N$ avec $M$ inversible. On pose

$$\begin{cases} Mx^{(k+1)} = Nx^{(k)} + b \\ x^{(0)} \text{ quelconque dans } \mathbb{R}^n \end{cases} \tag{3.5}$$

Si $M$ est facile à inverser (diagonale ou triangulaire), on a $x^{(k+1)}$ très facilement. Si $x^{(k)} \to x$ quand $k \to +\infty$ alors $x$ est solution de $Ax = b$.

**Proposition 3.3.1.** *La méthode converge si et seulement si $\rho(M^{-1}N) < 1$.*

*Démonstration.* **Si la méthode converge**. On a alors

$$\begin{aligned} Mx^{(k+1)} &= Nx^{(k)} + b \\ Mx &= Nx + b \end{aligned}$$

En soustrayant les deux égalités, on a

$$M(x^{(k+1)} - x) = N(x^{(k)} - x)$$

et donc

$$x^{(k+1)} - x = M^{-1}N((x^{(k)} - x)) = \ldots = \left(M^{-1}N\right)^k (x_0 - x).$$

Si la méthode converge, nécessairement $x^{(k+1)} - x \to 0$. Le vecteur $x_0$ étant arbitraire, il faut donc $\left(M^{-1}N\right)^k \to 0$ ce qui ne peut se faire que si $\rho(M^{-1}N) < 1$.

**Réciproque.** En effectuant un calcul du même type, on a quelques soient $p$ et $q$,

$$x^{(p)} - x^{(q)} = \left(M^{-1}N\right)^{p-q} (x^{(1)} - x_0),$$

soit

$$||x^{(p)} - x^{(q)}|| \leq ||\left(M^{-1}N\right)^{p-q}|| \; ||x^{(1)} - x_0||.$$

Soit $\varepsilon > 0$. Si $\rho(M^{-1}N) < 1$ alors $||\left(M^{-1}N\right)^N|| \leq \varepsilon$ dès que $N$ est assez grand. Donc si $p - q$ est assex grand, $||x^{(p)} - x^{(q)}|| \leq \varepsilon ||x^{(1)} - x_0||$. ce qui montre que la suite est une suite de Cauchy : elle est donc convergente. □

### 3.3.2 Méthode de Jacobi

On effectue la décomposition suivante

$$A = D - M$$

où $D$ est la matrice diagonale formée des éléments diagonaux de $A$. S'ils sont tous non nuls, $D$ est inversible. La méthode de Jacobi est définie par

$$Dx^{(k+1)} = Mx^{(k)} + b \tag{3.6}$$

soit

$$\begin{array}{rcccccccc} a_{11}x_1^{(k+1)} & = & & - & a_{12}x_2^2 & - & \dots & + & a_{1n}x_n^{(n)} \\ \vdots & & & & & & & & \\ a_{n1}x_n^{(k+1)} & = & a_{11}x^{(k)} & - & a_{n2}x_2^2 & - & \dots & + & a_{1n-1}x_{n-1}^{(n)} \end{array} \tag{3.7}$$

Une condition suffisante de convergence de la méthode est que la matrice soit à diagonale strictement dominante.

**Définition 3.3.2.** *On dit que $A = (a_{ij})$ est à diagonale dominante stricte si quelque soit $i$,*

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|.$$

**Proposition 3.3.3.** *Si $A$ est à diagonale dominante stricte, alors $\rho(D^{-1}M) < 1$ et la méthode de Jacobi converge.*

*Démonstration.* Pour montrer cela, il suffit de trouver une norme matricielle telle que $||D^{-1}M|| < 1$ car on sait que $\rho(D^{-1}M) \leq ||D^{-1}M||$.

La norme adaptée est la norme $L^\infty$. On sait (voir TD) que

$$||B||_\infty = \max_i \left( \sum_j |b_{ij}| \right).$$

Ici, $B = D^{-1}M$ et

$$||D^{-1}M|| = \max_i \left( \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} \right) < 1.$$

□

La méthode de Jacobi converge dans bien d'autre cas que nous n'étudierons pas. Voir par exemple la section 7.1.7 ou [2, 1].

## 3.4 Méthode de Gauss–Seidel

Ici, on définit la méthode par le découpage $A = M - N$ avec

$$M = \begin{pmatrix} a_{11} & & \\ \vdots & \ddots & 0 \\ a_{n1} & \dots & a_{nn} \end{pmatrix}$$

En fait, elle consiste à modifier la méthode de jacobi en utilisant à la ligne $i$ les $x_1^{(k+1)}, \dots, x_{i-1}^{(k+1)}$ au lieu de $x_1^{(k)}, \dots, x_{i-1}^{(k)}$, soit

$$\begin{aligned} x_1^{(k+1)} &= -\frac{1}{a_{11}} \sum_{j=2}^{b} a_{1i} x_j^{(k)} + \frac{b_1}{a_{11}} \\ &\vdots \\ x_i^{(k+1)} &= \frac{1}{a_{ii}} \left[ -\sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right] + \frac{b_i}{a_{ii}} \\ &\vdots \end{aligned}$$

Cette méthode permet de ne stocker qu'un seul vecteur de taille $n$ au lieu de deux vecteurs de taille $n$ pour Jacobi. Quand elle converge, elle converge plus vite que Jacobi. Par exemple, dans le cas de matrices à diagonale dominante stricte, les deux méthodes convergent, mais Gauss–Seidel converge beaucoup plus vite.

## 3.5 Méthodes de descente

L'idée de base est la suivante. Pour simplifier, on suppose $A$ symétrique définie positive. Considérons

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

pour qui

$$J'(x) = Ax - b$$

et $J'' = A$ qui est définie positive : $J$ est strictement convexe et admet donc un minimum global. Il y a donc un vecteur $\bar{x}$ tel que $A\bar{x} = b$.

Minimiser $J$ revient à résoudre le système $Ax = b$. On sait que $\nabla J(x)$ est orthogonal à la ligne de niveau $J(y) = J(x)$ passant par $x$. Soit $x_0$ fixé. Si $x_k$ est connu, on cherche $x_{k+1}$ tel que

$$x_{k+1} - x_k = -\theta_k J'(x_k)$$

et tel que $J(x_{k+1})$ soit le plus petit possible : on doit minimiser par rapport au paramètre $\theta_k$.

Notons $p_k = J'(x_k) = Ax_k - b$ et

$$J(x_{k+1}) - J(x_k) - \theta_k < p_k, p_k > + \frac{1}{2}\theta_k^2 < Ap_k, p_k >= f(\theta_k)$$

et

$$f'(\theta_k) = \theta_k < Ap_p, p_k > -||p_k||^2$$

Donc $f$ atteint son minimum pour

$$\theta_k = \frac{||p_k||^2}{< Ap_k, p_k >}$$

d'où

$$\begin{aligned} p_k &= Ax_k - b \\ x_{k+1} &= x_k - \frac{||p_k||^2}{< Ap_k, p_k >}\Big(Ax_k - b\Big) \end{aligned} \tag{3.8}$$

définit la méthode itérative. On a le théorème suivant

**Théorème 3.5.1.** *Soit $A$ symétrique définie positive. La méthode (3.8) converge vers la solution de $Ax = b$.*

# Chapitre 4

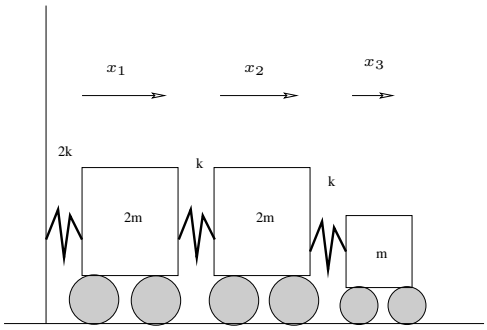
# Calcul de valeurs propres et de vecteurs propres

Dans ce chapitre, on ne va qu'effleurer le sujet.

## 4.1 Introduction : pourquoi calculer des valeurs propres ?

La recherche de valeurs propres et de vecteurs propres est un problème qui intervient naturellement dans l'étude de la dynamique des structures. On connait par exemple l'histoire de ce pont suspendu qui s'est écroulé au passage d'un régiment : le pas cadencé des soldats avait exité les fréquences propres du système mécanique que constitue le pont.

Considérons par exemple le système




Le système est constitué de trois corps rigides de masse $2m$, $2m$ et $m$. Ils sont reliés à un mur et entre eux par des ressorts de constante de raideur $2k$, $k$ et $k$.

Le mouvement du système est décrit par

$$\begin{aligned} 2mx_1'' + 2kx_1 + k(x_1 - x_2) &= 0 \\ 2mx_2'' + k(x_2 - x_1) + k(x_2 - x_3) &= 0 \\ mx_3'' + k(x_3 - x_2) &= 0 \end{aligned}$$

où $x_1$, $x_2$ et $x_3$ désignent les positions des centres de gravité. On peut réecrire ce système comme

$$MX'' + KX = 0$$

avec

$$X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \qquad M = \begin{pmatrix} 2m & 0 & 0 \\ 0 & 2m & 0 \\ 0 & 0 & m \end{pmatrix}, \qquad K = k \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Si on cherche des solutions de la forme $X(t) = \sin(\omega t)\hat{X}$ où $\hat{X}$ est un vecteur indépendant du temps, il faut que $\omega$ et $\hat{X}$ vérifient

$$(K - \omega^2 M)\hat{X} = 0$$

ce qui est un problème de valeurs propres pour la matrice $KM^{-1}$.

Il y a de très nombreux exemple en mécanique et en physique où le problème se ramène à la recherche de valeurs propres de système qui sont souvent de très grande taille.

## 4.2 Conditionnement du problème de valeurs propres

Comme pour la résolution de systèmes linéaires, il est important de savoir si des petites variations sur les données peuvent ou non entrainer d'importantes variations sur les résultats. L'esemple suivant est assez saisissant.

Considérons la matrice d'ordre $n$

$$A = \begin{pmatrix} 0 & 0 & \dots & \varepsilon \\ 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ \vdots & & & 0 \\ 0 & & 1 & 0 \end{pmatrix}$$

Son polynôme caractéristique est

$$P(\lambda) = \det(A - \lambda Id) = \lambda^n + (-1)^n \varepsilon.$$

Pour $\varepsilon = 0$, toutes les valeurs propres sont nulles. Par contre si $n = 40$ et $\varepsilon = 10^{-40}$, les valeurs propres ont toutes le même module, $10^{-1}$ ... Le problème est mal conditionné.

Plus généralement, on a les résultats suivants concernant le conditionnement d'un problème de valeur propres.

**Proposition 4.2.1.** *Soit $A$ une matrice diagonalisable, $P$ une matrice telle que $P^{-1}AP = D = diag(\lambda_i)$, et $||\ .\ ||$ une norme matricielle telle que pour toute matrice diagonale, $||diag(d_i)|| = \max_i |d_i|$.*

*Alors, pour toute matrice $\delta A$, l'ensemble des valeurs propres de $A + \delta A$ noté $Sp(A + \delta A)$ vérifie*

$$Sp(A + \delta A) \subset \cup_i D_i$$

*où*

$$D_i = \{z \in \mathbb{C}, |z - \lambda_i| \leq cond(P)\ ||\delta A||\}.$$

*Démonstration.* Soit $\lambda \in \mathrm{Sp}(A + \delta A)$. Si $\lambda = \lambda_j$ pour un indice $j$, le résultat est évident, sinon $\lambda \neq \lambda_j$ quelque soit $j$. La matrice $D - \lambda Id$ est donc inversible et on peut écrire

$$P^{-1}(A + \delta A - \lambda Id)P = D - \lambda Id + P^{-1}\delta AP$$
$$(D - \lambda Id)\left\{ Id - (D - \lambda Id)^{-1}P^{-1}\delta AP \right\}.$$

La matrice $A + \delta A - \lambda Id$ étant singuliere, la matrice $Id - (D - \lambda Id)^{-1} P^{-1} \delta A P$ l'est aussi et donc

$$1 \leq ||(D - \lambda Id)^{-1} P^{-1} \delta A P||$$

sinon, elle serait inversible (voir le corollaire 2.5.5). Comme

$$||(D - \lambda Id)^{-1}|| = \frac{1}{\min_i |\lambda_i - \lambda|},$$

il existe au moins un indice $i_0$ tel que

$$|\lambda - \lambda_{i_0}| \leq ||P||\, ||P^{-1}||\, ||\delta A||,$$

d'où le résultat. □

**Remarque 4.2.2.**

1. *L'hypothèse $||diag(d_i)|| = \max_i |d_i|$ est vérifiée par les normes $||\ .||_1$, $||\ .||_2$ et $||\ .||_\infty$.*
2. *Le conditionnement du problème est mesuré par $cond\ (P)$. En particulier, les matrices diagonalisables par des matrices orthogonales (unitaires) dont le conditionnement $L^2$ est égal à 1, comme les matrices symétriques ou les matrices normales, sont bien conditionnées.*

En ce qui concernes les matrices symétriques, on a un résultat beaucoup plus précis.

**Proposition 4.2.3.** *Soient $A$ et $B$ deux matrices symétriques de valeurs propres $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_n$ et $\beta_1 \leq \beta_2 \leq \ldots \leq \beta_n$. Soient $\gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_n$ les valeurs propres de $C = A + B$. Alors on a pour tout $i$, $1 \leq i \leq n$,*

$$\beta_1 \leq \gamma_i - \alpha_i \leq \beta_n.$$

La démonstration repose sur le théoreme min–max de Courant Fisher qui caractérise les valeurs propres d'une matrice symétrique,

**Théorème 4.2.1 (Courant-Fisher).** *Soit $A$ une matrice symétrique dont les valeurs propres sont rangées dans l'ordre $\lambda_n \leq \lambda_{n-1} \leq \ldots \leq \lambda_1$. Alors*

$$\lambda_k = \max_{V\ tel\ que\ dim\ V = k} \min_{x \in V - \{0\}} \frac{< Ax, x >}{< x, x >}$$

*et*

$$\lambda_k = \min_{W\ tel\ que\ dim\ W = n-k+1} \max_{x \in W - \{0\}} \frac{< Ax, x >}{< x, x >}$$

On a changé l'ordre des valeurs propres pour facilité les notations de la démonstration suivante.

*Démonstration.* Montrons d'abord la première inégalité. Soit $V_k$ l'espace de dimension $k$ engendré par les vecteurs propres orthonomés $(u_1, u_2, \ldots, u_k)$. Pour $x \in V_k - \{0\}$, i.e $x = \sum_{i=1}^{k} \alpha_i u_i$ avec $\sum_{i=1}^{k} |\alpha_i|^2 \neq 0$, on a

$$\frac{< Ax, x >}{< x, x >} = \frac{\sum_{i=1}^{n} \lambda_i |\alpha_i|^2}{\sum_{i=1}^{k} |\alpha_i|^2} \geq \lambda_k.$$

En prenant $x = u_k$, on a

$$\frac{< Ax, x >}{< x, x >} = \lambda_k$$

donc

$$\lambda_k = \min_{x \in V_k - \{0\}} \frac{< Ax, x >}{< x, x >}.$$

Soit maintenant un sous espace vectoriel de dimension $k$ et soit $V_k^N$ l'espace vectoriel engendré par $\{u_k, \ldots, u_N\}$ qui est de dimension $N - k + 1$. Comme dim $V +$ dim $V_k^N = N + 1 > N$, il existe $x \in V \cap V_k^N$ qui s'écrit

$$x = \sum_{i=k}^{N} \alpha_i u_i$$

vérifie

$$\frac{< Ax, x >}{< x, x >} = \frac{\sum_{i=k}^{N} \lambda_i |\alpha_i|^2}{\sum_{i=k}^{N} |\alpha_i|^2} \leq \lambda_k.$$

Donc

$$\min_{x \in V_k - \{0\}} \frac{< Ax, x >}{< x, x >} \leq \lambda_k$$

pour $V$ quelconque de dimension $k$. Ainsi, en combinant les deux résultats, on obtient la premiere égalité.

Démontrons la second inégalité. Soit $W$ de dimension $N-k+1$, $W \cap V_k \neq \{0\}$ puisque dim $W +$ dim $V_k = N + 1 > N$. On prend $x_k \in W \cap V_k$ donc comme précédemment,

$$\max_{x \in W - \{0\}} \frac{< Ax, x >}{< x, x >} \geq \frac{< Ax_k, x_k >}{< x_k, x_k >} \geq \lambda_k.$$

Puis, pour $W = V_{k-1}^{\perp} = \{u_k, \ldots, u_n\}$, on a

$$\max_{x \in W - \{0\}} \frac{< Ax, x >}{< x, x >} = \lambda_k$$

si bien qu'en combinant les deux résultats, on obtient la seconde égalité. □

*Démonstration de la proposition 4.2.3.* On applique le théorème de Courant–Fischer au sous espace engendré par $w_1, \ldots, w_i$ où $w_1, \ldots, w_n$ est une base orthonormale des vecteurs propres de $A$. On a alors

$$\gamma_i \leq \max_{x \in U_i, x \neq 0} \frac{< Cx, x >}{< x, x >} \leq \max_{x \in U_i, x \neq 0} \frac{< Ax, x >}{< x, x >} + \max_{x \in U_i, x \neq 0} \frac{< Bx, x >}{< x, x >} \leq \alpha_i + \beta_n.$$

On obtient ainsi l'inégalité de droite. Si on prend $D = -B$ et qu'on applique le même résultat à $A = C + D$, on obtient

$$\alpha_i \leq \gamma_i + \delta_n.$$

Comme $\delta_n = -\beta_1$, on obtient le résultat voulu. □

En conclusion, le problème du calcul des valeurs propres est plus facile à résoudre dans le cas d'une matrice symétrique que dans le cas d'une matrice diagonalisable quelconque. Dans le cas d'une matrice non diagonalisable, il peut être très instable comme le montre l'exemple plus haut.

## 4.3 Calcul de la plus grande valeur propre

Le problème général de la détermination des valeurs propres d'une matrice diagonalisable ne sera pas abordé ici. Un algorithme qui permet d'effectuer le calcul est l'algorithme de Givens–Householder qui est décrit dans [1] entre autres. Ici, on ne s'intéressera qu'à la détermination de la plus grande valeur propre d'une matrice diagonalisable au moyen de la méthode de la puissance itérée.

Il s'agit de l'algorithme

$$\begin{aligned} &q_0 \in \mathbb{R}^n \text{ quelconque tel que } ||q_0|| = 1 \\ &x^{(k)} = Aq^{(k-1)}, k \geq 1 \\ &q^{(k)} = \frac{x^{(k)}}{||x^{(k)}||}. \end{aligned} \tag{4.1}$$

Pour démontrer la convergence de la méthode, il faut faire une hypothèse essentielle

La valeur propre de plus grande module est unique (mais pas nécessairement simple).

Si ce n'est pas le cas, on peut encore prouver la convergence (qui est cependant plus lente).

Si on appelle «facteur de convergence d'une suite $y^{(k)}$ de limite $y$» le nombre

$$\rho = \lim_{k \to +\infty} \frac{||y^{(k+1)} - y||}{||y^{(k)} - y||}.$$

On a

**Proposition 4.3.1.** *Soit $A$ une matrice $(n, n)$ diagonalisable dont la valeur propre de plus grand module $\lambda_1$ est unique. Soit $q_0$ un vecteur de $\mathbb{R}^n$ qui n'est pas orthogonal aux sous–espace propre à gauche associé à $\lambda_1$, alors la suite définie par (4.1) vérifie*

1. $q = \lim_{k \to +\infty} \left( \frac{\lambda_1}{|\lambda_1|} \right)^k q^{(k)}$ *est un vecteur propre de norme 1 associé à $\lambda_1$,*
2. $\lim_{k \to +\infty} ||Aq^{(k)}|| = |\lambda_1|$,
3. *Si $q_j^{(k)}$ $(1 \leq j \leq n)$ est la $j^e$coordonnée de $q^{(k)}$ supposée non nulle, $\lim_{k \to +\infty} \frac{x_j^{(k+1)}}{q^{(k)}} = \lambda_1$.*

*De plus, le facteur de convergence de toute ces suites est $||\lambda_{p+1}/\lambda_1|$ où $p$ est la multiplicité de la valeur propre $\lambda_1$.*

Rappelons que si $v_j$ est vecteur propre à gauche de $A$ associé à la valeur propre $\lambda_j$ si $v_j^T A = \lambda_j v_j^T$. Si $u_i$ est vecteur propre (à droite) associé à la valeur propre $\lambda_i$, si $\lambda_i \neq \lambda_j$, alors $v_j^T u_i = 0$. De plus si $A$ est symétrique, les vecteurs propres à gauche et à droite coincident.

*Démonstration.* **Supposons dans un premier temps que $\lambda_1$ est une valeur propre simple.** $A$ étant diagonalisable, on note $u_1, \ldots, u_n$ une base de valeur propres associés à $A$ associé aux valeurs

propres $\lambda_1, \ldots, \lambda_n$ avec $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$. Par hypothèse,

$$q^{(0)} = \sum_{i=1}^{n} \alpha_i u_i$$

et si on note $v_i$ les vecteurs propres à gauche de $A$ vérifiant $v_i^T u_i = 1$, on a

$$\alpha_1 = v_1^T q^{(0)} \neq 0.$$

Soit $x^{(1)} = Aq^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i u_i$, alors $q^{(1)} = x^{(1)}/||x^{(1)}|| = Aq^{(0)}/||Aq^{(0)}||$. Montrons par récurence que pour tout $k$, $q^{(k)} = A^k q^{(0}/||Aq^{(0)}||$. On vient de voir que le résultat est vrai pour $k = 1$. Supposons le vrai pour $k$, et montrons le pour $k+1$. On a

$$\begin{aligned} q^{(k+1)} &= \frac{Aq^{(k)}}{||Aq^{(k)}||} \\ &= A\left(\frac{A^k q^{(k)}}{||A^k q^{(0)}||}\right)\left(\frac{||A^k(q^{(0)})||}{||A^k q^{(0)}||}\right)^{-1} \\ &= \frac{A^{k+1} q^{(0)}}{||A^{(k+1)} q^{(0)}||} \end{aligned}$$

Or, $A^k q^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i^k u_i$. Comme $\alpha_1 \neq 0$, on peut écrire

$$\begin{aligned} A^k q^{(0)} &= \alpha_i \lambda_i^k \left(u_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1}\left(\frac{\lambda_i}{\lambda_1}\right)^k u_i \right. \\ &= \alpha_i \lambda_i^k \left(u_1 + e^{(k)}\right) \end{aligned}$$

Par hypothèse $|\lambda_i| < |\lambda_1|$ pour $i \geq 2$ et donc

$$\lim_{k \to +\infty} (\lambda_i/\lambda_1)^k = 0, \qquad i = 2, \ldots, n.$$

Si $\alpha_2 \neq 0$, alors $e^{(k)}$ tend vers 0 comme $(\lambda_2/\lambda_1)^k$.

Il en résulte que

$$\begin{aligned} ||A^k q^{(0)}|| &= |\alpha_1|\ |\lambda_1|^k ||u_1 + e^{(k)}|| \\ &\quad |\alpha_1|\ |\lambda_1|^k (||u_1|| + \epsilon_k) \qquad \text{où } \epsilon_k = O\left((\lambda_2/\lambda_1)^k\right). \end{aligned}$$

On a donc

$$q^{(k)} = \frac{\alpha_1 \lambda_1^k (u_1 + e^{(k)})}{|\alpha_1|\ |\lambda_1|^k (||u_1|| + \epsilon_k)}$$

et donc

$$\lim_{k \to +\infty} \left(\frac{\lambda_1}{|\lambda_1|}\right)^k q^{(k)} = \frac{\alpha_1}{|\alpha_1|} \frac{u_1}{||u_1||}$$

d'où le premier point.

De même,

$$||Aq^{(k)}|| = \frac{||A^{k+1}q^{(0)}||}{||A^k q^{(0)}||}$$
$$= \frac{|\alpha_1|\ |\lambda_1|^{k+1}}{|\alpha_1|\ |\lambda_1|^k} \frac{||u_1|| + \epsilon_k}{||u_1|| + \epsilon_{k+1}}$$

d'où $\lim_{k\to+\infty} ||Aq^{(k)}|| = |\lambda_1|$ d'où le point 2.

Enfin, si $(u_1)_j \neq 0$, si $k$ est assez grand $(u_1)_j + e_j^{(k+1)} \neq 0$, on a

$$\frac{(Aq^{(k)})_j}{q_j^{(k)}} = \frac{(A^{k+1}q^{(0)})_j}{(A^k q^{(0)})_j}$$
$$\frac{\alpha_1\lambda_1^{k+1}\Big((u_1)_j + e_j^{(k+1)}\Big)}{\alpha_1\lambda_1^{k+1}\Big((u_1)_j + e_j^{(k)}\Big)}$$

et

$$\lim_{k\to+\infty} \frac{(Aq^{(k)})_j}{q_j^{(k)}} = \lambda_1$$

Pour chacune de ces suites, le facteur de convergence est $\lambda_2/\lambda_1$ si $\alpha_2 \neq 0$.

**Cas général où la valeur propre de plus grand module est unique mais de multiplicité $m$.** On a alors $\lambda_1 = \lambda_2 = \ldots = \lambda_m$. La démonstration se fait suivant le même schéma. On a

$$A^k q^{(0)} = \lambda_1^k \left( \sum_{i=1}^{m} \alpha_i u_i + \sum_{i=m+1}^{n} \alpha_i \big(\frac{\lambda_i}{\lambda_1}\big)^k u_i \right)$$
$$\lambda_i^k \Big( u + e^{(k)} \Big).$$

Par hypothèse, au moins un des $\alpha_i$, $i = 1, m$, est non nul et $Au = \lambda_1 u$. On a alors $e^{(k)}$ tend vers 0 comme $(\lambda_{p+1}/\lambda_1)^k$ si $\alpha_{p+1} \neq 0$. □

**Remarque 4.3.2.** *Bien souvent, en calcul des structures par exemple, on a besoin de connaître la plus petite valeur propre. On applique cette méthode à $A^{-1}$ dont la plus grande valeur propre est $1/\min(\lambda(A))$.*

## 4.4 Calcul de vecteurs propres : méthode de la puissance inverse

Elle permet de calculer rapidement une approximation d'un vecteur propre assocé à une valeur propre dont on connaît déjà une bonne approximation. C'est une méthode itérative définie par

$$u^{(0)} \text{ arbitraire,}$$
$$(A - \tilde{\lambda} Id)u^{(k+1)} = u^{(k)}$$

où $\tilde{\lambda}$ est une approximation dúne valeur propre de $A$.

**Proposition 4.4.1.** *Soit $A$ une matrice diagonalisable et $\lambda$ une valeur propre quelconque de cette matrice. On suppose que $\tilde{\lambda}$ vérifie*

$$\tilde{\lambda} \neq \lambda \text{ et } |\tilde{\lambda} - \lambda| < |\tilde{\lambda} - \mu| \qquad \mu \in Sp(A) - \{\lambda\}.$$

*Alors*

$$\lim_{k \to +\infty} \frac{(\lambda - \tilde{\lambda})^k}{|\lambda - \tilde{\lambda}|^k} \frac{u^{(k)}}{||u^{(k)}||} = q$$

*où $q$ est un vecteur propre associé à la valeur propre $\lambda$.*

*Démonstration.* Soit $\mu_i$, $1 \leq i \leq m$ les valeurs propres de $A$ qui sont différentes de $\lambda$ (multiplicité comprise), et notons $q_i$ les vecteurs propres associés. On écrit

$$u^{(0)} = \tilde{q} + \sum_{i=1}^{m} \alpha_i q_i$$

avec $\tilde{q} \neq 0$ vecteur propre de $A$ associé à $\lambda$. On a alors

$$u^{(k)} = \frac{\tilde{q}}{(\lambda - \tilde{\lambda})^k} + \sum_{i=1}^{m} \alpha_i \frac{q_i}{(\mu_i - \tilde{\lambda})^k}$$

ou encore

$$(\lambda - \tilde{\lambda})^k u^{(k)} = \tilde{q} + \sum_{i=1}^{m} \alpha_i \left[ \frac{\lambda - \tilde{\lambda}}{\mu_i - \tilde{\lambda}} \right]^k q_i.$$

Comme $|\tilde{\lambda} - \lambda| < |\tilde{\lambda} - \mu_i|$, on a donc

$$(\tilde{\lambda} - \lambda)^k u^{(k)} = \tilde{q} + \varepsilon_k$$

avec $\lim_{k \to +\infty} \varepsilon_k = 0$.

En continuant comme dans la proposition précédente, on obtient le résultat. □

**Remarque 4.4.2.**

1. *A chaque itération, on a besoin de résoudre un système linéaire qui ne diffère que par son second membre.*
2. *La convergence est d'autant plus rapide que $|\lambda - \tilde{\lambda}$ est petit. Cependant, il ne faut pas que $\lambda - \tilde{\lambda} = 0$ car la matrice $A - \tilde{\lambda} Id$ est alors singulière.*
3. *Cette méthode est utilisable pour accélérer une autre méthode, par exemple la méthode de Givens–Householder, voir [1] par exemple.*

# Chapitre 5

# Interpolation de fonctions.

## 5.1 Introduction

Dans ce chapitre, on s'intéresse à l'approximation de fonctions. Deux cas de figures se présentent :

1. On a un résultat de calcul ou d'expérience sous la forme d'un ensemble fini de points du plan $\{(x_i, y_i), i = 1, \ldots, N\}$. On cherche la «meilleure fonction» $f$ telle que $y_i = f(_i)$ (ou $y_i \simeq f(x_i)$) $i = 1, \ldots, N$. C'est un problème d'interpolation.
2. On a un problème où l'inconnue est une fonction. Pour la calculer, il faut chercher, a priori sous une forme donnée, calculable en un nombre fini d'opération.

Les fonctions polynômiales (ou polynômiale par morceau) sont de bons outils pour aborder ce problème.

On note par $C^k(]a, b[)$ l'ensemble des fonctions $k$ fois dérivables sur $]a, b[$ telle que $f$ est continue sur $[a, b]$. On note

$$C^\infty(]a, b[) = \cap_k C^k(]a, b[)$$

l'ensemble des fonctions indéfiniment différentiables sur $]a, b[$.

On dit qu'une fonction est analytique sur $]a, b[$ si quelque soit $x_0 \in ]a, b[$, il existe $\epsilon > 0$ tel que quelque soit $x \in ]x - \epsilon, x_0 + \epsilon[$,

$$f(x) = \sum_{i=0}^{\infty} a_n (x - x_0)^n.$$

Un tel développement n'est généralement pas global.

Si $f$ est analytique sur $]a, b[$, alors $f \in C^\infty(]a, b[)$. La réciproque est fausse. Un contre–exemple est donné par

$$f(x) = \begin{cases} 0 & \text{si } x <\leq 0 \\ e^{-1/x^2} & \text{sinon.} \end{cases}$$

Un polynôme est une série entière n'ayant qu'un nombre fini de termes "il existe $a_0, \ldots, a_n$ tel que

$$P(X) = \sum_{k=0}^{n} a_k X^k.$$

Le degré de $P$ est le plus petit entier $n$ tel que $a_k = 0$ si $k \geq n$. La fonction polynômiale associée à $P$, encore notée $P$, est l'application

$$x \mapsto \sum_{k=0}^{n} a_k x^k.$$

On note par $\mathbb{P}_k$ l'ensemble des polynômes de degrè $\leq k$. C'est un espace vectoriel sur $\mathbb{R}$ de dimension $n+1$ dont une base est $\{1, X, \ldots, X^n\}$.

## 5.2 Fonctions et polynômes.

Revenons aux deux cas de figures évoqués dans l'introduction : $f$ est connue seulement sur un nombre fini de points, $f$ est inconnue.

Dans ces deux cas, on va essayer d'utiliser des polynôme. On peut

1. Dans le premier cas, soit
   (a) Interpoler $f$ par un polynôme, c'est à dire trouver un polynôme $p$ tel que

   $$\text{quel que soit } i \in \{1, \ldots, N\}, \qquad p * x_i) = y_i.$$

   Nécessairement, le degré de $p$ est $\geq N - 1$.
   (b) Soit trouver un polynôme $q \in \mathbb{P}_k$ tel que l'erreur soit minimale, par exemple

   $$\text{quel que soit } \tilde{q} \in \mathbb{P}_k, \sum_{i=1}^{N} |y_i - \tilde{q}(x_i)|^2 \geq \sum_{i=1}^{N} |y_i - q(x_i)|^2.$$

   Ici, il n'y a pas de condition sur le degré $k$.
   (c) ou encore chercher chercher une fonction de la régularité voulue ($C^1$ par exemple) et qui est un polynôme sur chaque intervalle $[x_i, x_{i+1}]$ (splines, interpolation d'Hermite)
2. Dans le deuxième case (fonction inconnue), on peut soit
   (a) chercher $f$ sur un nombre finis de points puis appliquer les techniques ci–dessus,
   (b) chercher $f$ dans un espace vectoriel de dimension finie (construit généralement avec des polynômes) de la forme

   $$f = \sum_{k=1}^{N} \alpha_k e_k$$

   où $\{e_1, \ldots, e_k\}$ est une base de $V$. Alors

   $$\forall x \in I, \qquad f(x) = \sum_{j=1}^{N} \alpha_j e_j(x).$$

## 5.3 Interpolation de Lagrange

Soit $f$ une fonction définie sur $\{x_1, \ldots, x_n\}$, les points $x_i$ étant deux à deux distincts. Existe–t–il $\varphi \in \mathbb{P}_{n-1}$ tel que

$$\varphi(x_1) = f(x_1), \ldots, \varphi(x_n) = f(x_n)\,?$$

**Lemme 5.3.1.** *Soit $N \in \mathbb{N}$. Soient $t_0, \ldots, t_N$ des points deux à deux distincts de $\mathbb{R}$ et $b_0, \ldots, b_N$ des réels. Il existe un polynôme $\varphi \in \mathbb{P}_N$ tel que $\varphi(t_i) = b_i$, $i = 1, \ldots, N$.*

*Démonstration.* On chercher $\varphi$ de la forme

$$\varphi(x) = \sum_{i=0}^{N} a_i x^i$$

avec la contrainte

$$\varphi(t_j) = \sum_{i=0}^{N} a_i t_j^i = b_j.$$

On obtient un système linéaire dont la matrice est

$$U = \begin{pmatrix} 1 & t_0 & \ldots & t_0^N \\ \vdots & \vdots & & \vdots \\ 1 & t_N & \ldots & t_N^N \end{pmatrix}$$

et le second membre est

$$b \begin{pmatrix} b_0 \\ \vdots \\ b_N \end{pmatrix}.$$

Le déterminant de cette matrice (dite de Vandermonde) est

$$\det U = \Pi_{i \neq j}(x_i - x_j).$$

Il est non nul, d'où le résultat. □

Comment le calculer ? Une réponse (non satisfaisante) est donnée par les résultats suivants.

**Définition 5.3.2.** *Le polynôme $\varphi \in \mathbb{P}_{m-1}$ tel que $\varphi(x_i) = f(x_i)$ pour $i = 1, \ldots, m$ est le polynôme d'interpolation de Lagrange de $f$ aux points $x_1, \ldots, x_m$.*

**Proposition 5.3.3.** *Soit $\ell_i \in \mathbb{P}_{m-1}$ tel que $l_i(x_j) = \delta_{ij}$. $\{\ell_1, \ldots, \ell_m\}$ est une base de $\mathbb{R}^{m-1}$, appelée base de Lagrange aux points $x_1, \ldots, x_m$. On a*

$$\ell_j(x) = \frac{\Pi_{i \neq j}(x - x_i)}{\Pi_{i \neq j}(x_j - x_i)}$$

*et*

$$\varphi(x) = \sum_{i=1}^{m} f(x_i)\ell_i(x).$$

*Démonstration.* Le degrè de $\ell_i$ est $m - 1$ donc $\ell_i \in \mathbb{P}_{m-1}$. On a $\ell_i(x_j) = \delta_{ij}$.

C'est bien une base car d'une part $\dim \mathbb{P}_{m-1} = m = \text{ card } \{\ell_1, \ldots, \ell_m\}$ et d'autre part c'est une famille libre

$$\sum_{i=1}^{m} \alpha_i \ell_i = 0 \rightarrow \forall j, \sum_{i=1}^{m} \alpha_i \ell_i(x_j) = \alpha_j = 0.$$

□

**Remarque 5.3.4.** *La définition du polynôme d'interpolation ne fait pas référence à un ordre quelconque des points $x_i$. En particulier, on ne suppose pas que $x_1 < x_2 < \ldots < x_m$.*

### 5.3.1 Forme de Newton, différences divisées

La formule utilisant les $\ell_i$ n'est pas toujours pratique, en particulier si on veut ajouter des points. On obtient une formule plus utilisable en remarquant qu'une base de $\mathbb{P}_{m-1}$ est aussi constitué de

$$\{1, x - x_1, \ldots, (x - x_1)\ldots(x - x_i), \ldots, (x - x_1)\ldots(x - x_N)\}$$

L'idée est donc d'écrire

$$\varphi(x) = \sum_{i=0}^{m} \alpha_i \Pi_{j=1}^{i}(x - x_j) \tag{5.1}$$

et d'evaluer successivement les coefficients. Il s'agit, d'une certaine manière, d'une triangulation du système.

**Définition 5.3.5.** *On note $f[x_1, \ldots, x_k]$ le coefficient $\alpha_k$ du développement (5.1).*

Le calcul se fait de proche en proche :
- degrè 0 : $f[x_1] = fx_1$,
- degrè 1 :
$$\varphi_1(x) = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1),$$
donc $f[x_1, x_2] = \frac{f(x_2) - f(x_1)}{x_2 - x_1}$.
- degrè 2. $\varphi_2 - \varphi_1$ est de degré 2 et s'annule en $x_1$ et $x_2$ par construction. Puisque
$$\varphi_2(x) = \varphi_1(x) + \alpha_2(x - x_1)(x - x_2),$$
on a
$$\alpha_2 = \frac{\varphi_2(x_2) - \varphi_1(x_3)}{(x_3 - x_2)(x_3 - x_1)} = \frac{f(x_3) - f(x_1) - f[x1, x2](x_3 - x_1)}{(x_3 - x_2)(x_3 - x_1)}$$
soit
$$f[x_1, x_2, x_3] = \frac{f[x_3, x_1] - f[x_2, x_1]}{x_3 - x_2}.$$

On a en fait la proposition suivante

**Proposition 5.3.6.** *Soit $\mathcal{S}_n$ l'ensemble des permutations de $\{1, \ldots, n\}$, alors*

1. *quelque soit $\sigma \in \mathcal{S}_n$, $f[x_1, \ldots, x_n] = f[x_{\sigma(1)}, \ldots, x_{\sigma(n)}]$,*
2. $$f[x_1, \ldots, x_n] = \frac{f[x_2, \ldots, x_n] - f[x_1, \ldots, x_{n-1}]}{x_n - x_1}.$$

*Démonstration.* 1. On écrit le polynôme d'interpolation pour les points $\{x_1, \ldots, x_n\}$ et $\{x_{\sigma(1)}, \ldots, x_{\sigma(n)}\}$ On adonc

$$\begin{aligned}\varphi_{n-1}(x) &= a_0 + a_1(x - x_1) + \ldots + a_{n-1}(x - x_1)\ldots(x - x_{n-1}) \\ &= b_0 + b_1(x - x_{\sigma(1)}) + \ldots + b_{n-1}(x - x_{\sigma(1)})\ldots(x - x_{\sigma(n-1)})\end{aligned}$$

On regarde ensuite le coefficient de $x^{n-1}$. C'est $a_{n-1}$ pour la première expression et $b_{n-1}$ pour la seconde. On a donc $a_{n-1} = b_{n-1}$ d'où le premier point.

2. On prend $\sigma$ tel que $\sigma(j) = n + 1 - j$. On a déjà $a_{n-1} = b_{n-1}$ et on sait que

$$a_{n-2} = f[x_1, x_2, \ldots, x_{n-1}] \text{ et } b_{n-2} = f[x_2, \ldots, x_n].$$

On regarde ensuite les facterus de $x^{n-2}$. On trouve

$$a_{n-2} - a_{n-1}\Big(x_1 + \ldots x_{n-1}\Big) = b_{n-2} - a_{n-1}\big(x_2 + \ldots + x_n\big)$$

d'où

$$b_{n-2} - a_{n-2} = f[x_1, \ldots, x_n](x_n - x_1),$$

c'est à dire le point deux.

□

**Évaluation du nombre d'opérations.** On range les différences divisées en un tableau

$$\begin{array}{ccccccc}
f[x_1, \ldots, x_m] & & & & & & \\
f[x_1, \ldots, x_{m-1}] & f[x_2, \ldots, x_m] & & & & & \\
f[x_1, \ldots, x_{m-2}] & f[x_2, \ldots, x_{m-1}] & f[x_3, \ldots, x_m] & & & & \\
\ldots & \ldots & \ldots & \ldots & & & \\
f[x_1, x_2] & f[x_2, x_3] & \ldots & & f[x_{m-1}, x_m] & & \\
f(x_1) & f(x_2) & \ldots & \ldots & \ldots & \ldots & f(x_m)
\end{array}$$

Pour chacun de ces coefficients, on fait 3 opération sauf pour la première ligne, soit

$$3\Big[1 + 2 + \ldots + m\Big] 3\frac{m(m-1)}{2} = \alpha$$

opérations.

Si maintenant, on veut calculer $\varphi_{m-1}(x)$, soit

$$\varphi_{m-1}(x) = a_0 + a_1(x - x_1) + \ldots a_m(x - x_1)\ldots(x - x_m)$$

il faut

$$\alpha + m + \sum_{j=1}^{m} 2j = m^2 + \frac{3}{2}m(m-1) = \frac{m(5m-3)}{2}$$

opérations.

On peut aussi écrire

$$\varphi_{m-1} = a_0 + (x - x_1)\Big[a_1 + (x - x_2)\Big[a_2 + \ldots + a_m(x - x_m)\Big]\Big]$$

soit

$$3(m-1) + \alpha = \frac{3}{2}(m+2)(m-1)$$

opérations, soit asymptotiquemen $5/3 \simeq 1.666$ fois moins que la méthode précédente.

Si on calcul $\varphi_{m-1}$ avec la forme initiale, chaque $\ell_i$ demande $2(m-1)$ additions, $2(m-1)$ multiplications et 1 division, soit $4m - 5$ opérations, et donc en tout $m(4m-5)$ opérations. Ceci demande asymptotiquement $8/3 \simeq 2.33$ fois plus d'opérations.

Cette procédure se révèle aussi plus instable dans la pratique. Dans la section 7.1.7, on propose une autre manière (plus économique) d'évaluer l'interpolé de Lagrange en un point $x$ quelconque sans avoir à déterminer les différences divisées. Il s'agit de l'algorithme d'Aitken.

### 5.3.2 Etude de l'erreur d'interpolation

Si $f$ est connue sur $[a,b]$, on cherche à évaluer

$$E_X(f) = \sup_{x\in[a,b]} |f(x) - \varphi(x)|$$

où $X = \{x_1, \ldots, x_m\}$.

On a le théorème suivant :

**Théorème 5.3.1.** *Soit $X = \{x_1, \ldots, x_m\}$ un ensemble de points deux à deux distincts de $[a,b]$, et soit $f \in C^m([a,b])$. On pose $M_m(f) = \sup_{x\in[a,b]} |f^{(m)}(x)|$, $\Pi_X(x) = \Pi_{j=1}^m (x - x_j)$ et $\varphi$ désigne l'interpolée de lagrange de $f$. Alors $|\varphi(x) - f(x)| \leq \frac{M_m(f)}{m!} |\Pi_X(x)|$ et*

$$E_X(f) \leq \frac{M_m(f)}{m!}(b-a)^m.$$

Ce résultat ne donne pas une évolution de l'erreur en fonction de $m$.

Pour montrer ce résultat, on commence par montrer le lemme suivant.

**Lemme 5.3.7.** *Sous les hypothèses du théorème 5.3.1, quelque soit $x \in [1,b]$, il existe $\xi_x \in ]a,b[$ tel que*

$$f(x) - \varphi(x) = \frac{f^{(m)}(\xi_x)}{m!}\Pi_X(x).$$

*Démonstration.* Si $x \in X$, $\Pi_X(x) = 0$ et la preuve est finie.

Supposons $x \notin X$. On considère $Y = X \cup \{x\}$. Le polynôme et $t$ définit par

$$\Psi(t) = \varphi(x) + \frac{f(x) - \varphi(x)}{\Pi_X(x)}\Pi_X(t)$$

est de degré $m$ et interpole $f$ sur $Y$. Ainsi, $F$ définie par $F(t) = \Psi(t) - f(t)$ s'annule sur $Y$. En appliquant $m-1$ fois le théorème de Rolle, il existe donc $\xi_x \in ]a,b[$ tel que

$$F^{(m)}(\xi_x) = 0.$$

Eant donné que $\varphi$ est de degré $m-1$, $\varphi^{(m)}(t) = 0$. Le coefficient de degré $m$ de $\Pi_X(t)$ étant $t^m$, on voit donc que $\Pi_X^{(m)}(t) = m!$, et donc

$$F^{(m)}(t) - f^{(m)}(t) - m!\frac{f(x) - \varphi(x)}{\Pi_X(x)}$$

d'où

$$\frac{f^{(m)}(\xi_x)}{m!} = \frac{f(x) - \varphi(x)}{\Pi_X(x)}.$$

□

*Démonstration.* Démonstration du théorème 5.3.1 Soit $x \in [a,b]$. D'après le lemme 5.3.7, il existe $\xi_x \in ]a,b[$ tel que

$$f(x) - \varphi(x) = \frac{f^{(m)}(\xi_x)}{m!}\Pi_X(x),$$

d'où le premier point du théorème. On conclue en remarquant que

$$\left| f^{(m)}(\xi_x) \right| \leq M_m(f)$$

et

$$|\Pi_X(x)| \leq (b-a)^m.$$

□

Ce résultat pourrait faire croire que pour obtenir une bonne approximation de $f$, il suffit d'augmenter le nombre de points d'interpolation. La situation est loin d'être aussi simple comme le montre l'exemple suivant.

Considérons $f$ définie par

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

c'est une fonction analytique. On l'interpole sur $[-8, 8]$ avec des points équidistants, donc $x_i =$

$-8+i\dfrac{16}{m}$. On prend $m=30$. Le polynôme d'interpolation de Lagrange (calculé avec MAPLE) est

$$\varphi(x) = -\frac{9243510314271494963119243761269881003778749009989181441}{10777089472025258132295648328333818052654310010499866625}\,x^2$$

$$+\frac{28042236452669743348774386627186740803273051026964488197}{53885447360126290661478241641669090263271550052499333125}\,x^4$$

$$-\frac{10412329558090989834571410744856827110137980688343160805 3}{51812930153967587174498309270835663714684182742787820 3125}\,x^6$$

$$+\frac{25640268674698693135097765078468648121104252836002426613}{51812930153967587174498309270835663714684182742787820 3125}\,x^8$$

$$-\frac{20234121942897430086331444606507409259119893399930569}{25274600075106140085121126473578372543748381825750156 25}\,x^{10}$$

$$+\frac{90805213170065379024774793204389774321535399094392369}{10362586030793517434899661854167132742936836548557564 0625}\,x^{12}$$

$$-\frac{52560173939089644536745794091844710724438747245951 17}{79243304941362192149232708296572191563634632430146078125}\,x^{14}$$

$$+\frac{27863976177506162430963419424794714717767401948919 7}{79243304941362192149232708296572191563634632430146078125}\,x^{16}$$

$$-\frac{13568529080225395409135201906810658063224994534961}{10362586030793517434899661854167132742936836548557564 0625}\,x^{18}$$

$$+\frac{32281183737697454119163324012652801516969969 49}{95069596612784563622932677561166355439787491271170312 5}\,x^{20}$$

$$-\frac{49610496680895515635972432073658418293272261}{82900688246348139479197294833337061943494692388460512 5}\,x^{22}$$

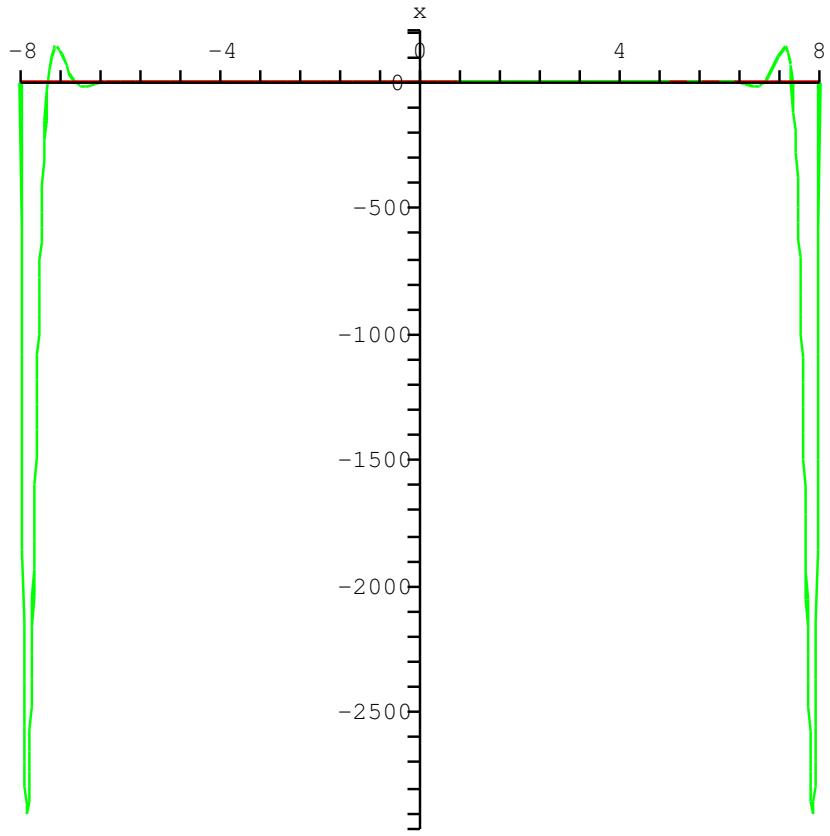
$$+\frac{14122518891830006080049820710427563212984637}{20725172061587034869799323708334265485873673097115128125}\,x^{24}$$

$$-\frac{30375635597860748606130355420165136078934641}{67356809200157863326847802052086362829089437565624166406 25}\,x^{26}$$

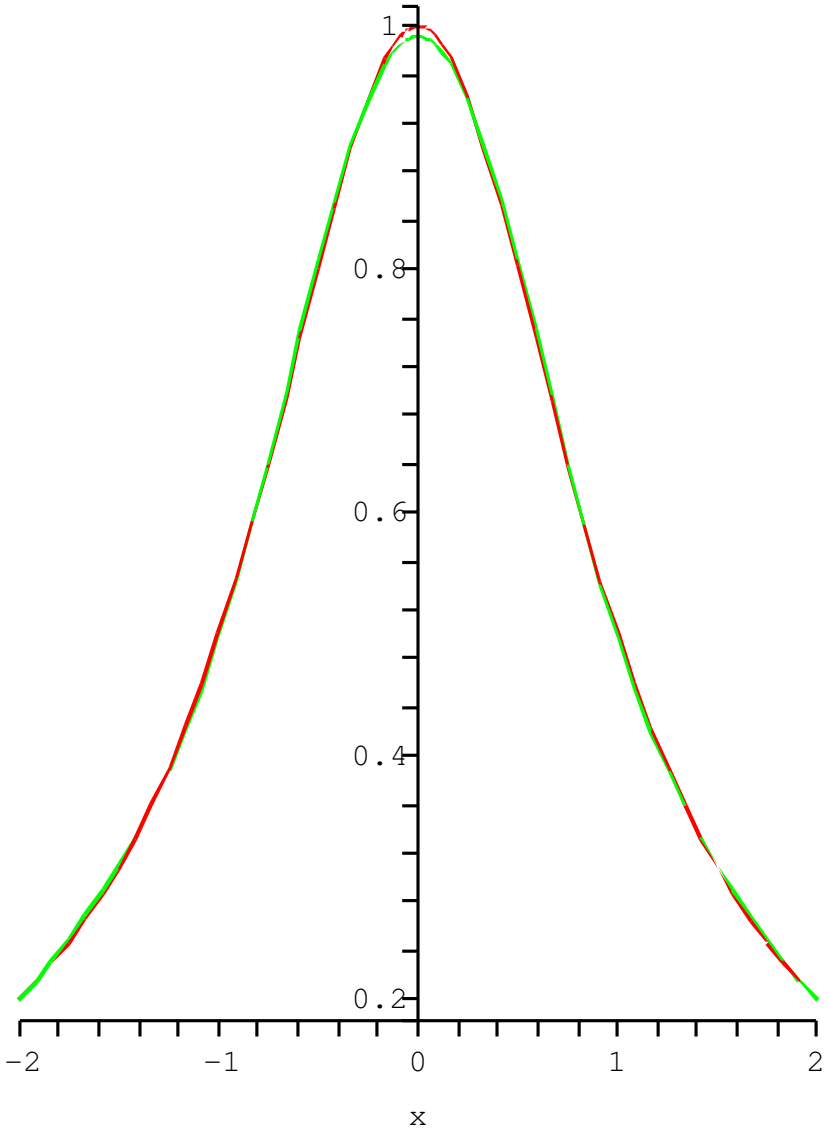
$$+\frac{17504703021214327481252729074691557634448968795 33}{17652528772179534624263464430841003341707680040 13}$$

$$+\frac{88540901833145211536614766025207452637361}{67356809200157863326847802052086362829089437565624166406 25}\,x^{28}$$

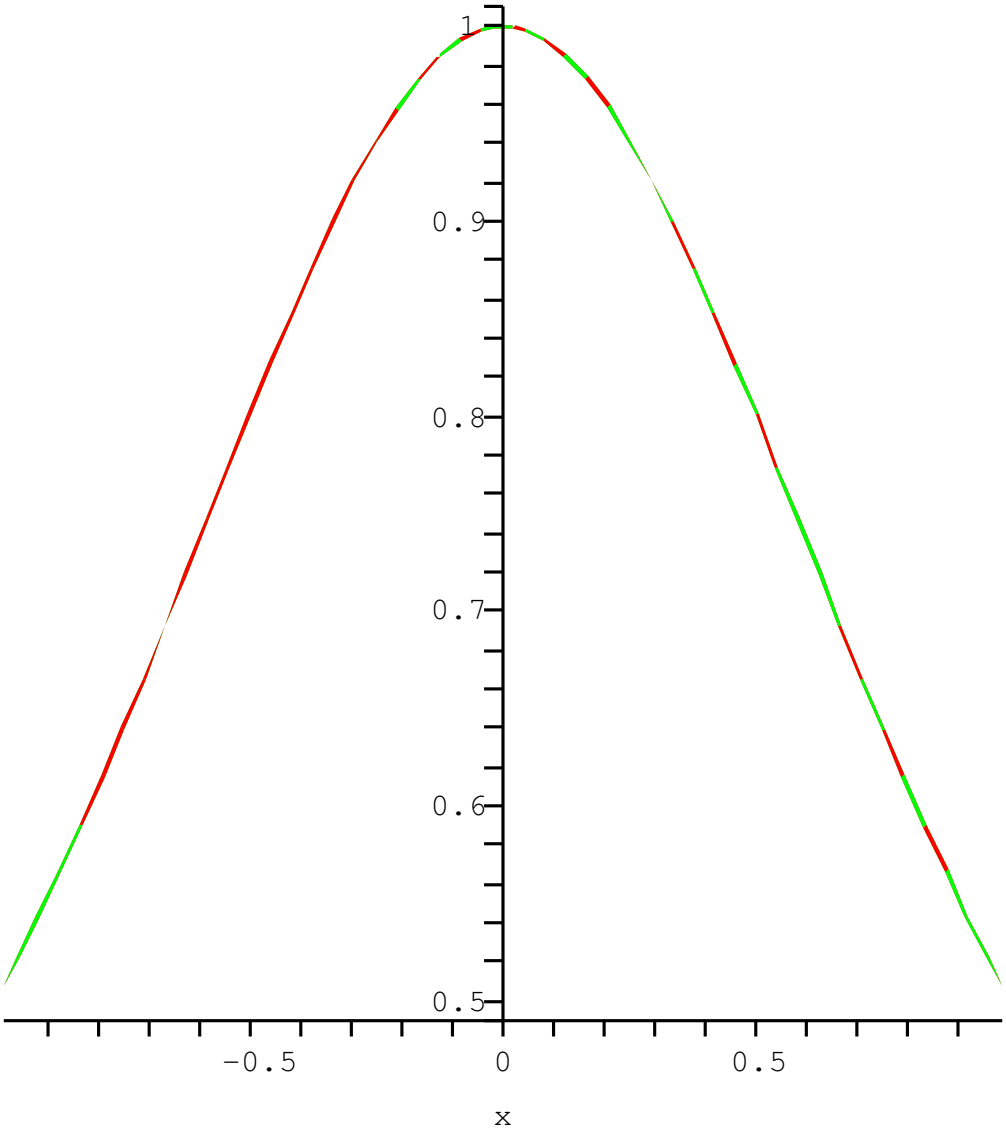
dont le graphe est




Quand on ne représente la fonction que dans $[-2, 2]$, on obtient




où sont superposées $\varphi$ et $f$. On voit que l'interpolation est meilleure, mais pas extraordinaire. Le problème, subtile, provient du fait qu'on a choisi des points équidistants. Voici, par exemple, ce qu'on obtient en prenant comme point d'interpolation les points $\tan(x_i)$, et q'on représente la fonction et son interpolée sur $[-1, 1]$ :




C'est nettement meilleur, mais quand on représente la fonction jusqu'aux bornes de $X$, les résultats ne sont pas bon (et du même type que sur la première figure). L' explication de ce phénomène de Runge est subtile, mais ceci motive le paragraphe suivant.

## 5.4 Eléments finis de Lagrange

Ici, on essaie de déterminer une approximation précise de la fonction $f$ au moyen de fonctions polynômiale tout en cherchant à ne pas faire croître le degré des polynômes. Un moyen pour réaliser cela est de ne pas chercher une approximation globale mais plutôt une approximation local.

Dans ce paragraphe, on va donner un moyen d'y arriver. Plus précisément, soit $[a,b] \subset \mathbb{R}$ un intervalle et $a = y_1 < y_2 < \ldots < y_{m-1} < y_m = b$ une subdivision de cet intervalle. On pose $Y = \{y_1, \ldots, y_m\}$. On définit l'espace des fonctions continues sur $[a,b]$ dont la restriction à $[y_i, y_{i+1}], i = 1, \ldots, m-1$ est un polynôme de degré un,

$$V_Y^1 = \{\varphi \in C^0([a,b]), \varphi_{|[y_i,y_{i+1}]} \in \mathbb{P}_1\}.$$

La méthode qui va être décrite ici peut se généraliser à des polynômes de degré plus élevés. On peut consulter [8] pour un cadre généralisant ce que l'on fait ici.

L'idée est alors de rechercher la précision en augmentant le nombre de point dans la subdivision $Y$. On a le théorème suivant

**Théorème 5.4.1.** *L'espace $V_Y^1$ est un espace vectoriel de dimension $m$ dont une base est donné par $\{p_1, \ldots p_m\}$ où*

- *Si $1 < i < m$,*

$$p_i(y) = \begin{cases} 0 & si\ y \notin [y_{i-1}, y_{i+1}], \\ \frac{y - y_{i-1}}{y_i - y_{i-1}} & si\ y \in [y_{i-1}, y_i], \\ \frac{y_{i+1} - y}{y_{i+1} - y_i} & si\ y \in [y_i, y_{i+1}]. \end{cases}$$

- *Si $i = 1$,*

$$p_1(y) = \begin{cases} \frac{y_2 - y}{y_2 - y_1} & si\ y \in [y_1, y_2] = [a, y_2] \\ 0 & sinon. \end{cases}$$

- *Si $i = m$,*

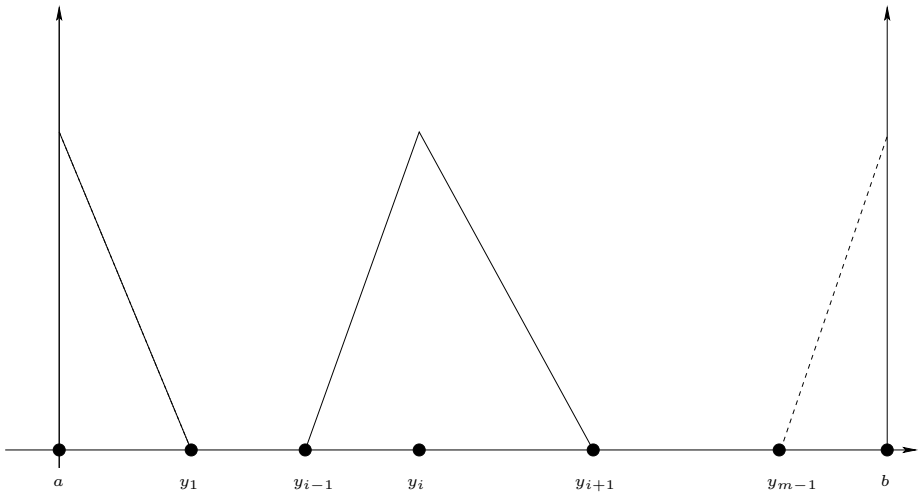
$$p_m(y) = \begin{cases} \frac{y_m - y}{y_m - y_{m-1}} & si\ y \in [y_{m-1}, y_m] = [y_{m-1}, b] \\ 0 & sinon. \end{cases}$$

On a le corollaire suivant

**Corollaire 5.4.1.** *Pour toute fonction $f$ définie sur $[a,b]$, il existe un unique $\varphi \in V_Y^1$ tel que pour tout $i$, $\varphi(y_i) = f(y_i)$ et*

$$\varphi = \sum_{i=1}^{m} y_i p_i.$$

Avant de faire la démonstration du théorème et du corollaire, on donne une représentation graphique des fonctions $p_i$.




*Démonstration du théorème 5.4.1.* Il est clair que $V_Y^1$ est un sous espace vectoriel de $C^0(a,b)$. Montrons que les fonctions $p_i$ forment une famille libre. Si

$$\sum_{j=1}^{m} \alpha_j p_j = 0$$

alors, en évaluant cette somme en $x = x_i$, on a

$$0 = \sum_{j=1}^{m} \alpha_j p_j(y) = \alpha_i.$$

Montrons que c'est une famille génératrice. Soit $\varphi \in V_Y^1$. On considère

$$\Psi = \sum_{j=1}^{m} \varphi(y_j) p_j,$$

et on va montrer que $\Psi = \varphi$. Déjà $\Psi \in V_Y^1$, donc $\Psi_{|[y_i,y_{i+1}]}$ est un polynôme de degré 1 qui vérifie $\Psi(y_i) = \varphi(y_i)$ et $\Psi(y_i) = \varphi(y_i)$ donc

$$\Psi_{|[y_i,y_{i+1}]}(y) = \varphi(y_i)\frac{y - y_{i+1}}{y_i - y_{i+1}} + \varphi(y_{i+1})\frac{y - y_i}{y_{i+1} - y_i}.$$

En appliquant le même raisonnement à $\varphi \in V_Y^1$, on voit que

$$\varphi_{|[y_i,y_{i+1}]}(y) = \varphi(y_i)\frac{y - y_{i+1}}{y_i - y_{i+1}} + \varphi(y_{i+1})\frac{y - y_i}{y_{i+1} - y_i}$$

et donc $\Psi_{|[y_i,y_{i+1}]} = \varphi_{|[y_i,y_{i+1}]}$ d'où $\Psi = \varphi$. Ceci montre que la famille est génératrice, donc c'est une base et dim $V_Y^1 = m$. □

*Démonstration.* Démonstration du corollaire 5.4.1 Soit $f$ définie sur $[a,b]$. On cherche $\varphi = \sum_{i=1}^{m} \alpha_i p_i$ telle que $\varphi(y_i) = f(y_i)$. On remarque que $p_i(y_j) = \delta_{ij}$ donc $\alpha_i = f(y_i)$ et

$$\varphi = \sum_{i=1}^{m} f(y_i) p_i.$$

□

Enfin, on donne une estimation d'erreur.

**Théorème 5.4.2.** *Soit $h = \max_j |y_{j+1} - y_j|$. Soit $f \in C^2([a,b])$ et $\varphi_h$ son approximation dans $V_Y^1$. Alors*

$$\max_{x\in[a,b]} |f(x) - \varphi(x)| \leq \frac{h^2}{8} M_2(f)$$

*où $M_2(f) = \max_{x\in[a,b]} |f''(x)|$.*

Ce résultat montre donc que quand $h \to 0$, $\varphi_h$ converge uniformément vers $f$.

*Démonstration.* Sur $[y_i, y_{i+1}]$, $\varphi$ est l'interpolé de Lagrange de $f$. En utilisant le résultat du théorème 5.3.1, on a

$$\begin{aligned}
|\varphi(x) - f(x)| &\leq \frac{M_2(f)}{2} \Pi_{|[y_i,y_{i+1}]}(x) \\
&\leq \frac{M_2(f)}{2} |(x - y_i)(x - y_{i+1}| \\
&\leq \frac{M_2(f)}{2} \left(\frac{y_{i+1} - y_i}{2}\right)^2 \\
&\leq \frac{M_2(f)}{8} h^2
\end{aligned}$$

d'où le résultat. □

## 5.5 Interpolation d'Hermite et éléments finis d'Hermite

Ce type d'interpolation ne fournit qu'un interpolant continu. Même en généralisant l'approche présentée ici en augmentant le degré du polynome employé dans $[y_i, y_{i+1}]$, on n'aura pas mieux que la continuité de l'interpolant (même si l'erreur d'approximation s'améliore). Afin de définir des interpolants plus régulier, on introduit un cadre plus général.

Remarquons que dire que

$$\text{trouver } \varphi \in V_Y^1 \text{ tel que } \varphi(x_i) = f(x_i), \forall x_i \in Y$$

peut se réécrire

$$\text{trouver } \varphi \in V_Y^1 \text{ tel que } \forall \ell \in L, \ell(\varphi) = \ell(f)$$

où $L$ est un ensemble de formes linéaires

$$L = \{\ell_1, \ldots, \ell_m\}$$

avec

$$\begin{aligned}
\ell_i \quad : \quad & C^0([a,b]) \to \mathbb{R} \\
& f \mapsto f(y_i).
\end{aligned}$$

On peut généraliser ce cadre en considérant $L$ un ensemble de formes linéaires définies sur un espace espace de fonctions $V$, et le problème d'interpolation est défini par la donnée d'un sous–espace de dimension finie de $V$ noté $V'$ et la recherche de $\varphi \in V'$ telle que

$$\forall \ell \in L, \ell(\varphi) = \ell(f).$$

Un exemple est donné par l'interpolation d'Hermite. Ici, $V = C^{d-1}([a,b])$ et

$$L = \{l_{j,i}, j = 1, \dots, m; i = 0, \dots, d-1\}$$

avec

$$\begin{array}{rcl} l_{j,i} & : & C^{d-1}([a,b]) \to \mathbb{R} \\ & & f \mapsto f^{(i)}(x_j) \end{array}$$

On voit que card $L = md =$ dim $\mathbb{P}_{md-1}$. On a le résultat suivant :

**Théorème 5.5.1.** *Pour tous $j = 1, \dots, m$ et $i = 0, \dots, d-1$, il existe un unique $\varphi_{j,i} \in \mathbb{P}_{md-1}$ tel que*

$$\begin{array}{l} \ell_{j,i}(\varphi) = 1 \\ \ell_{j',i'}(\varphi) = 0 si\ (j',i') \neq (j,i). \end{array}$$

*La famille $\{\varphi_{j,i}, j = 1, ldots, m; i = 0, \dots, d-1\}$ est une base de $\mathbb{P}_{md-1}$.*

Un exemple très utile en pratique est le cas $m = d = 2$ : il s'agit d'une interpolation cubique. Dans l'intervalle $[a,b]$, elle est définie par

$$\begin{array}{ll} \varphi(a) = f(a) & \varphi'(a) = f'(a) \\ \varphi(b) = f(b) & \varphi'(b) = f'(b) \end{array} \tag{5.2}$$

On verra en TD comment calculer $\varphi$ connaissant $(f(a), f'(a), f(b), f'(b))$. On peut démontrer le résultat suivant

**Théorème 5.5.2.** *Si $f \in C^4([a,b])$, alors $\varphi$ définie par (5.2) vérifie*

$$f(x) = \varphi(x) = \frac{1}{4!}(x-a)^2(x-b)^2 f^{(4)}(\xi_x)$$

*où $\xi_x \in [a,b]$ et*

$$\max_{x \in [a,b]} |f(x) - \varphi(x)| \leq C M_4(f)(b-a)^4$$

*où $M_3(f) = \max_{[a,b]} |f'''(x)|$ et $C$ est une constante indépendante de $a$ et $b$.*

# Chapitre 6

# Intégration numérique.

Le problème qui est abordé ici est le suivant. Soit $I$ un intervalle et $w > 0$, $f$ deux fonctions telles que $wf$ est intégrable sur $I$. Comment calculer ou approximer l'intégrale

$$\int_I f(x)w(x)dx \ ?$$

Ce problème apparaît très souvent : par exemple, si on veut résoudre

$$\frac{dy}{dt} = f(y(t), t), \qquad y(0) = y_0 \text{ donné},$$

on sait que

$$y(x) = y(0) + \int_0^x f(y(s), s)ds.$$

A partir de cette formule, on peut construire des formules approchées (le but de ce chapitre), puis des méthodes d'approximation de cette équation différentielle (non abordé dans ce cours faute de temps). De même, pour simplifier l'exposé, on supposera que $I$ est borné, $I = [a, b]$ ou $I = ]a, b[$. Par contre $w$ peut avoir des singularités comme dans $w(x) = 1/\sqrt{1 - x^2}$ si $I = ]-1, 1[$.

## 6.1 Quelques exemples

On débute en donnant quelques exemples de formules bien connues. Supposons $f$ continue dans $[a, b]$. Considérons une subdivision de $I$, $a = x_0 < x_1 < \ldots < x_{n-1} < x + n = b$. On a les formules de Riemann

$$I_n^G(f) = \sum_{j=0}^{n-1} f(x_j)(x_{j+1} - x_j),$$

ou

$$I_n^D(f) = \sum_{j=0}^{n-1} f(x_{j+1})(x_{j+1} - x_j).$$

On a la formule du point milieu

$$I_n^M(f) = \sum_{j=0}^{n-1} f\left(\frac{x_j + x_{j+1}}{2}\right)(x_{j+1} - x_j),$$

ou la formule des trapèzes

$$I_n^T(f) = \sum_{j=0}^{n-1} \frac{f(x_j) + f(x_{j+1})}{2}(x_{j+1} - x_j),$$

ou encore la formule de Simpson

$$I_{2n+1}^S = \sum j = 0^{n-1} \frac{f(x_j) + 4f\Big((x_j + x_{j+1})/2\Big) + f(x_{j+1})}{6}(x_{j+1} - x_j).$$

Les points $x_j$ et éventuellement $\frac{x_j+x_{j+1}}{2}$ sont appelés point d'intégration.

Les questions sont :

1. Comment fabriquer des formules de quadratures ? Il y en a de deux sortes : les formules simples et les formules composées.
2. Comment les comparent-t-on ? Comment voir si elles sont précises ou non ? Il faut introduire la notion d'ordre d' une formule de quadrature.
3. Pour un nombre de points d'intégration donné, comment fabriquer des formules d'ordre maximal ?

## 6.2 Ordre d'une formule de quadrature

De façon générale, une formule de quadrature est une expression de la forme

$$I(f) = \sum_{j=1}^{n} \lambda_j f(x_j) \tag{6.1}$$

où les points de quadrature $x_j$ sont $n$ points deux à deux distincts donnés dans $[a, b]$ et les scalaires $\lambda_j$ sont choisis de manière à ce que l'erreur de quadrature

$$\varepsilon_n(f) = \int_a^b f(x)w(x)dx - \sum_{j=1}^{n} \lambda_j f(x_j)$$

ne soit pas trop grande.

On a la définition suivante

**Définition 6.2.1.** *On dit que la formule de quadrature* (6.1) *est d'ordre $m$ si $m$ est le plus grand entier tel que la formule soit exacte sur $\mathbb{P}_m$.*

Avec cette définition, on voit que $I^G$ et $I^D$ sont d'ordre 0, $I^T$ et $I^M$ sont d'ordre 1, la formule de Simpson $I^S$ est d'ordre 2 au moins. On peut le vérifier en exercice.

Pour estimer l'ordre d'une formule, il suffit de chercher $m$ tel que quelque soit $k \leq m$,

$$\sum_{j=0} \lambda_j x_j^k = \int_a^b x^k w(x)dx.$$

Ceci définit un système de $m+1$ équations à $n$ inconnues (les $\lambda_i$). Pour pouvoir le résoudre, il faut qu'on ait au moins autant d'équations que d'inconnues. On supposera donc $n \leq m+1$. Le rang du système est celui de la sous–matrice formée de ses $n$ premières colonnes, il faut donc que la matrice de Vandermonde

$$\begin{pmatrix} 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^n \end{pmatrix}$$

soit inversible, ce qui est le cas puisque les $x_i$ sont deux à deux distincts.

Pour que la formule de quadrature soit exacte pour les polyômes de degré $\leq n-1$, il suffit que

$$\lambda_j = \int_a^b w(x)\ell_p(x)dx \tag{6.2}$$

où $\ell_p$ est le p$^{\text{e}}$polynôme de Lagrange associé aux nœuds $\{x_1, \dots, x_n\}$. Ceci donne $n$ relations.

Dans ce cas, la formule (6.1) peut être réécrite comme

$$\begin{aligned} I_n(f) &= \sum_{j=1}^n \left( \int_a^b \ell_j(x) w(x) dx \right) \\ &= \int_a^b \left( \sum_{j=1}^n f(x_j)\ell_j(x) \right) w(x) dx \\ &= \int_a^b \varphi(x) w(x) dx \end{aligned}$$

où $\varphi$ est le polynôme de Lagrange associé à $\{x_1, \dots, x_n\}$.

**Remarque 6.2.2.** *On peut se demander pourquoi on considère le poids $w$ dans l'intégrale et qu'on ne se contente pas de n'utiliser que des formules du type Simpson. C'est en fait pour mieux gérer les singularités de l'intégrande. Supposons qu'on veuille intégrer $f(x) = x^\alpha$ sur $[0,1]$. On sait que si $\alpha > -1$,*

$$\int_0^1 x^\alpha dx = \frac{1}{\alpha+1}$$

*. On réécrit cette relation comme*

$$\int_0^1 x^\alpha dx = \int_0^1 \frac{x^{\alpha+1/2}}{\sqrt{x}} dx$$

*et On va prendre $\alpha > -1/2$ dans ces exemples. Le poids set donc $w(x) = 1/\sqrt{x}$.*

*La formule des trapèze donne*

$$\int_0^1 x^\alpha dx \simeq \frac{1}{2}\Big(0^\alpha + 1^\alpha) = \frac{1}{2}.$$

*Considérons maintenant $w(x) = 1/sqrtx$. Ainsi $\alpha + 1/2 > 0$, $x^{\alpha+1/2}$ est continue. On cherche une formule de quadrature associée à ce poids*

$$\int_0^1 w(x)g(x)dx = \lambda_0 g(0) + \lambda_1 g(1).$$

*On a* $\ell_0(x) = 1 - x$ *et* $\ell_1(x) = x$ *et donc*

$$\lambda_1 = \int_0^1 \frac{1-x}{\sqrt{x}} dx = \frac{4}{3}$$

*et*

$$\lambda_2 = \int_0^1 \frac{x}{\sqrt{x}} dx = \frac{2}{3}.$$

*La formule est donc*

$$\int_0^1 w(x)g(x)dx = \frac{4}{3}g(0) + \frac{2}{3}g(1)$$

*qu'on applique à* $x^{\alpha+1/2}$. *On trouve donc*

$$\int_0^1 x^\alpha dx \simeq \frac{2}{3}.$$

*La formule à poids donne une erreur* $\varepsilon = |\frac{1}{\alpha+1} - 2/3|$, *la formule des trapèze donne* $\epsilon' = |\frac{1}{\alpha+1} - 1/2|$. *On a donc* $\varepsilon < \epsilon'$ *si*

$$\frac{1}{\alpha+1} > \frac{1}{2}(\frac{1}{2} + \frac{2}{3}) = \frac{5}{12}$$

*soit* $\alpha < 7/5$. *L'erreur donnée par la formule à poids est inférieure à celle de la formule des trapèzes si* $-1/2 \leq \alpha < 7/5$. *Plus généralement, les formules à poids permettent de mieux prendre en compte les singularités intégrables.*

## 6.3 Formules simples et formules composées

Les formules simples sont obtenues en considérant directement l'intervalle $[a, b]$. Les formules composées sont obtenues en considérant une subdivision de $[a, b]$, et sur chaque intervalle $[y_j, y_{j+1}]$, on considère une formule simple. Elles sont donc obtenues en juxtaposant des formules simples.

### 6.3.1 Exemples de formules simples

Des exemples sont donnés par $I_1^G(f) = f(0), \qquad I_1^D(f) = f(1)$.
La formule simple du point milieu est

$$I_T^2(f) = \frac{f(0) + f(1)}{2}.$$

La formule simple des trapèze est

$$I_3^S(f) = \frac{f(0) + 4f(1/2) + f(1)}{6}.$$

On va calculer l'ordre de $I^T$. On a

$$\begin{aligned}
I_3^S(1) &= 1 = \int_0^1 1dx \\
I_3^S(x) &= \tfrac{1}{6}(4\tfrac{1}{2}+1) = \tfrac{1}{2} = \int_0^1 xdx \\
I_3^S(x^2) &= \tfrac{1}{6}(4\tfrac{1}{2^2}+1) = \tfrac{1}{3} = \int_0^1 x^2dx \\
I_3^S(x^3) &= \tfrac{1}{6}(4\tfrac{1}{2^3}+1) = \tfrac{1}{4} = \int_0^1 x^3dx \\
I_3^S(x^4) &= \tfrac{1}{6}(4\tfrac{1}{2^4}+1) = \tfrac{5}{24} = \int_0^1 x^4dx
\end{aligned}$$

La formule de Simpson est donc d'ordre 3.

On a aussi les formules de Newton–Cotes obtenues en considérant le poids $w = 1$ et les points

$$x_1 = a, \ldots, x_j = a + \frac{(j-1)(b-a)}{n-1}, \ldots x_n = b$$

ce sont les formules fermées ou les formules ouvertes avec

$$x_1 = a + \frac{b-a}{n+1}, \ldots, x_j = a + \frac{(j-1)(b-a)}{n+1}, \ldots x_n = b - \frac{b-a}{n+1}.$$

Les poids sont obtenus par la formule (6.2). Les formules ouvertes permettent de traiter des problèmes avec des singularités en $a$ ou $b$.

### 6.3.2 Formules composées

Si on a une formule de quadrature

$$I(f) = \sum_{j=1}^{n} \lambda_j f(x_j)$$

sur $[0,1]$, on peut obtenir une formule sur $[a,b]$ grâce à la transformation $x \mapsto a + x(b-a)$, et donc

$$\int_a^b g(x)dx \simeq (b-a)\sum_{j=1}^{n} \lambda_j g(a + x_j(b-a)). \tag{6.3}$$

Les formules des trapèzes et de Simpson sont des exemples de formules fermées, la formule du point milieu est une formule ouverte.

Pour obtenir une formule composée, on considère une subdivision de $[a,b]$, $a = a_0 < a_1 < \ldots < a_{n-1} < a_n = b$. Dans chacun des intervalles $[a_j, a_{j+1}]$, on applique la formule (6.3) pour obtenir

$$I_{n,p}(f) = \sum_{i=0}^{p-1}(a_{i+1} - a_i)\sum_{j=1}^{n} \lambda_j f(a_i + x_j(a_{j+1} - a_j)).$$

L'ordre de la formule $I_{n,p}$ est au moins égale à celui de la formule (6.3).

## 6.4 Estimations d'erreur

### 6.4.1 Cas des formules simples.

On a le théorème suivant

**Théorème 6.4.1.** *Soit une formule de quadrature d'ordre $m$ sur $[a,b]$ notée*

$$I_n(f) = \sum_{j=1}^{n} \lambda_j f(x_j).$$

*On note $t_+ = \max(t,0)$, et on pose $t_+^0 = 1$ si $t > 0$ et $t_+^0 = 0$ si $t \leq 0$. On définit $G$ par*

$$G(y) = \int_a^b (x-y)_+^m dy - \sum_{j=1}^{m} \lambda_j (x_i - y_j)_+^m.$$

*Alors, quelque soit $f \in C^{m+1}(a,b)$, on a*

$$\int_a^b f(y)w(y)dy - I_n(f) = \frac{1}{m!} \int_a^b f^{(m+1)}(y)G(y)dy.$$

*La fonction $G$ est appelée noyau de Peano.*

Avant de commencer la démonstration, remarquons que $G(y)$ est l'erreur de quadrature qu'on fait avec $(x-y)_+^m$.

*Démonstration.* On commence à écrire la formule de Taylor avec reste intégral

$$f(x) = P(x) + R(x)$$

avec

$$P(x) = \sum_{j=0}^{m} \frac{f^{(j)}(a)}{j!}(x-a)^j$$

et

$$R(x) = \frac{1}{m!} \int_a^x f^{(m+1)}(y)(x-y)^m dy.$$

Puisque la formule est d'ordre $m$,

$$\int_a^b P(x)dx = I_n(P).$$

De plus, si $y \in [a,x]$, $x \leq y$ donc $x - y \geq 0$ et $(x-y)^m = (x-y)_+^m$. Si $y \in ]x,b]$, $x - y \leq 0$ et donc $(x-y)_+^m = 0$, si bien que

$$\int_a^x f^{(m+1)}(y)(x-y)^m dy = \int_a^b f^{(m+1)}(y)(x-y)_+^m dy$$

et

$$\begin{aligned}\int_a^b w(x)R(x)dx &= \int_a^b \frac{w(x)}{m!}\left(\int_a^x f^{(m+1)}(y)(x-y)^m dy\right)dx \\ &= \int_a^b \frac{w(x)}{m!}\left(\int_a^b f^{(m+1)}(y)(x-y)_+^m dy\right)dx \\ &= \int_a^b \frac{1}{m!}\left(\int_a^b w(x)(x-y)_+^m dx\right)f(y)dy.\end{aligned}$$

De même,

$$\begin{aligned}I_n(R) &= \sum_{j=1}^m \lambda_j R(x_j) \\ &= \sum_{j=1}^n \lambda_j \frac{1}{m!}\int_a^{x_j} f^{(m+1)}(y)(x_j-y)^m dy \\ &= \sum_{j=1}^n \lambda_j \frac{1}{m!}\int_a^b f^{(m+1)}(y)(x_j-y)_+^m dy \\ &= \frac{1}{m!}\int_a^b \left(\sum_{j=1}^n \lambda_j (x_j-y)_+^m\right)f(y)dy.\end{aligned}$$

En utilisant ces deux relations, on obtient le résultat voulu. □

On obtient à partir de là l'estimation d'erreur suivante :

**Corollaire 6.4.1.** *Soit $I+n$ une formule de quadrature d'ordre $m$. Alors, pour toute $f \in C^{m+1}([a,b])$, on a*

$$\left|I_n(f) - \int_a^b f(s)ds\right| \leq \frac{1}{m!}\max_{x\in[a,b]}|f^{(m+1)}(x)| \int_a^b |G(y)|dy.$$

dont la démonstration est immédiate.

**Exemples de noyaux de Peano.** On considère

$$I_1(f) = f(c)$$

pour $f$ définie sur $[0,1]$ et $c \in [0,1]$. Si $c \neq 1/2$, la formule est d'ordre 1, et d'ordre deux si $c = 1/2$.

Si $c \neq 1/2$, on a donc

$$\begin{aligned}G(y) &= \int_0^1 (x-y)_+^0 dx - (c-y)_+^0 \\ &= \int_y^1 dx - 1_{[0,c]}(y) \\ &= 1 - y - 1_{[0,c]}(y)\end{aligned}$$

Si $c = 1/2$, on a

$$\begin{aligned} G(y) &= \int_0^1 (x-y)_+^1 dx - (1/2-y)_+^1 \\ &= \int_y^1 (x-y)dx - (1/2-y)_+ \\ &= \frac{1-y^2}{2} - (1-y)y - (1/2-y)_+ \end{aligned}$$

On a donc deux cas différents
– Si $y < 1/2$, $G(y) = y^2/2$,
– Si $y > 1/2$, $G(y) = (1-y)^2/2$
Le corrolaire 6.4.1 fournit donc les estimations suivantes
– Si $c \neq 1/2$,

$$\left| I_1(f) - \int_0^1 f(s)ds \right| \leq \max_{x\in[a,b]} |f'(x)|\, c(1-c),$$

– Si $c = 1/2$,

$$\left| I_1(f) - \int_0^1 f(s)ds \right| \leq \frac{1}{24} \max_{x\in[a,b]} |f''(x)|.$$

### 6.4.2 Cas des formules composées.

On se donne une formule simple pour $[0,1]$,

$$I_n(f) = \sum_{j=1}^n \lambda_j f(x_j) \tag{6.4}$$

d'où l'on tire la formule composée

$$I_{p,m}(f) = \sum_{i=1}^p (b_i - a_i) \left[ \sum_{j=1}^n \lambda_j f\big(a_i + x_j(b_i - a_i big)\right] \tag{6.5}$$

associé à la subdivision $[a,b] = \cup_{i=1}^p [a_i, b_i]$ où on a posé par comodité $b_i = a_{i+1}$.

Du théorème 6.4.1, on déduit le résultat suivant.

**Théorème 6.4.2.** *Si la formule (6.4) est d'ordre $m$, la formule composée (6.5) vérifie*

$$\left| I_{p,m}(f) - \int_a^b f(y)dy \right| \leq (b-a) \frac{\max_i (b_i - a_i)^{m+1}}{m!} \max_{x\in[a,b]} |f^{(m+1)}(x)| \int_0^1 |G(y)|dy.$$

*Démonstration.* Ce résultat se déduit simplement de la démonstration du théorème 6.4.1 en faisant le changement de variable $x \mapsto a + x(b-a)$ qui transforme $[0,1]$ en $[a,b]$. □

## 6.5 Formules gaussiennes

### 6.5.1 Théorie générale

Soit $[a, b]$ un intervalle compact et $w$ une fonction intégrable positive sur $[a, b]$. Etant donné de $n$ points deux à deux distincts, on sait qu'il existe une formule d'ordre au moins $n-1$. Elle est construite en considérant l'interpolant de Lagrange sur ces nœuds. La question qui se pose est la suivante. Comment déterminer ces $n$ nœuds et les poids $\lambda_i$ de manière à ce que la formule soit de degré le plus haut possible, c'est à dire que

$$\int_a^b f(x)w(x)dx - \sum_{i=1}^{n} \lambda_i f(x_i)$$

s'annule sur $\mathbb{P}_k$ avec $k$ le plus grand possible. Une fois les $x_i$ déterminés on sait que

$$\lambda_i = \int_a^b \Pi_{p\neq i} \frac{x - x_p}{x_i - x_p} dx,$$

donc seuls les $x_i$ sont à déterminés.

Si $P \in \mathbb{P}_k$, on doit avoir

$$\int_a^b P(x)w(x)dx = \sum_{j=1}^{n} \lambda_j P(x_j).$$

Considérons le polynôme $Q$ définit par

$$Q(x) = \Pi_{j=1}^{n}(x - x_j).$$

On fait la division euclidienne de $P$ par $Q$,

$$P = Q\ M + r,$$

où le degré de $r$ est $\leq n$. On a donc

$$\int_a^b P(x)w(x)dx = \sum_{j=1}^{n} \lambda_j P(x_j) = \sum_{j=1}^{n} \lambda_j r(x_j),$$

et donc nécessairement

$$\int_a^b Q(x)M(x)w(x)dx = 0 \tag{6.6}$$

quelque soit $M$ polynôme de degré $k-n$ car la formule est d'ordre $k$. Par induction (i.e. division euclidienne successive), on voit qu'il suffit que (6.6) soit vrai pour tout polynôme de degrè $\leq n$

On est donc conduit à étudier la forme bilinéaire

$$(P, Q) \in \mathbb{P}_n \times \mathbb{P}_n \mapsto \int_a^b P(x)Q(x)w(x)dx := (P, Q). \tag{6.7}$$

**Définition 6.5.1.** *On appelle polynôme orthogonaux relativement à un poids $w$ positif et intégrable sur un intervalle $[a, b]$ la suite suite des polynômes $P_0, P_1, \ldots P_n, \ldots$ ayant les propriétés suivantes*

1. *Quel que soit $n$, $P_n$ est de degré $n$, le coefficient de plus haut degré vaut* 1,
2. *Quel que soit $n$, quel que soit $P \in \mathbb{P}_k$ avec $k \leq n-1$,*

$$\int_a^b P_n(x)P(x)w(x)dx = 0.$$

Montrons d'abord que de telles suites existent

**Lemme 6.5.2.** *Que que soit le poids $w > 0$ intégrable sur $[a,b]$, il existe une unique suite de polynômes orthogonaux qui satisfont à*

$$P_n = x^n - \sum_{i=0}^{n-1} c_{in} P_i$$

*avec*

$$c_{in} = \frac{(x^n, P_i)}{(P_i, P_i)}.$$

*Démonstration.* On a $P_0 = 1$. Supposons connus les $n-1$ polynômes orthogonaux. On a nécessairement (si existence)

$$P_n = x^n - \sum_{j=0}^{n-1} \alpha_j P_j$$

On détermine $\alpha_j$ en écrivant que $(P_n, P_j) = 0$ si $j < n$.

Une autre façon de faire le calcul est d'appliquer la méthode d'orthogonalisation de Gramm Schmidt à $\{1, x, \ldots, x^n\}$. □

Donnons une propriété qualitative des polynômes orthogonaux qui permet de construire les formules de quadrature : les zéros sont deux 1a deux distincts et tous dans $[a,b]$. Plus précisément

**Théorème 6.5.1.** *Soit $w$ un poids intégrable et strictement poistif dans $[a,b]$. Alors, quelque soit $n$, toutes les racines du $n^e$polynôme orthogonal $P_n$ sont simples et sont dans $]a,b[$.*

*Démonstration.* Soient $x_1$, $x_2, \ldots, x_j$ les racines de $P_n$ appartenant à $]a,b[$, et comptées avec leur multiplicité. On a $j \leq n$ et peut être nul. Supposons que $j < n$. Comme les coefficients de $P_n$ sont réels, $P_n$ va changer de signe à toutes les racines de multiplicité impaire. Posons, si $j > 0$,

$$Q(x) = \Pi_{k=1}^{j}(x - x_k).$$

Si $j = 0$, on prend $Q = 1$. Le produit $P_n Q$ ne change pas de signe dans $]a,b[$, et $Q$ est de degré $\leq n-1$, donc

$$\int_a^b P_n(x)Q(x)w(x)dx = 0.$$

Comme l'expression $P_n(x)Q(x)w(x)$ est positive, elle doit être nulle ce qui est impossible : les racines de $P_n$ sont donc toutes dans $[a,b]$. Il faut montrer maintenant qu'elles sont simples.

Soit $x_1$ supposée multiple. Alors $P_n(x) = p(x)(x - x_1)^2$ et $P_n$ et $p$ sont de même signe. Comme $p$ est de degré au plus $n-2$, et donc

$$\int_{-1}^{1} P_n(x)p(x)w(x)dx = 0$$

ce qui est absurde.

Les racines sont donc simples. □

On finit par donner la réponse à la question posée :

**Théorème 6.5.2.** *L'unique formule à $n$ points d'ordre maximale est la formule par interpolation construite en prenant pour nœuds les zéros du $n^e$polynôme orthogonal par rapport au poids $w$. Elle est d'ordre $2n-1$. La formule ainsi déterminée est dite gaussienne.*

*Démonstration.* On part de (6.7). Cette formule exprime que les points d'interpolations $x_1, \ldots, x_n$, s'ils existent, sont tels que

$$Q(x) = \Pi_{j=1}^{n}(x - x_j)$$

est orthogonal à tout polynôme de degré $n-1$. Au vu de ce qui vient d'être dit sur les polynômes orthogonaux, on voit que nécessairement les $x_i$ sont les zéros d'un polynôme orthogonal pour le poids $w$. C'est forcément $P_n$, et on a l'existence et l'unicité.

Son ordre est $2n-1$ : Soit $k$ l'ordre de cette formule. Nécessairement $k \geq 2n-1$ car si $P$ est de degré $2n-1$, on effectue la division euclidienne avec $P_n$, $P = P - np + r$ et comme précédement, $r$ est de degré $n-1$ donc

$$\int_a^b P(x)w(x)dx = \int_a^b r(x)w(x)dx = \sum_{j=1}^{n} \lambda_j r(x_j) = \sum_{j=1}^{n} \lambda_j P(x_j)$$

la première égalité étant conséquence de l'orthogonalité de $P_n$ et de de $p$ car son degré est au plus $n-1$.

Si $k > 2n-1$, on prend $P(x) = P_n^2(x)$, et on a nécessairement

$$\sum_{j=0}^{n} \lambda_j P_n^2(x_j) = 0$$

donc

$$\int_a^b P_n^2(x)w(x)dx = 0$$

ce qui est absurde. □

### 6.5.2 Exemples

Avant de continuer, on va donner deux exemples, l'un pour le poids $w = 1$ et l'autre pour le poids

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

Dans les deux cas $[a, b] = [-1, 1]$.

**Exemple du poids $w = 1$ : les polynômes de Legendre.**

On a dans ce cas

$$P_n(x) = \sqrt{n + \frac{1}{2}} \frac{1}{2^n n!} \frac{d^n}{dx^n} \left[ (x^2 - 1)^n \right].$$

*Démonstration.* On pose pour simplifier les notations

$$R_n(x) = \frac{d^n}{dx^n} \left[ (x^2 - 1)^n \right]$$

qui est un polynôme de degré $n$ de coefficient dominant $2n(2n-1)\dots(n+1)$. On vérifie ensuite que si $p < n$, il existe un polynôme $r_p(x)$ tel que

$$\frac{d^p}{dx^p} \left[ (x^2 - 1)^n \right] = r_p(x)(x^2 - 1)^{n-p}.$$

Cette relation est vraie si $p = 0$. Supposons la vraie pour $p - 1$. Alors

$$\begin{aligned} \frac{d^p}{dx^p} \left[ (x^2 - 1)^n \right] &= \frac{d}{dx} \left[ r_{p-1}(x)(x^2 - 1)^{n-p+1} \right] \\ &\left( r'_{p-1}(x)(x^2 - 1) + 2x(n - p + 1) r_{p-1}(x) \right) (x^2 - 1)^{n-p}. \end{aligned}$$

A partir de là, et supposant $m \leq n$, on intègre par partie

$$\begin{aligned} \int_{-1}^{1} R_n(x) R_m(x) dx &= \frac{d^{n-1}}{dx^{n-1}} [(x^2 - 1)^n] \frac{d^m}{dx^m} [(x^2 - 1)^m] \Big|_{x=-1}^{x=1} \\ &- \int_{-1}^{1} \frac{d^{n-1}}{dx^{n-1}} [(x^2 - 1)^n] \frac{d^{m+1}}{dx^{m+1}} [(x^2 - 1)^m] dx \end{aligned}$$

Le premier terme s'annule en $\pm 1$ grâce au résultat préliminaire, et de proche en proche, on a

$$\begin{aligned} &\int_{-1}^{1} R_n(x) R_m(x) dx \\ &\quad = (-1)^p \int_{-1}^{1} \frac{d^{n-p}}{dx^{n-p}} [(x^2 - 1)^n] \frac{d^{m+p}}{dx^{m+p}} [(x^2 - 1)^m] dx \end{aligned}$$

si $p \leq n$. Si $m < n$, on prend $p = m + 1$ et la dérivée d'ordre $m + p = 2m + 1$ de $(x^2 - 1)^m$ est nulle d'où

$$\int_{-1}^{1} R_n(x) R_m(x) dx = 0.$$

Si $m = n$, on prend $p = n$ et

$$\int_{-1}^{1} R_n(x) R_m(x) dx = (2n)! \int_{-1}^{1} (1 - x^2)^n dx = (2n)! I_n.$$

En intégrant par partie, on voit que

$$\int_{-1}^{1}(1-x^2)^n dx = x(1-x^2)^n\Big|_{-1}^{1} + 2n\int_{-1}^{1} x^2(1-x^2)^{n-1}dx$$

et donc par récurence,

$$I_n = 2n(I_n - I_{n-1})$$

soit

$$I_n = \frac{2n}{2n+1}I_{n-1}$$

et donc

$$I_n = \frac{2n(2n-2)\dots 2}{(2n+1)(2n-1)\dots 3.1}2 = \frac{2^{n+1}n!}{(2n+1)(2n)!/2^n n!} = \frac{2^{2n+1}(n!)^2}{(2n+1)(2n)!}.$$

Finalement,

$$\int_{-1}^{1} R_n(x)^2 dx = \frac{2^{2n+1}(n!)^2}{(2n+1)(2n)!}$$

Ceci montre que les $P_n$ forme une famille orthonormée, le coefficient dominant de $P_n$ est strictement positif. □

On appele polynôme de Legendre les polynômes $Q_n$ donnés par

$$Q_n(x) = \frac{1}{n!2^n}\frac{d^n}{dx^n}\left[(1-x^2)^2\right].$$

Donnons quelques valeurs.

1. Si $n = 1$, on a
$$\int_{-1}^{1} f(x)dx \simeq 2f(0).$$
C'est la formule du point milieu qui est précise à l'ordre 1.
2. Si $n = 2$,
$$\int_{-1}^{1} f(x)dx \simeq f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$
qui est précise à l'ordre 3.
3. Si $n = 3$,
$$\int_{-1}^{1} f(x)dx \simeq \frac{5}{9}f\left(-\frac{\sqrt{15}}{5}\right) + \frac{8}{9}f(0) + \frac{5}{9}f\left(\frac{\sqrt{15}}{5}\right).$$
qui est précise à l'ordre 5
4. Si $n = 4$,
$$\int_{-1}^{1} f(x)dx \simeq \frac{1}{180}\left(-5+3\sqrt{30}\right)\sqrt{30}\left[f\left(-1/35\sqrt{525+70\sqrt{30}}\right) + f\left(1/35\sqrt{525+70\sqrt{30}}\right)\right]$$
$$+ \frac{1}{180}\left(5+3\sqrt{30}\right)\sqrt{30}\left[f\left(-1/35\sqrt{525-70\sqrt{30}}\right) + f\left(1/35\sqrt{525-70\sqrt{30}}\right)\right]$$
qui est d'ordre 7,
5. etc

**Exemple du poids $w(x) = 1/sqrt1 - x^2$ : les polynômes de Tchébychev**

Le poids

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

est singulier en $\pm 1$ mais intégrable sur $[-1, 1]$.

Si arccos désigne la détermination inverse de cos dans $[0, \pi]$,

$$\theta = \arccos(x) \text{ si et seulement si } x \in [0, \pi] \text{ et } \cos\theta = x,$$

les fonctions $Q_n(x) = \cos\Big(\arccos x\Big)$ définies sur $[-1, 1]$ sont orthogonales relativement au poids $w(1) = \frac{1}{\sqrt{1-x^2}}$. De plus, $Q_n$ est un polynôme de degré $n$ et

$$\int_{-1}^{1} Q_n^2(x)\frac{1}{\sqrt{1-x^2}}dx = \begin{cases} \pi/2 & \text{si } n \geq 1 \\ \pi & \text{si } n = 0 \end{cases}$$

*Démonstration.* En faisant le changement de variable $x = \cos(\theta)$, on a

$$\int_{-1}^{1} Q_n^(x)Q_m(x)\frac{1}{\sqrt{1-x^2}}dx = \int_0^{\pi} \cos(nx)\cos(mx)dx$$

En utilisant les formules d'addition, on voit qu'il ne reste plus qu'à prouver que $Q_n$ est un polynôme.

On a $Q_0 = 1$, $Q_1 = x$ et

$$Q_2(x) = \cos(2\arccos(x)) = 2\cos^2(\arccos x) - 1 = 2x^2 - 1.$$

En posant $\theta = \arccos x$, on a

$$\begin{aligned} Q_{n-1}(x) + Q_{n+1}(x) &= \cos(n-1)\theta + \cos(n+1)\theta \\ &= 2\cos\theta\ \cos n\theta = 2xQ_n(x) \end{aligned}$$

Ainsi,

$$Q_{n+1}(x) = 2xQ_n(x) - Q_{n-1}(x).$$

□

Donnons quelques valeurs.

1. Si $n = 1$, on a
$$\int_{-1}^{1} f(x)\frac{1}{\sqrt{1-x^2}}dx \simeq 2f(0).$$
C'est encore la formule du point milieu qui est précise à l'ordre 1.
2. Si $n = 2$,
$$\int_{-1}^{1} f(x)\frac{1}{\sqrt{1-x^2}}dx \simeq \frac{\pi}{2}\left[f\Big(-\frac{1}{\sqrt{2}}\Big) + f\Big(\frac{1}{\sqrt{2}}\Big)\right],$$
3. Si $n = 3$,
$$\int_{-1}^{1} f(x)\frac{1}{\sqrt{1-x^2}}dx \simeq \frac{\pi}{3}\left[f\Big(\frac{\sqrt{3}}{2}\Big) + f(0) + f\Big(-\frac{\sqrt{3}}{2}\Big)\right]$$

On voit que si $x_j = \cos\left(j\frac{\pi}{n}\right)$ pour $j = 1, \dots n$, on a

$$\int_{-1}^{1} f(x) \frac{1}{\sqrt{1-x^2}} \simeq \frac{\pi}{n} \sum_{j=1}^{n} f(x_j).$$

Comment peut–on expliquer ce résultat ? Ceci sera fait en TD.

# Chapitre 7

# Exercices et Travaux dirigés

## 7.1 Exercices

### 7.1.1 Précision des calculs et problèmes d'arrondis

1. On vérifie que *%eps*$= 2^{-52}$ est la précision maximale pour les calculs en Scilab : pour cela on compare à 1 les nombres : 1+ *%eps* et 1+ *%eps* /2.
2. On vérifie qu'en informatique, l'addition n'est pas une opération associative à cause des erreurs d'arrondis : écrire un programme qui calcule $z_1 = ((y + x) - x)/y$ et $z_2 = (y + (x - x))/y$ et le tester avec des valeurs *bien choisies* de $x$ et de $y$.
3. Écriture d'un programme qui calcule le cœfficient binômial $\binom{n}{k}$ pour $n$ et $k$ donnés :
   (a) Écrire la formule utilisant les factorielles et la simplifier afin d'obtenir un nombre de facteurs minimal au numérateur comme au dénominateur.
   (b) Écrire une fonction de deux variables avec ou sans boucle qui calcule $\binom{n}{k}$.
4. Écrire un programme qui calcule de trois façons différentes $(1 - x)^7$ :
   (a) En calculant $y = 1 - x$ puis $y^7$.
   (b) En utilisant la formule du binôme de Newton : $(1 - x)^7 = 1 - 7x + 21x^2...$
   (c) En utilisant la méthode de Hörner
   (d) Tracer sur un même graphique les deux premières fonctions dans l'intervalle $[0, 99; 1, 01]$. Commentaire.
   (e) Tracer sur un autre graphique,sur le même intervalle,les première et troisième fonctions. Commentaire.
   (f) En utilisant la fonction qui calcule les cœfficients binômiaux (exercice 3), généraliser à $(1 - x)^n$.
5. (a) **Définitions**
   – Soit $f$ une fonction définie et $n$ fois dérivable sur un intervalle $I$ de $\mathbb{R}$ ; $f$ est solution d'une équation différentielle d'ordre $n$ sur $I$ s'il existe une relation fonctionnelle entre $f$ et ses $n$ premières dérivées sur $I$.
   – Cas particulier : $f$ est solution d'une équation différentielle linéaire d'ordre $n$ sur $I$ si la relation fonctionnelle entre $f$ et ses $n$ premières dérivées est linéaire : $\exists(a_0, \ldots, a_n, b)$, fonctions au moins continues sur $I$ telles que $\forall x \in I$ , $a_n(x).f^{(n)}(x)+\cdots+a_1(x).f'(x)+ a_0.f(x) = b(x)$ **(1)**

(b) **Principe de superposition**
Soient $f_1$ et $f_2$ deux solutions de **(1)**, alors $f_2 - f_1$ est solution de $a_n(x).y^{(n)} + \cdots + a_1(x).y' + a_0.y = 0$ **(2)** qui s'appelle équation homogène (ou sans second membre) associée à **(1)**. Une solution quelconque de **(1)** s'obtient donc en ajoutant à une solution particulière de **(1)** n'importe quelle solution de l'équation sans second membre associée. On procède alors en deux temps : résolution de **(2)**, puis recherche d'une solution particulière de **(1)** et synthèse.
Proposition : l'ensemble des solutions de **(2)** est un espace vectoriel sur $\mathbb{R}$.

(c) **Étude d'un cas particulier : résolution de l'équation homogène $\mathbf{y'' + a.y' + b.y = 0}$ $(\star)$**
L'ensemble des solutions de $y'' + a\,y' + b\,y = 0$ $(\star)$ est un plan vectoriel sur $\mathbb{R}$.
On cherche des solutions sous la forme $f(x) = e^{rx}$ où $r \in \mathbb{C}$. On constate que $r$ est solution de l'équation caractéristique $r^2 + ar + b = 0$. On distingue donc trois cas selon la valeur du discriminant $\Delta$.

i. $\underline{\Delta > 0 :}$ il y a deux racines réelles distinctes $r_1$ et $r_2$, donc deux fonctions solutions : $x \longmapsto e^{r_1 x}$ et $x \longmapsto e^{r_2 x}$. Soit $x \longmapsto z(x)$ une autre solution ; on étudie la fonction $y : x \longmapsto e^{-r_1 x}.z(x)$ et on montre en dérivant que $y$ est solution de l'équation différentielle $y'' + (2\,r_1 + a)y' = 0$ d'où $y(x) = K.e^{-(2r_1+a)x} + Q) = K.e^{(r_2 - r_1)x} + Q$ ; en effet $-(2\,r_1 + a) = r_2 - r_1$
Donc toute solution est de la forme $z(x) = K.e^{r_2 x} + Q.e^{r_1 x}$ et l'ensemble des solutions est bien un espace vectoriel de dimension deux ayant pour base les deux fonctions obtenues à partir de l'équation caractéristique.

ii. $\underline{\Delta = 0 :}$ il y a une racine double réelle $r_0$ ; on considère de même une solution quelconque $z(x)$ et $y$ définie par $y(x) = e^{-r_0 x}.z(x)$ et on a $(2\,r_1 + a) = 0$ donc l'équation différentielle dont $y$ est solution devient $y'' = 0$ donc $y(x) = K.x + Q$ et $z(x) = e^{r_0 x}(K.x + Q)$. Une base de solutions est alors $(x \mapsto e^{r_0 x}, x \mapsto x.e^{r_0 x})$.

iii. $\underline{\Delta < 0 :}$ il y a deux racines complexes conjuguées $r_1 = \alpha + i\beta$ et $r_2 = \alpha - i\beta$ ; une base de solutions réelles est alors $(x \mapsto e^{\alpha x}.\cos(\beta x), x \mapsto e^{\alpha x}.\sin(\beta x))$

(d) **Exemples :** trouver l'ensembles des solutions réelles ou complexes des équations différentielles suivantes :
$(a)$ $y'' + y' - 6\,y = 0$ $\quad$ $(b)$ $4\,y'' - 4\,y' + y = 0$
$(c)$ $y'' + y' + y = 0$ $\quad$ $(d)$ $i\,y'' + y' + 2\,i\,y = 0$

6. Calculer le produit matriciel $A \times B$ suivant :

$$A = \begin{pmatrix} 4 & -3 & 0 & 2 \\ -5 & 2 & -1 & 7 \\ 1 & 0 & -1 & 4 \end{pmatrix} \text{ et } B = \begin{pmatrix} 3 & 2 & 2 & -3 & -1 \\ 1 & 2 & 3 & 4 & 5 \\ -2 & -2 & -1 & 0 & -3 \\ 7 & 6 & 5 & 4 & 3 \end{pmatrix}$$

(a) À la main.
(b) En SCILAB.

7. Résoudre le système suivant :

$$\begin{cases} -x + 2\,y + 2\,z = 4 \\ 2\,x - y - 2\,z = -5 \\ 2\,x + 2\,y + z = 0 \end{cases}$$

(a) Sur feuille en détaillant les diverses étapes (opérations sur les lignes)
(b) En écriture matricielle (en respectant les mêmes étapes que précédemment).

### 7.1.2 Normes vectorielles et matricielles

Rappels

- Rappeler les propriétés qui définissent une norme vectorielle.
- Quelle propriété supplémentaire a-t-on pour une norme matricielle ?
- On rappelle que pour une norme vectorielle donnée $||.||$, la norme matricielle induite est définie par :
  $||A|| = \sup_{x \neq 0} \frac{||Ax||}{||x||} = \sup_{||x||=1} ||Ax||.$

On note $A = (a_{i,j})_{1 \leq i,j \leq n}$ une matrice de $\mathcal{M}_n(\mathbb{C})$ et $x = (x_1, \ldots, x_n)$ un vecteur de $\mathbb{C}^n$.

1. Équivalence des normes en dimension finie
   (a) Montrer les inégalités suivantes et vérifier qu'elles sont optimales :
   - $\forall x \in \mathbb{C}^n$ , $||x||_1 \leq n\,||x||_\infty$
   - $\forall x \in \mathbb{C}^n$ , $||x||_\infty \leq ||x||_1$
   - $\forall x \in \mathbb{C}^n$ , $||x||_\infty \leq ||x||_2$
   - $\forall x \in \mathbb{C}^n$ , $||x||_2 \leq \sqrt{n}\,||x||_\infty$
   - $\forall x \in \mathbb{C}^n$ , $||x||_2 \leq ||x||_1$
   - $\forall x \in \mathbb{C}^n$ , $||x||_1 \leq \sqrt{n}\,||x||_2$

   (b) En déduire les cœfficients $\alpha$ et $\beta$ qui montrent l'équivalence de ces différentes normes sur $\mathbb{C}^n$.
2. Vérifier que la norme qui à la matrice $A$ associe $\max_{1 \leq i,j \leq n} |a_{i,j}|$ est bien une norme sur l'espace $\mathcal{M}_n(\mathbb{C})$ mais pas une norme matricielle.
   *On pourra par exemple utiliser la matrice dont tous les cœfficients sont égaux à 1.*
3. (a) Calcul de la norme matricielle induite pour $||.||_1$ : montrer que $||A||_1 = \max_{j \in [[1,n]]} \left( \sum_{i=1}^{n} |a_{i,j}| \right)$.

   (b) Calcul de la norme matricielle induite pour $||.||_\infty$ : montrer que $||A||_\infty = \max_{i \in [[1,n]]} \left( \sum_{j=1}^{n} |a_{i,j}| \right)$.

   *On remarque que* $||A||_1 = ||\,{}^t A\,||_\infty$

   (c) Si $||.||$ est une norme induite qulconque, que vaut la norme de l'identité ?
4. On définit la norme de Frobénius d'une matrice $A$ par : $||A|| = \left( \sum_{1 \leq i,j \leq n} |a_{i,j}|^2 \right)^{1/2}$
   (a) Vérifier que la norme de Frobénius est bien une norme matricielle.
   (b) Que vaut la norme de l'identité ?
   (c) La norme de Frobénius peut-elle être une norme induite ?
5. Définition : Une norme matricielle $||.||_M$ et une norme vectorielle $||.||_V$ sont compatibles si et seulement si $\forall A \in \mathcal{M}_n(\mathbb{C})$ , $\forall x \in \mathbb{C}^n; ||A\,x||_V \leq ||A||_M.||x||_V$
   Remarque : Une norme vectorielle $||.||_V$ et sa norme matricielle induite $||.||_M$ sont toujours compatibles.
   (a) Montrer que la norme de Frobénius et la norme $||.||_2$ sont compatibles.

(b) Prouver le théorème suivant : pour toute norme matricielle, il existe une norme vectorielle qui lui est compatible.

*Indication : à $x$ un vecteur donné, on associe la matrice carrée $X$ dont la première colonne est $x$ et les colonnes suivantes sont nulles ; on définit alors une norme vectorielle par $||x||_V = ||X||_M$.*

Il s'agit de vérifier que $||.||_V$ est bien une norme, puis la compatibilité.

### 7.1.3 Problèmes de conditionnement

Rappels

- Soit $A$ une matrice inversible et $||.||$ une norme matricielle, le conditionnement de $A$ associé à cette norme est cond$(A)= ||A|| \times ||A^{-1}||$.
- Si $A$ est une matrice ayant pour valeurs propres $\lambda_1$,...,$\lambda_n$, son rayon spectral est $\rho(A) = \max\limits_{i\in[[1,n]]} |\lambda_i|$.

1. (a) Soit $A$ une matrice, montrer que la suite $A^n$ converge vers la matrice nulle si et seulement si $\rho(A) < 1$.

   (b) En déduire que la série $\sum\limits_{k\in\mathbb{N}} A^k$ converge si et seulement si $\rho(A) < 1$, vers $(I - A)^{-1}$.

2. Montrer que pour toute norme matricielle, $\lim\limits_{k\to+\infty} ||A^k||^{1/k} = \rho(A)$.

3. Calculer le rayon spectral de la matrice $A = \begin{pmatrix} 5 & 9/2 & -3/2 \\ -2 & -3/2 & 3/2 \\ 2 & 7/2 & 1/2 \end{pmatrix}$

4. En utilisant l'équivalence des normes $||.||_1$, $||.||_2$ et $||.||_\infty$, montrer que les conditionnements correspondants sont équivalents et déterminer les cœfficients $\alpha$ et $\beta$ correspondants.

5. Décomposition "$UDR$" : soit $A$ une matrice carrée à cœfficients réels, on veut montrer qu'il existe deux matrices unitaires $U$ et $R$ et une matrice diagonale $D$ à termes positifs telles que $A = U.D.R$

   (a) Justifier l'existence d'une matrice diagonale $\Delta$ et d'une matrice unitaire $R$ telles que : ${}^tA \times A = {}^tR \,.\, \Delta \,.\, R$

   (b) À quoi est égale la matrice $D$ ?

   (c) Vérifier que $A.\ {}^tR\ .\ D^{-1}$ est une matrice unitaire et conclure.

6. Montrer que pour toute matrice $A$ de $\mathcal{M}_n(\mathbb{R})$, $||A||_2 = \sqrt{\rho({}^tA\ A)}$.

7. Montrer que pour la norme 2, le conditionnement d'une matrice est le quotient de la plus grande valeur singulière de $A$ par la plus petite valeur singulière de $A$.

   *les valeurs singulières d'une matrice $A$ sont les racines carrées des valeurs propres de ${}^tA \times A$.*

8. Montrer que pour une norme induite quelconque, le conditionnement vérifie : cond$(A) \geq \rho(A) \,.\, \rho(A^{-1})$.

9. Pour $n \in \mathbb{N}^*$, on définit la matrice de Hilbert d'ordre $n$ la matrice $H_n = (h_{i,j})_{1\leq i,j\leq n}$ telle que $h_{i,j} = \frac{1}{i+j-1}$

   *En* SCILAB*, l'inverse de la matrice de Hilbert d'ordre $n$ est obtenue par la commande testmatrix('hilb',n).*

(a) Calculer le conditionnement de $H_3$, de $H_{10}$ pour la norme 2.

(b) À l'aide d'un programme SCILAB, tracer le logarithme du conditionnement de $H_n$ en fonction de $n$ ($n \leq 150$ environ)

### 7.1.4 Localisation des valeurs propres d'une matrice

**1re partie : théorème de Gerschgörin – Hadamard**

Soit $A = (a_{i,j})_{1 \leq i,j \leq n}$ une matrice carrée d'ordre $n$.

Pour $k \in [[1, n]]$, on définit les disques de Gerschgörin $D_k$ associés à $A$ par :

$$z \in D_k \Leftarrow : |z - a_{k,k}| \leq \Lambda_k = \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{k,j}|$$

Toutes les valeurs propres de $A$ appartiennent à la réunion des disques de Gerschgörin associés à $A$.

1. Démonstration du théorème :

   Soit $\lambda$ une valeur propre quelconque de $A$ et $u$ un vecteur propre associé à $\lambda$. Quitte à prendre un multiple de $u$, on peut supposer que le plus grand module de ses composantes est égal à 1 ; on note $k$ un indice tel que $|u_k| = 1$.

   Montrer que $|\lambda - a_{k,k}| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{k,j}|$ et conclure.

2. Étude d'un exemple :

   Soit $A = \begin{pmatrix} 1+i & i & 2 \\ -3 & 2+i & 1 \\ 1 & i & 6 \end{pmatrix}$

   (a) Dessiner dans le plan complexe, les trois disques de Gerschgörin associés à la matrice $A$.

   (b) Remarquer que $A$ et ${}^tA$ ont les mêmes valeurs propres, et représenter de même les trois disques de Gerschgörin associés à la matrice ${}^tA$.

   (c) En déduire une majoration du rayon spectral de $A$.

3. Une application : les matrices à diagonale dominante

   Définition : Soit $A \in \mathcal{M}_n(\mathbb{C})$, $A$ est à diagonale dominante (resp. strictement dominante) si et seulement si $\forall i \in [[1, n]]$, $|a_{i,i}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|$ (resp. $|a_{i,i}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|$)

   (a) Montrer, en utilisant le théorème de Gerschgörin – Hadamard, que si $A$ est à diagonale strictement dominante, alors elle est inversible.

   (b) Vérifier que la matrice suivante est à diagonale dominante, mais non inversible :

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

(c) Montrer qu'on peut généraliser ce résultat à une matrice à cœfficients réels, à diagonale dominante, et qui vérifie :

$$\forall i \in [[1,n]]\,,\ a_{i,i} \geq 0, \forall j \neq i\,,\ a_{i,j} \leq 0 \ \text{ et } \sum_{j=1}^{n} a_{i,j} = 0$$

**2e partie : nombre d'opérations dans la résolution d'un système linéaire**

Soit le système $(\mathcal{S})$ : $AX = B$ où $A \in \mathcal{M}_n(\mathbb{C})$, $B \in \mathbb{C}^n$ et $X = (x_1, \ldots, x_n) \in \mathbb{C}^n$.

On se propose d'évaluer le nombre d'opérations nécessaires à la résolution du système $(\mathcal{S})$ par les méthodes de Gauss et de Cholesky ; cette évaluation a été faite en cours pour la méthode du déterminant.

1. Méthode de Gauss : On néglige dans un premier temps les permutations éventuelles dues à la recherche du meilleur pivot.
   (a) Élimination de $x_1$ dans les équations 2 à $n$ : compter séparément le nombre de divisions, de multiplications et d'additions nécessaires.
   Une division coûtant *plus cher* qu'une multiplication, expliquer comment on peut remplacer la plupart des divisions par des multiplications.
   (b) Élimination de $x_k$ dans les équations $k+1$ à $n$ (pour $1 \leq k \leq n-1$) : procéder de même et compter le nombre total d'opérations de chaque sorte pour la phase de triangulation.
   (c) Pour la phase de remontée, déterminer de même le nombre d'opérations de chaque sorte.
   (d) Évaluer le nombre maximal de permutations nécessaires, puis déterminer le nombre total de chaque type d'opération.
2. Méthode de Cholesky (cas des matrices symétriques définies positives) : On cherche une matrice triangulaire inférieure $L$ telle que $A = L \,.\, {}^tL$. On a vu en cours que cette décomposition existe si et seulement si $A$ est symétrique définie positive ; qu'elle est alors unique et que de plus, $L$ est à diagonale strictement positive.
   (a) En posant $A = (a_{i,j})_{1 \leq i,j \leq n}$ et $L = (b_{i,j})_{1 \leq j \leq i \leq n}$, ($b_{i,j} = 0$ si $i < j$), montrer que l'on doit résoudre :
   $$\forall (i,j) \in [[1,n]]^2\,,\ a_{i,j} = \sum_{k=1}^{n} b_{i,k} b_{j,k} = \sum_{k=1}^{\min(i,j)} b_{i,k} b_{j,k}$$
   (b) En raisonnant par ligne (ou par colonne, ce qui revient aau même puisque $A$ est symétrique), compter les opérations nécessaires à la détermination de $L$.
   (c) Déterminer ensuite le nombre d'opérations pour substituer et résoudre le système.
3. Peut-on comparer les deux méthodes ?

### 7.1.5 Formes quadratiques

**1re partie : Formes quadratiques**

Définition 1 : Soit $A$ une matrice symétrique à cœfficients réels, on appelle *forme quadratique* associée
à $A$ l'application de $\mathbb{R}^n$ vers $\mathbb{R}$ : $x \longmapsto {}^tx\,A\,x$.

1. Expliciter la formule ${}^tx\,A\,x$ en fonction des cœfficients de $A$ et des composantes $(x_1, \ldots, x_n)$ de $x$.
2. Définition 2 : Une base $\mathcal{B} = (e_i)_{1 \leq i \leq n}$ est dite orthogonale pour une forme quadratique $q$ si et seulement si $q$ s'écrit dans la base $\mathcal{B}$ $q(x) = \sum_{i=1}^{n} \alpha_i\, x_i^2$, les $\alpha_i$ étants des réels à déterminer.

Procédé de réduction de Gauss : *(rien à voir avec la méthode du pivot de Gauss)*
À toute forme quadratique on peut associer des bases orthogonales (une infinité), et le procédé suivant permet d'en construire une.

On commence par les exemples suivants : $q_1(x) = 2\,x_1^2 + 4\,x_2^2 + x_3^2 + 8\,x_1\,x_2 - 4\,x_1\,x_3 - 6\,x_2\,x_3$
et $q_2(x) = 2\,x_1\,x_2 - 4\,x_1\,x_3 + 6\,x_1\,x_4 + 8\,x_2\,x_3 - 4\,x_2\,x_4 + 4\,x_3\,x_4$

Méthode générale : on procède par récurrence sur le nombre $k$ de variables, et pour un nombre donné, on regarde s'il existe un cœfficient diagonal non nul ou pas. Dans chacun de ces deux cas, en utilisant un genre de forme canonique, on s'arrange pour "diminuer" de 1 ou 2 le nombre de variables.

**2e partie : Matrices symétriques définies positives**

Définition 3 : Soit $A$ une matrice symétrique à cœfficients réels, elle est positive (resp définie positive) si et seulement si sa forme quadratique associée vérifie :

$$\forall x \in \mathbb{R}^n \setminus \{0\}\,,\ {}^t x\,A\,x \geq 0 \quad (\text{resp. } {}^t x\,A\,x > 0)$$

1. Montrer qu'une telle matrice a tous ses termes diagonaux positifs (strictement positifs dans le cas d'une matrice symétrique définie positive).
2. Montrer de même que toutes les sous–matrices principales de $A$ sont aussi positives (resp. définies positives).
   *On peut obtenir d'autres relations entre les cœfficients de la matrice, ce qui est en accord avec les conditions d'existence de la matrice $L$ dans la factorisation de Cholesky.*
3. Montrer qu'il existe une matrice unitaire $U$ et une matrice diagonale $\Delta$ dont les termes diagonaux sont positifs (resp. strictement positifs) telles que $A = {}^t U\,\Delta\,U$. Autrement dit, il existe une base orthonormée de vecteurs propres pour $A$, de plus ses valeurs propres sont réelles positives (resp. réelles strictement positives). *Il est à noter que cette diagonalisation n'est pas en général la même que celle obtenue par le procédé de réduction de Gauss.*

### 7.1.6 Interpolation polynômiale

Rappels :

- Soit $f$ une fonction définie sur un intervalle $I = [a, b]$ et $x_0, \ldots, x_k$, $k+1$ points distincts de $I$. On cherche s'il existe un polynôme $P_k$ de degré minimal vérifiant $\forall i \in [[0, k]]\,,\ P_k(x_i) = f(x_i)$.
  On pose $\forall i \in [[0; i]]\ ,\ L_i(X) = \prod_{j=0}^{k} \left(\frac{X - x_j}{x_i - x_j}\right)$.
- Les polynômes $L_i$ s'appellent les polynômes d'interpolation de Lagrange associés à $f$ et à $(x_0, x_1, \ldots, x_k)$.
  On vérifie que $\forall (i, j) \in [[0; k]]^2\,,\ L_i(x_j) = \delta_{i,j}$ *(symbole de Kronecker)*, donc le polynôme $P = \sum_{i=0}^{k} f(x_i)\,L_i$ est solution du problème posé.
- La méthode de Lagrange pose le problème suivant : si on souhaite rajouter des points, il faut recommencer tous les calculs. On fait alors la remarque suivante : Pour calculer le polynôme

$P_{k+1}$ à partir de $P_k$, on définit $R_{k+1} = a_{k+1}\prod_{i=0}^{k}(X - x_i)$ qui s'annule pour tous les $x_i$, $0 \leq i \leq k$ et vaut 1 pour $X = x_{k+1}$ ; cela entraine que $a_{k+1} = \prod_{i=0}^{k} \frac{f(x_i) - f(x_{k+1})}{x_i - x_{k+1}}$

On peut alors écrire $P_{k+1} = P_k + R_{k+1}$

Le cœfficient $a_k$ s'appelle la $k^{\text{e}}$ différence divisée de $f$ et se note $f[x_0, \ldots, x_k]$.

De plus, on a la relation de récurrence $f[x_0, \ldots, x_{k+1}] = \frac{f[x_0, \ldots, x_k] - f[x_1, \ldots, x_{k+1}]}{x_0 - x_{k+1}}$.

Cette notation peut se généraliser à des points non distincts : $f[x_0, \ldots, x_0] = \frac{f^{(n)}(x_0)}{n!}$

1. Les polynômes d'interpolation d'Hermite sont une généralisation de ceux de Lagrange : on considère $k+1$ points d'un intervalle $I = [a, b]$, $f$ une fonction suffisamment dérivable sur $I$ et $\alpha_0, \ldots, \alpha_k$, $k+1$ entiers naturels. Enfin, on pose $n = k + \alpha_0 + \cdots + \alpha_k$. Si $f$ admet des dérivées d'ordre $\alpha_i$ aux points $x_i$, alors il existe un unique polynôme $P_n$ de degré inférieur ou égal à $n$ tel que :
$$\forall i \in [[0, k]]\,,\ \forall j \in [[0, \alpha_i]]\,,\ P_n^{(j)}(x_i) = f^{(j)}(x_i)$$
Le calcul explicite de ce polynôme est assez compliqué dans le cas général, on va donc se limiter à un exemple :
2. Soit $f$ une fonction dérivable sur $[a, b]$, $h_3$ est le polynôme d'Hermite d'ordre 3 associé à $f$, à savoir :
$$h_3(a) = f(a)\ ,\ h_3'(a) = f'(a)\ ,\ h_3(b) = f(b)\ ,\ h_3'(b) = f'(b)$$
   (a) Déterminer ce polynôme pour $a = 0$ et $b = 1$
   *on écrira $h_3$ sous la forme : $\alpha_0 + \alpha_1\, x + \alpha_2\, x^2 + \alpha_3\, x^2(x-1)$*
   (b) Généraliser à $a$ et $b$ quelconques, puis écrire les cœfficients $\alpha_i$ en fonction des différences divisées à arguments éventuellement répétés de $f$.
   (c) Donner une majoration de l'erreur sur $[0, 1]$, puis dans le cas général.
   (d) On considère la fonction sinus hyperbolique et on donne sh $(1) \simeq 1,1752$, sh'$(1) \simeq 1,5431$ sh $(0) = 0$ et sh'$(0) = 1$
   Comparer l'interpolation linéaire et celle avec le polynôme d'Hermite déterminé à la question 2a pour déterminer une valeur approchée de sh $(0,5)$.
   (e) Vérifier que la majoration de l'erreur calculée prédédemment est en accord avec les résultats obtenus à la question 2c.
3. Trouver un polynôme de degré minimal vérifiant : $P(1) = -1$, $P'(1) = 2$, $P''(1) = 1$, $P(2) = 1$ et $P'(2) = -2$
4. Un petit exercice pour s'entrainer
   Déterminer la décomposition de Cholesky pour la matrice tridiagonale :

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

### 7.1.7 Divers

**Exercice 1.** Soit $A \in M_n(\mathbb{R})$. On suppose $A$ inversible. Soit $\omega > 0$, on considère la méthode itérative suivante

$$x_0 \in \mathbb{R}^n \text{quelconque},$$
$$x_{n+1} = x_n - \omega(Ax_n - b)$$

A quelle condition sur $\omega$ la méthode converge ?

**Exercice 2.** Soit $B$ la matrice carrée d'ordre $n$

$$B = \begin{pmatrix} 0 & F \\ F^T & 0 \end{pmatrix}$$

où $F$ est une matrice à $k$ lignes et $n-k$ colonnes. On considère le système $(Id - B)x = b$.

1. Ecrire la matrice $J$ de la méthode de Jacobi et la matrice $\mathcal{L}$ de la méthode de Gauss Seidel.
2. Que peut–on dire de $\rho(J)$ et $\rho(\mathcal{L})$ ?
3. Montrer que
$$||\mathcal{L}^k||_2 = \rho(B)^{2k-1}\sqrt{1+\rho^2(B)}.$$

**Exercice 3.** Soient $x_0, x_1, x_2$ trois réels distincts.

1. Ecrire le polynôme d'interpolation de Lagrange $P$ associé à $(x_0, y_0)$, $(x_1, y_1)$, $(x_2, y_2)$.
2. On suppose que $y_i = f(x_i)$, où $f$ est une fonction de classe $C^3$. On admet que $P'$ et $P''$ est une bonne approximation de $f'$ et $f''$. Déterminer diverses approximations de $f'(x_1)$ et $f''(x_1)$.

**Exercice 4** Cet exercice propose un algorithme de calcul du polynôme d'interpolation de Lagrange différent de celui reposant sur l'emploi de différence divisées. Il s'agit de la méthode d'Aitken.
**Notations.** On note par $x_0, x_1, \dots, x_n$ les points d'interpolations et $y_0, y_1, \dots, y_n$ les valeurs en ces points. On note par $P_{p,\dots,q}$ le polynôme interpolant aux nœuds $\{x_p, \dots, x_q\}$ (il est de degré $q-p-1$, par $P_{p,\dots\hat{\pi}\dots,q}$ le polynôme interpolant aux nœuds $\{x_p, \dots, x_q\}$ le nœuds $\pi$ étant exclu. C'est un polynôme de degré $q-p-2$. Le polynôme est indépendant de la numérotation des points le définissant.

1. Montrer que quelque soit $x_\pi$ et $x_\sigma$ distincts,
$$(x_\pi - x_\sigma)P_{p\dots q} = \det \begin{pmatrix} x_\sigma - x & P_{p\dots\hat{\pi}\dots q} \\ x_\pi - x & P_{p\dots\hat{\sigma}\dots q} \end{pmatrix}.$$
On examinera les cas $x = x_\sigma$, $x = x_\nu$ et $x \notin \{x_\sigma, x_\nu\}$.
2. En déduire une méthode de calcul.

3. Le point $x \in [\min_i x)i, \max_i x_i]$ étant donné, comment ordonner les calculs pour avoir les résultats les plus stables.
4. Quel est le coût de cette méthode ?

**Exercice 5.** On considère les zéros du $n+1^{\text{e}}$polynôme de Legendre $P_{n+1}$,

$$P_{n+1} = a_{0,n}\Pi_{j=0}^{n}(x - x_j).$$

On pose

$$Q_k(x) = \frac{P_{n+1}(x)}{x - x_k}.$$

On sait que les poids de la formule de Gauss qui lui est associée sont donnés par

$$\lambda_k = \frac{1}{Q_k(x_k)} \int_{-1}^{1} Q(x)dx.$$

Le but de cet exercice est de donner deux expressions «faciles» de $\lambda_k$, c'est à dire ne nécessitant pas de calcul d'intégrales. On admet la relation suivante (voir Guilpin, Méthodes du Calcul Numérique appliqué, page 113)

$$\text{quelque soit } k,\ (1-x^2)P_k' - kxP_k + kP_{k-1} = 0. \tag{7.1}$$

1. Vérifier que
$$\lambda_k = \frac{1}{P_{n+1}'(x_k)} \int_{-1}^{1} \frac{P_{n+1}(x)}{x - x_k}dx$$
puis que
$$\int_{-1}^{1} P_{n+1}(x)\frac{P_n(x) - P_n(x_k)}{x - x_k}dx = 0.$$
2. Vérifier que
$$\int_{-1}^{1} \frac{P_{n+1}(x)P_x(x)}{x - x_k}dx = a_{0,n}\int_{-1}^{1} x^n P_n dx$$
et en déduire que
$$\int_{-1}^{1} \frac{P_{n+1}(x)P_x(x)}{x - x_k}dx = \frac{a_{0,n+1}}{a_{0,n}}\frac{2}{n+1}$$
3. En déduire que
$$\lambda_k = \frac{1}{P_{n+1}'(x_k)}\frac{1}{P_n(x_k)}\frac{2}{n+1}.$$
4. En utilisant ce résultat et la relation (7.1), montrer que
$$\lambda_k = \frac{2}{1-x_k^2}\left(P_n'(x_k)\right)^2.$$
On voit en particulier que $\lambda_k > 0$.

**Solution de l'exercice 2.** On a

$$Id - B = \begin{pmatrix} Id_k & -F \\ -F^T & Id_{N-k} \end{pmatrix}$$

où on a posé $Id_p$ = matrice identité de $\mathbb{R}^p$.

L' itération de Jacobi est donc

$$x^{(n+1)} = Bx^{(n)} + b$$

d'où $J = B$, et celle de Gauss–Seidel est

$$\begin{pmatrix} Id_k & 0 \\ -F^T & Id_{N-k} \end{pmatrix} x^{(n+1)} = \begin{pmatrix} 0 & F \\ 0 & 0 \end{pmatrix} x^{(n)} + b$$

Puisque

$$\begin{pmatrix} Id_k & 0 \\ -F^T & Id_{N-k} \end{pmatrix}^{-1} = \begin{pmatrix} Id_k & 0 \\ F^T & Id_{N-k} \end{pmatrix},$$

on a

$$\mathcal{L} = \begin{pmatrix} Id_k & 0 \\ -F^T & Id_{N-k} \end{pmatrix}^{-1} \begin{pmatrix} 0 & F \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} Id_k & 0 \\ F^T & Id_{N-k} \end{pmatrix} \begin{pmatrix} 0 & F \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & F \\ 0 & F^TF \end{pmatrix}$$

Le rayon spectral de $J$ est celui de $B$. Le rayon spectral de $\mathcal{L}$ est $\rho(FF^T)$.

On calcule directement $\rho(B)$.

$$\begin{pmatrix} 0 & F \\ F^T & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

donne

$$\begin{aligned} Fx_2 &= \lambda x_1 \\ F^Tx_1 &= \lambda x_2 \end{aligned}$$

soit $F^TFx_2 = \lambda^2 x_2$ Comme les valeurs propres de $B$ sont réelles (matrice symétrique), on voit que

$$\rho(B)^2 = \rho(FF^T). \tag{7.2}$$

Finalement, on a $\rho(\mathcal{L}) = \rho(J)^2$. Les deux méthodes convergent ou divergent simultanément. Si elles convergent, Gauss Seidel converge deux fois plus vite.

On calcule par récurence $\mathcal{L}^k$. On a

$$\mathcal{L}^2 = \begin{pmatrix} 0 & F \\ 0 & F^TF \end{pmatrix}$$

On montre que

$$\mathcal{L}^k = \begin{pmatrix} 0 & F(F^TF)^{k-1} \\ 0 & (F^TF)^k \end{pmatrix}$$

C'est vrai pour $k = 1, 2$. On voit que

$$\begin{pmatrix} 0 & F(F^TF)^{k-1} \\ 0 & (F^TF)^k \end{pmatrix} \begin{pmatrix} 0 & F \\ 0 & F^TF \end{pmatrix} = \begin{pmatrix} 0 & F(F^TF)^k \\ 0 & (F^TF)^{k+1} \end{pmatrix}$$

d'où le résultat.

Enfin si $A = F^T F$ (qui est une matrice symétrique positive)

$$(\mathcal{L}^k)^T \mathcal{L}^k = \begin{pmatrix} 0 & 0 \\ A^{k-1}F^T & A^k \end{pmatrix} \begin{pmatrix} 0 & FA^{k-1} \\ 0A^k & \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & A^{k-1}F^T F A^{k-1} + A^2 k \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & A^{2k-1} + A^2 k \end{pmatrix}$$

La matrice $A$ est positive, donc

$$\rho((\mathcal{L}^k)^T \mathcal{L}^k) = \rho(A)^{2k-1} + \rho(A)^{2k}$$

car $A^{2k-1}$, $A^{2k}$ et $A$ commutent. Puisque $\rho(A) = \rho(B)^2$, on a le résultat.

### 7.1.8 Problème

**Question de cours** On considère les normes suivantes définies sur $\mathbb{R}^n$ :

$$||x||_\infty = \max_{1 \leq i \leq n} |x_i| \,, \qquad ||x||_1 = \sum_{i=1}^{n} |x_i| \quad \text{et} \qquad ||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Soit $A = (a_{ij})_{1 \leq i,j \leq n}$ une matrice $(n, n)$. Rappelez *et démontrez* les formules des normes induites associées.

**Exercice 1** On considère la matrice :

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

1. Déterminer le conditionnement $L^2$ de $A$.
2. Soit
$$D = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$
Déterminer le conditionnement $L^2$ de $D^{-1}A$.
3. Quelle remarque peut-on faire ?

**Exercice 2** Soit $A \in M_{n+p}(\mathbb{R})$ la matrice bloc suivante :

$$A = \begin{pmatrix} C & D^T \\ D & 0 \end{pmatrix}$$

où $C \in \mathcal{M}_n(\mathbb{R})$ et $D \in \mathcal{M}_{p,n}(\mathbb{R})$.[1] Les matrices $C$ et $D$ ne sont pas nécessairement inversibles.

Dans la suite on suppose $C$ inversible.

[1] Un exemple d'une telle matrice est

$$B = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{où} \quad C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ et } D = (1, 0)$$

.

1. Montrer que $A$ est inversible si et seulement si $K = DC^{-1}D^T$ est inversible.
   *On pourra par exemple résoudre le système matriciel suivant :*
   $\begin{cases} C.X + D^T.Y &= 0 \\ D.X &= 0 \end{cases}$ où $X \in \mathbb{R}^n$ et $Y \in \mathbb{R}^p$
   *en se souvenant que $D$ n'est pas en général inversible.*
2. On suppose à présent que $K$ est inversible et on considère la matrice $M^{-1}A$ où

$$M = \begin{pmatrix} C & 0 \\ 0 & K \end{pmatrix}$$

   (a) Soit $\lambda$ une valeur propre de $M^{-1}A$ et $(X, Y) \in R^n \times \mathbb{R}^p$ un vecteur propre associé à $\lambda$. Montrer que :

$$(\lambda^2 - \lambda - 1)D.X = 0$$

   (b) En déduire les valeurs propres de $M^{-1}A$.

   (c) On suppose à présent que $M^{-1}A$ est symétrique.
   – À quelle condition cela est–il possible ? *(cette question est indépendante de la suivante)*
   – Déterminer un majorant du conditionnement $L^2$ de $M^{-1}A$.

## 7.2 Suggestion de programmes SCILAB

### 7.2.1 Factorisation LU

```
  //Essai de factorisation LU

//Choix de la matrice A :
A=[2,4,-4,1;3,7,1,-2;-1,1,2,3;1,1,-4,1]

n=4;
for k=1:n-1     //kiÃÍme Ãl'tape
   for i=k+1:n  //iiÃÍme ligne de A
      for j=k+1:n
          A(i,j)=A(i,j)-A(i,k)*A(k,j)/A(k,k);    //combinaison des lignes i et k
      end
      A
   end
   for i=k+1:n
      A(i,k)=A(i,k)/A(k,k)  //calcul de la kiÃÍme colonne de L
   end
end
A
```

### 7.2.2 Erreurs d'arrondi

```
//Feuille d'exercices 1
//Exercice 2
```

```
//Calcul de ((y+x)-x)/y et de (y+(x-x))/y
//et tests pour diffÃĽrentes valeurs de x et de y

deff('[A]=somme1(x,y)','A=((y+x)-x)/y')

deff('[B]=somme2(x,y)','B=(y+(x-x))/y')


x=input('entrer une valeur de x : ');
y=input('entrer une valeur de y : ');
somme1(x,y)
somme2(x,y)
//Pour des valeurs de y très faibles par rapport àă x, on obtient 0 pour la quantité A
//et 1 pour B, alors que A et B sont en fait tous les deux égaux Ãă 1
```

### 7.2.3 Calcul de coefficients binomiaux

```
//Feuille d'exercices 1
//Exercice 3
//Calcul des coefficients binomiaux :

//simplification : n!/(k!(n-k)!)= prod(n-k+1:n)/prod(1:k) ou bien prod(k+1:n)/prod(1:n-k)
//Ãă choisir selon les valeurs de k.

//Fonction qui calcule le coefficient binomial :
function [calcul]=binome(n,k)
if (k>n|k<0) then
calcul='erreur'
elseif k <=(n/2) then
calcul=prod(n-k+1:n)/prod(1:k)
else
calcul=prod(k+1:n)/prod(1:n-k)
end
endfunction
```

### 7.2.4 Erreurs d'arrondi : calcul de $(1-x)^n$, méthode d'Horner

```
//Feuille d'exercices 1
//Exercice 4 -- généralisation
//Calcul de (1-x)^n de trois fçons différentes

//Calcul de y=1-x puis y :

function [fonction1]=puissance1(x,n)
```

```
y=1-x;
fonction1=y.^n;
endfunction

//Calcul en utilisant la formule du binôme de Newton :
function [fonction2]=puissance2(x,n)
aux=0;
for i=0:n
aux=aux+binome(n,i)*(-x).^i;
end
fonction2=aux
endfunction

//Méthode de Horner :
function [fonction3]=puissance3(x,n)
aux=(-1)^n;
for i=1:n
aux=aux.*x+(-1)^(n-i)*binome(n,i)
end
fonction3=aux;
endfunction

n=input('entrer une valeur de n : ');
u=0.99:0.00001:1.01;

plot2d(u,puissance2(u),style=2)
plot2d(u,puissance1(u),style=3)
//pause
plot2d(u,puissance3(u),style=4)
plot2d(u,puissance1(u),style=3)
```

### 7.2.5 TD 1,Exercice 4, suite

```
//Feuille d'exercices 1
//Exercice 4
//Calcul de (1-x)^7 de trois façons différentes

//Calcul de y=1-x puis y :

function [fonction1]=puissance1(x)
y=1-x;
fonction1=y.^7;
endfunction

//Calcul en utilisant la formule du binôme de Newton :
function [fonction2]=puissance2(x)
```

```
aux=0;
for i=0:7
aux=aux+binome(7,i)*(-x).^i;
end
fonction2=aux
endfunction

//MÃ©thode de Horner :
function [fonction3]=puissance3(x)
aux=-1;
for i=1:7
aux=aux.*x+(-1)^(7-i)*binome(7,i)
end
fonction3=aux;
endfunction

//TracÃ©s sur l'intervalle [0,99 ; 1,01]
u=0.99:0.00001:1.01;

plot2d(u,puissance2(u),style=2)
plot2d(u,puissance1(u),style=3)
//pause
plot2d(u,puissance3(u),style=4)
plot2d(u,puissance1(u),style=3)
```

### 7.2.6 Conditionnement : Matrice de Hilbert

```
//Calcul du conditionnement de la matrice de Hilbert d'ordre n
//Comportement asymptotique

clear all
n=30;

for i=1:n
   H=testmatrix('hilb',i);
   conditionnement(i)=log(cond(H));
end

//TracÃ© de ln(cond(H_n)) en fonction de n :

plot2d(1:n,conditionnement)

xtitle('Trace de ln(Hn) en fonction de n')
```

### 7.2.7 Théorème de Gersgorin

```
//Calcul des disques de Gershgörin pour la matrice de la partie 1 :

A=[1+%i,%i,2;-3,2+%i,1;1,%i,6]

u=spec(A)

//TracÃľ des disques de Gershgörin associés Ãă A :
// D1 : |z-(1+i)|<=3
// D2 : |z-(2+i)|<=4
// D3 : |z-6|<=2

plot2d(0,0,0,"032"," ",[0,-12,12,9])

D= [-2 -2  4; //abscisses des coins en haut àă gauche,
     4  5  2; //ordonnÃľes des coins en haut àă gauche,
     6  8 12 ; //largeurs,
     6  8 12 ; //hauteurs
     0  0  0 ; //angles dÃľbut,
  360*64 360*64 360*64]; //angles fin
  //xarcs(D,[1,2,3])
  xfarcs(D,[4,4,4])
//TracÃľ des disques de GershgÃűrin associées àă la transposée de A :
// Delta1 : |z-(1+i)|<=4
// Delta2 : |z-(2+i)|<=2
// Delta3 : |z-6|<=3

Delta=[-3  0  3; //abscisses des coins en haut àă gauche,
        5  3  3; //ordonnÃľes des coins en haut àă gauche,
        8  4  6 ; //largeurs,
        8  4  6 ; //hauteurs
        0  0  0 ; //angles dÃľbut,
  360*64 360*64 360*64]; //angles fin

  xfarcs(Delta,[7,7,7])
  xarcs(D,[1,1,1])
  //xarcs(Delta,[2,2,2])
  plot2d([-10 20],[0 0])
plot2d([0 0],[-10 10])

  xtitle("Disques de Gershgoring de la matrice A")
```

### 7.2.8 Interpolation

```
//TracÃľ de la fonction sinus hyperbolique sur [0,1], de son interpolation linéaire
```

```
//et de son polynÃťme d'Hermite

deff('[approx1]=affine(x)','approx1=1.1752*x ')
deff('[approx3]=hermite(x)','approx3=x+0.1752*(x.^2)+0.1927*(x.^2).*(x-1)')
v=0:0.01:1;

plot2d(v,[sinh(v'),affine(v'),hermite(v')],[1,2,3])
legends(['sh(x)';'affine(x)';'h3(x)'],[1,2 3],opt="lr")


//t=0:0.1:2*%pi;
//plot2d(t,[cos(t'),cos(2*t'),cos(3*t')],[-1,2 3]);
//legends(['cos(t)';'cos(2*t)';'cos(3*t)'],[-1,2 3],opt="lr")
//xset("line style",2);plot2d(t,cos(t),style=5);
//xset("line style",4);plot2d(t,sin(t),style=3);
//legends(["sin(t)";"cos(t)"],[[5;2],[3;4]], with_box=%f, opt="?")
```

# Bibliographie


[1] P. Lascaux and R. Théodor. *Analyse Numérique matricielle appliquée à l'art de l'ing'enieur, Tomes 1 et 2.* Masson, 1986.

[2] R. S. Varga. *Matrix Iterative Analysis.* Series in Automatic Computation. Prentice–Hall Inc, 1962.

[3] G. Allaire and S.M. Kaber. *Algèbre linéaire numérique.* Ellipse, mathématiques 2$^{e}$cycle edition, 2002.

[4] P.G. Ciarlet. *Analyse numérique matricielle et optimisation.* Collection Mathématiques Appliquées pour la maîtrise. Masson, 1984.

[5] M. Crouzeix and A.L. Mignot, editors. *Analyse numérique des équations différentielles.* Collection Mathématiques Appliquées pour la maîtrise. Masson, 1984.

[6] M. Schatzman. *Analyse Numérique.* Enseignement des Mathématiques. Masson/InterEdition, 1991.

[7] C. Guilpin. *Manuel de calcul numérique appliqué.* EDP Sciences, 1999.

[8] P.A. Raviart and J.M. Thomas. *Analyse numérique des équations aux dérivées partielles.* Collection Mathématiques Appliquées pour la maîtrise. Masson, 1983.