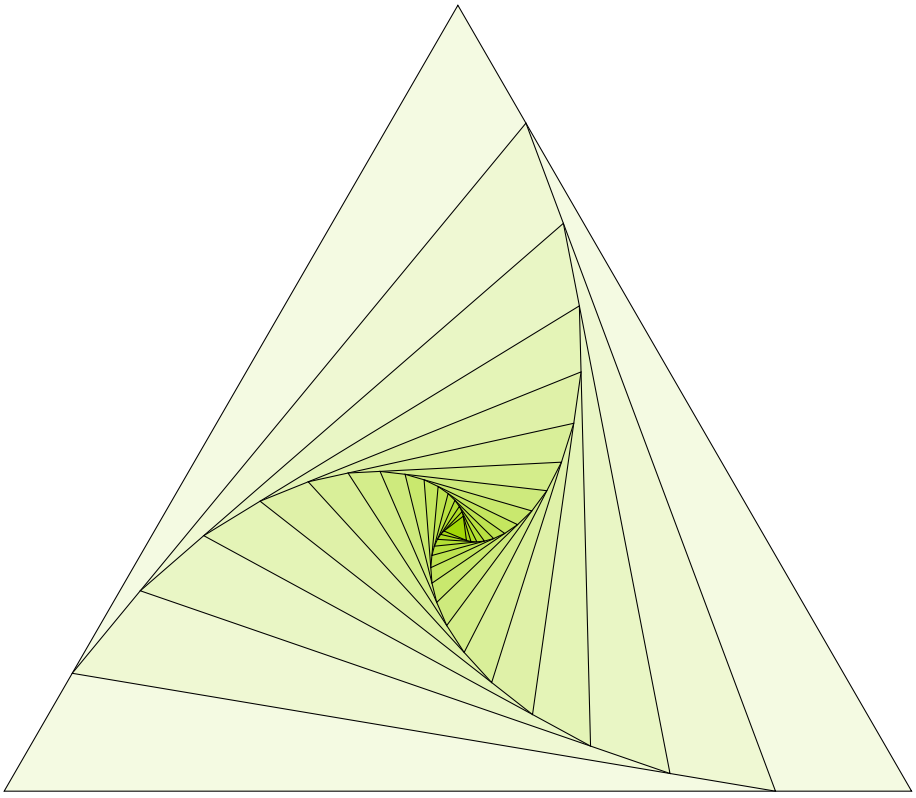
Analyse numérique élémentaire

Notes de cours

Sup Galilée, Ingénieurs MACS 1ère année & L3-MIM

Version du 2022/09/15



Francois Cuvelier
Université Paris XIII / Institut Galilée
L.A.G.A./Département de Mathématiques
http://www.math.univ-paris13.fr/~cuvelier
cuvelier@math.univ-paris13.fr

Table des matières

Chapitre 1

Représentation des nombres en machine, erreurs d'arrondis

 Toute cette partie est le contenu quasi-intégral d'un document réalisé par C. Japhet

Ce chapitre est une introduction à la représentation des nombres en machine et aux erreurs d'arrondis, basé sur [5], [4].

1.1 Un exemple : calcul approché de π

Cet exemple est extrait de [5], [4]. Le nombre π est connu depuis l'antiquité, en tant que méthode de calcul du périmètre du cercle ou de l'aire du disque. Le problème de la quadrature du cercle étudié par les anciens Grecs consiste à construire un carré de même aire qu'un cercle donné à l'aide d'une règle et d'un compas. Ce problème resta insoluble jusqu'au $19^{\text{ème}}$ siècle, où la démonstration de la transcendance de π montra que le problème ne peut être résolu en utilisant une règle et un compas.

Nous savons aujourd'hui que l'aire d'un cercle de rayon r est $\mathcal{A} = \pi r^2$. Parmi les solutions proposées pour approcher $\mathcal{A}$, une méthode consiste à construire un polygône dont le nombre de côté augmenterait jusqu'à ce qu'il devienne équivalent au cercle circonscrit. C'est Archimède vers 250 avant J-C qui appliquera cette propriété au calcul des décimales du nombre π, en utilisant à la fois un polygône inscrit et circonscrit au cercle. Il utilise ainsi un algorithme pour le calcul et parvient à l'approximation de π dans l'intervalle $(3 + \frac{1}{7}, 3 + \frac{10}{71})$ en faisant tendre le nombre de côtés jusqu'à 96.

Regardons l'algorithme de calcul par les polygônes inscrits. On considère un cercle de rayon $r = 1$ et on note $\mathcal{A}_n$ l'aire associée au polygône inscrit à n côtés. En notant $\alpha_n = \frac{2\pi}{n}$, $\mathcal{A}_n$ est égale à n fois l'aire du triangle ABC représenté sur la figure 1.1, c'est-à-dire

$$\mathcal{A}_n = n \cos\frac{\alpha_n}{2} \sin\frac{\alpha_n}{2},$$

que l'on peut réécrire

$$\mathcal{A}_n = \frac{n}{2}(2\cos\frac{\alpha_n}{2}\sin\frac{\alpha_n}{2}) = \frac{n}{2}\sin\alpha_n = \frac{n}{2}\sin(\frac{2\pi}{n}).$$

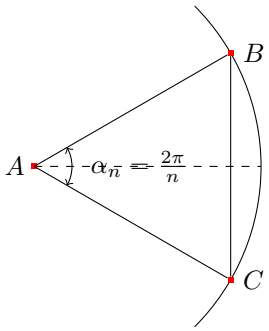


Figure 1.1: Quadrature du cercle

Comme on cherche à calculer π à l'aide de A_n, on ne peut pas utiliser l'expression ci-dessus pour calculer A_n, mais on peut exprimer A_{2n} en fonction de A_n en utilisant la relation

$$\sin\frac{\alpha_n}{2} = \sqrt{\frac{1-\cos\alpha_n}{2}} = \sqrt{\frac{1-\sqrt{1-\sin^2\alpha_n}}{2}}.$$

Ainsi, en prenant $n = 2^k$, on définit l'approximation de π par récurrence

$$x_k = A_{2^k} = \frac{2^k}{2}s_k, \quad \text{avec } s_k = \sin(\frac{2\pi}{2^k}) = \sqrt{\frac{1-\sqrt{1-s_{k-1}}}{2}}$$

En partant de $k = 2$ (i.e. $n = 4$ et $s = 1$) on obtient l'algorithme suivant:

Algorithme 1.1 Algorithme de calcul de π, version naïve

```
1: s ← 1, n ← 4                                  ▷ Initialisations
2: Tantque s > 1e − 10 faire                     ▷ Arrêt si s = sin(α) est petit
3:     s ← sqrt((1 − sqrt(1 − s ∗ s))/2)         ▷ nouvelle valeur de sin(α/2)
4:     n ← 2 ∗ n                                 ▷ nouvelle valeur de n
5:     A ← (n/2) ∗ s                             ▷ nouvelle valeur de l'aire du polygône
6: Fin Tantque
```

On a $\lim_{k\longrightarrow+\infty} x_k = \pi$. Ce n'est pourtant pas du tout ce que l'on va observer sur machine! Les résultats en Python (sous Sage) de la table 1.1 montre que l'algorithme commence par converger vers π puis pour $n > 65536$, l'erreur augmente et finalement on obtient $A_n = 0$!! "Although the theory and the program are correct, we obtain incorrect answers" ([5]).

Ceci résulte du codage des valeurs réelles sur un nombre fini de bits, ce que nous allons détailler dans ce chapitre.

1.2 Représentation scientifique des nombres dans différentes bases

Dans cette section nous introduisons les notions de mantisse, exposant, et la façon dont sont représentés les nombres sur une calculatrice ou un ordinateur.

1.2.1 Partie entière, mantisse et exposant

Exemple en base 10

La base 10 est la base naturelle avec laquelle on travaille et celle que l'on retrouve dans les calculatrices.

Un nombre à virgule, ou nombre décimal, a plusieurs écritures différentes en changeant simplement la position du point décimal et en rajoutant à la fin une puissance de 10 dans l'écriture de ce nombre. La partie à gauche du point décimal est la partie entière, celle à droite avant l'exposant s'appelle la mantisse. Par exemple le nombre $x = 1234.5678$ a plusieurs représentations :

$$x = 1234.5678 = 1234.5678\ 10^0 = \mathbf{1.2345678\ 10^3} = 0.0012345678\ 10^6, \tag{1.1}$$

n	A_n	$\lvert A_n - \pi \rvert$	$\sin(\alpha_n)$
4	2.00000000000000	1.141593e+00	1.000000e+00
8	2.82842712474619	3.131655e-01	7.071068e-01
16	3.06146745892072	8.012519e-02	3.826834e-01
32	3.12144515225805	2.014750e-02	1.950903e-01
64	3.13654849054594	5.044163e-03	9.801714e-02
128	3.14033115695474	1.261497e-03	4.906767e-02
256	3.14127725093276	3.154027e-04	2.454123e-02
512	3.14151380114415	7.885245e-05	1.227154e-02
1024	3.14157294036788	1.971322e-05	6.135885e-03
2048	3.14158772527996	4.928310e-06	3.067957e-03
4096	3.14159142150464	1.232085e-06	1.533980e-03
8192	3.14159234561108	3.079787e-07	7.669903e-04
16384	3.14159257654500	7.704479e-08	3.834952e-04
32768	3.14159263346325	2.012654e-08	1.917476e-04
65536	3.14159265480759	1.217796e-09	9.587380e-05
131072	3.14159264532122	8.268578e-09	4.793690e-05
262144	3.14159260737572	4.621407e-08	2.396845e-05
524288	3.14159291093967	2.573499e-07	1.198423e-05
1048576	3.14159412519519	1.471605e-06	5.992115e-06
2097152	3.14159655370482	3.900115e-06	2.996060e-06
4194304	3.14159655370482	3.900115e-06	1.498030e-06
8388608	3.14167426502176	8.161143e-05	7.490335e-07
16777216	3.14182968188920	2.370283e-04	3.745353e-07
33554432	3.14245127249413	8.586189e-04	1.873047e-07
67108864	3.14245127249413	8.586189e-04	9.365235e-08
134217728	3.16227766016838	2.068501e-02	4.712161e-08
268435456	3.16227766016838	2.068501e-02	2.356080e-08
536870912	3.46410161513775	3.225090e-01	1.290478e-08
1073741824	4.00000000000000	8.584073e-01	7.450581e-09
2147483648	0.000000000000000	3.141593e+00	0.000000e+00

Table 1.1: Calcul de π avec l'algorithme naïf 1.1

avec

- *Partie entière* : 1234
- *Mantisse* : 0.5678 ou 1.2345678 ou 0.0012345678
- *Exposant* : 4 ou 6

Selon le décalage et l'exposant que l'on aura choisi, le couple mantisse-exposant va changer mais le nombre représenté est le même. Afin d'avoir une représentation unique, la troisième dans (1.1) sera utilisée, avec 1.2345678 pour mantisse et 3 pour exposant. Elle correspond à l'écriture avec un premier chiffre avant le point décimal non nul dans la mantisse.

Exemple en base 2

La base 2 est la base utilisée par les ordinateurs. Les chiffres utilisables dans cette base sont 0 et 1 que l'on appelle *bit* pour *binary digit*, les ordinateurs travaillent en binaire. Par exemple

$$39 = 32 + 4 + 2 + 1 = 2^5 + 2^2 + 2^1 + 2^0 = (100111)_2,$$
$$3.625 = 2^1 + 2^0 + 2^{-1} + 2^{-3} = (11.101)_2 = (1.1101)_2\ 2^1$$

Représentation d'un nombre en machine : nombres flottants

De façon générale tout nombre réel x sera représenté dans une base b ($b = 10$ pour une calculatrice $b = 2$ pour un ordinateur) par son signe ($+$ ou $-$), la mantisse m (appelée aussi significande), la base b et un

exposant e tel que le couple (m, e) caractérise le nombre. En faisant varier e, on fait « flotter » la virgule décimale. La limitation fondamentale est que la place mémoire d'un ordinateur est limitée, c'est-à-dire qu'il ne pourra stocker qu'un ensemble fini de nombres. Ainsi un nombre machine réel ou *nombre à virgule flottante* s'écrira :

$$\begin{aligned} \tilde{x} &= \pm m \cdot b^e \\ m &= D.D\cdots D \\ e &= D\cdots D \end{aligned}$$

où $D \in \{0, 1, ..., b-1\}$ représente un chiffre. Des représentations approchées de π sont : (0.031,2), (3.142,0), (0.003,3) et on observe qu'elles ne donnent pas la même précision. Pour rendre la représentation unique et garder la meilleure précision, on utilisera une mantisse *normalisée* : le premier chiffre avant le point décimal dans la mantisse est non nul. Les nombres machine correspondants sont appelés *normalisés*. En base 2, le premier bit dans la mantisse sera donc toujours 1, et on n'écrit pas ce 1 ce qui permet d'économiser un bit. L'exposant est un nombre variant dans un intervalle fini de valeurs admissibles : $L \leqslant e \leqslant U$ (typiquement $L < 0$ et $U > 0$). Le système est donc caractérisé par quatre entiers :

- la base b ($b = 2$),
- le nombre de chiffres t dans la mantisse (en base b),
- l'exposant minimal L et maximal U.

En mathématiques on effectue les calculs avec des nombres réels x provenant de l'intervalle continu $x \in [-\infty, \infty]$. A cause de la limitation ci-dessus, la plupart des réels seront approchés sur un ordinateur. Par exemple, $\frac{1}{3}, \sqrt{2}, \pi$ possèdent une infinité de décimales et ne peuvent donc pas avoir de représentation exacte en machine. Le plus simple des calculs devient alors approché. L'expérience pratique montre que cette quantité limitée de nombres représentables est largement suffisante pour les calculs. Sur l'ordinateur, les nombres utilisés lors des calculs sont des nombres machine $\tilde{x}$ provenant d'un ensemble discret de nombres machine $\tilde{x} \in \{\tilde{x}_{min}, \ldots, \tilde{x}_{max}\}$. Ainsi, chaque nombre réel x doit être transformé en un nombre machine $\tilde{x}$ afin de pouvoir être utilisé sur un ordinateur. Un exemple est donné sur la figure 1.1.

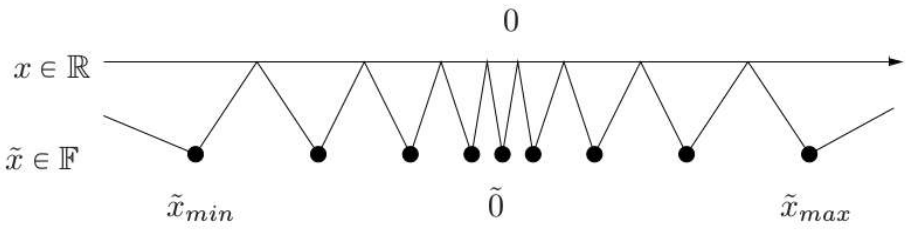


Figure 1.2: Représentation des nombres réels $\mathbb{R}$ par les nombres machine $\mathbb{F}$

Prenons un exemple, beaucoup trop simple pour être utilisé mais pour fixer les idées: $t = 3,\ L = -1,\ U = 2$. Dans ce cas on a 3 chiffres significatifs et 33 nombres dans le système $\mathbb{F}$. Ils se répartissent avec 0 d'une part, 16 nombres négatifs que l'on ne représente pas ici, et 16 nombres positifs représentés sur la figure 1.3.

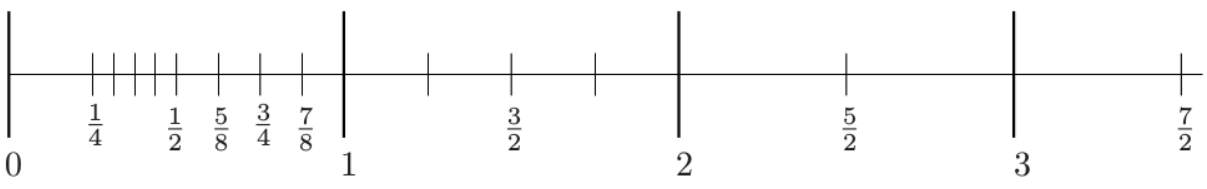


Figure 1.3: Nombres positifs de $\mathbb{F}$ dans le cas $t = 3,\ L = -1,\ U = 2$

Dans cet exemple, l'écriture en binaire des nombres entre $\frac{1}{2}$ et 1 est

$$\frac{1}{2} = (0.100)_2, \quad \frac{3}{4} = \frac{1}{2} + \frac{1}{4} = (0.110)_2,$$
$$\frac{5}{8} = \frac{1}{2} + \frac{1}{8} = (0.101)_2, \quad \frac{7}{8} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = (0.111)_2.$$

On obtient ensuite les autres nombres en multipliant par une puissance de 2. Le plus grand nombre représentable dans ce système est $\frac{7}{2}$ (en particulier 4 n'est pas représentable). On remarque que les nombres ne sont pas espacés régulièrement. Ils sont beaucoup plus resserrés du côté de 0 entre $\frac{1}{4}$ et $\frac{1}{2}$ que entre 1 et 2 et encore plus qu'entre 2 et 3. Plus précisément, chaque fois que l'on passe par une puissance de 2, l'espacement absolu est multiplié par 2, mais l'espacement relatif reste constant ce qui est une bonne chose pour un calcul d'ingéniérie car on a besoin d'une précision absolue beaucoup plus grande pour des nombres petits (autour de un millième par exemple) que des nombres très grands (de l'ordre du million par exemple). Mais la précision ou l'erreur relative sera du même ordre. L'erreur absolue ou relative est définie à section 1.4.1.

Précision machine. elle est décrite par le nombre machine *eps*. *eps* est le plus petit nombre machine positif tel que $1 + eps > 1$ sur la machine. C'est la distance entre l'entier 1 et le nombre machine $\tilde{x} \in \mathbb{F}$ le plus proche, qui lui est supérieur. Dans l'exemple précédent $eps = 1/4$.

Sur une calculatrice

Le système utilisé est la base 10 ($b = 10$). Typiquement, il y a 10 chiffres pour la mantisse et 2 pour l'exposant ($L = -99$ et $U = 99$).

- Le plus grand nombre machine

$$\tilde{x}_{max} = 9.999999999 \times 10^{+99}$$

- Le plus petit nombre machine

$$\tilde{x}_{min} = -9.999999999 \times 10^{+99}$$

- Le plus petit nombre machine strictement positif

$$\tilde{x}_{+} = 1.000000000 \times 10^{-99}$$

  Notez qu'avec des nombres dénormalisés, ce nombre serait $0.000000001 \times 10^{-99}$, c'est-à-dire avec seulement un chiffre significatif!

Les différences de représentation des nombres flottants d'un ordinateur à un autre obligeaient à reprendre les programmes de calcul scientifique pour les porter d'une machine à une autre. Pour assurer la compatibilité entre les machines, depuis 1985 une norme a été proposée par l'IEEE (Institute of Electrical and Electronics Engineers), c'est la norme 754.

1.3 Nombres flottants : le système IEEE 754

Le système IEEE 754 est un standard pour la représentation des nombres à virgule flottante en binaire. Il définit les formats de représentation des nombres à virgule flottante (signe, mantisse, exposant, nombres dénormalisés) et valeurs spéciales (infinis et NaN). Le bit de poids fort est le bit de signe. Cela signifie que si ce bit est à 1, le nombre est négatif, et s'il est à 0, le nombre est positif. Les N_e bits suivants représentent l'exposant décalé, et les N_m bits suivants représentent la mantisse.

L'exposant est décalé de $2^{N_e-1} - 1$ (N_e représente le nombre de bits de l'exposant), afin de le stocker sous forme d'un nombre non signé.

1.3.1 Simple précision

C'est le format 32 bits : 1 bit de signe, $N_e = 8$ bits d'exposant (-126 à 127), 23 bits de mantisse comme sur le tableau 1.2. L'exposant est décalé de $2^{N_e-1} - 1 = 2^7 - 1 = 127$.

$\overbrace{\textit{signe } s}$	$\overbrace{\textit{exposant } e}$	$\overbrace{\textit{mantisse } m}$
$\overbrace{S}$	$\overbrace{EEEEEEEE}$	$\overbrace{FFFFFFFFFFFFFFFFFFFFFFF}$
0	1 à 8	9 à 31

Table 1.2: Représentation en simple précision

$\tilde{x}$	exposant e	mantisse m
$\tilde{x} = 0$ (si $S = 0$) $\tilde{x} = -0$ (si $S = 1$)	$e = 0$	$m = 0$
Nombre *normalisé* $\tilde{x} = (-1)^S \times 2^{e-127} \times 1.m$	$0 < e < 255$	quelconque
Nombre *dénormalisé* $\tilde{x} = (-1)^S \times 2^{e-126} \times 0.m$	$e = 0$	$m \neq 0$
$\tilde{x} =$ `Inf` (si $S = 0$) $\tilde{x} = -$`Inf` (si $S = 1$)	$e = 255$	$m = 0$
$\tilde{x} =$ `NaN` *(Not a Number)*	$e = 255$	$m \neq 0$

Table 1.3: Représentation en simple précision

- Le *plus petit nombre positif normalisé* différent de zéro, et le *plus grand nombre négatif normalisé* différent de zéro sont :
  $\pm 2^{-126} = \pm 1,175494351 \times 10^{-38}$
- Le *plus grand nombre positif fini*, et le *plus petit nombre négatif fini* sont : $\pm(2^{24} - 1) \times 2^{104} = \pm 3,4028235 \times 10^{38}$

$\tilde{x}$	signe	exposant e	mantisse m	valeur
zéro	0	0000 0000	000 0000 0000 0000 0000 0000	0,0
	1	0000 0000	000 0000 0000 0000 0000 0000	-0,0
1	0	0111 1111	000 0000 0000 0000 0000 0000	1,0
Plus grand nombre normalisé	0	1111 1110	111 1111 1111 1111 1111 1111	$3,4 \times 10^{38}$
Plus petit $\tilde{x} \geqslant 0$ normalisé	0	0000 0001	000 0000 0000 0000 0000 0000	2^{-126}
Plus petit $\tilde{x} \geqslant 0$ dénormalisé	0	0000 0000	000 0000 0000 0000 0000 0001	2^{-149}
Infini	0	1111 1111	000 0000 0000 0000 0000 0000	`Inf`
	1	1111 1111	000 0000 0000 0000 0000 0000	`-Inf`
NaN	0	1111 1111	010 0000 0000 0000 0000 0000	`NaN`
2	0	1000 0000	000 0000 0000 0000 0000 0000	2,0
exemple 1	0	1000 0001	101 0000 0000 0000 0000 0000	6,5
exemple 2	1	1000 0001	101 0000 0000 0000 0000 0000	-6,5
exemple 3	1	1000 0101	110 1101 0100 0000 0000 0000	-118,625

Table 1.4: Exemples en simple précision

Détaillons l'exemple 3 : on code le nombre décimal -118,625 en utilisant le système IEEE 754.

1. C'est un nombre négatif, le bit de signe est donc "1",
2. On écrit le nombre (sans le signe) en binaire. Nous obtenons 1110110,101,
3. On décale la virgule vers la gauche, en laissant seulement un 1 sur sa gauche (nombre flottant normalisé): $1110110,101 = 1,110110101 \times 2^6$. La mantisse est la partie à droite de la virgule, remplie de 0 vers la droite pour obtenir 23 bits. Cela donne 11011010100000000000000,
4. L'exposant est égal à 6, et nous devons le convertir en binaire et le décaler. Pour le format 32-bit IEEE 754, le décalage est 127. Donc 6 + 127 = 133. En binaire, cela donne 10000101.

1.3.2 Double précision

C'est le format 64 bits : 1 bit de signe, 11 bits d'exposant (-1022 à 1023), 52 bits de mantisse. L'exposant est décalé de $2^{N_e-1}-1 = 2^{10}-1 = 1023$.

$\overbrace{S}^{signe\ s}$	$\overbrace{EEEEEEEEEEE}^{exposant\ e}$	$\overbrace{FFFFF\ldots FFFFFF}^{mantisse\ m}$
0	1 à 11	12 à 63

Table 1.5: Représentation en double précision

$\tilde{x}$	exposant e	mantisse m
$\tilde{x} = 0$ (si $S = 0$) $\tilde{x} = -0$ (si $S = 1$)	$e = 0$	$m = 0$
Nombre *normalisé* $\tilde{x} = (-1)^S \times 2^{e-1023} \times 1.m$	$0 < e < 2047$	quelconque
Nombre *dénormalisé* $\tilde{x} = (-1)^S \times 2^{e-1022} \times 0.m$	$e = 0$	$m \neq 0$
$\tilde{x} =$ `Inf` (si $S = 0$) $\tilde{x} = -$`Inf` (si $S = 1$)	$e = 2047$	$m = 0$
$\tilde{x} =$ `NaN` *(Not a Number)*	$e = 2047$	$m \neq 0$

Table 1.6: Représentation en double précision

- La précision machine est $eps = 2^{-52}$.
- Le *plus grand nombre positif fini*, et le *plus petit nombre négatif fini* sont : $\tilde{x}^{\pm}_{max} = \pm(2^{1024} - 2^{971}) = \pm 1,7976931348623157 \times 10^{308}$
- **Overflow**: L'intervalle de calcul est $[\tilde{x}^{-}_{max}, \tilde{x}^{+}_{max}]$. Si un calcul produit un nombre x qui n'est pas dans cet intervalle, on dit qu'il y a *overflow*. La valeur de x est alors mise à $\pm$`Inf`. Cela peut arriver au milieu du calcul, même si le résultat final peut être représenté par un nombre machine.
- Le *plus petit nombre positif normalisé* différent de zéro, et le *plus grand nombre négatif normalisé* différent de zéro sont :
  $\tilde{x}^{\pm}_{min} = \pm 2^{-1022} = \pm 2,2250738585072020 \times 10^{-308}$
- Le système IEEE permet les calculs avec des nombres dénormalisés dans l'intervalle $[\tilde{x}^{+}_{min} * eps, \tilde{x}^{+}_{min}]$.
- **Underflow**: Si un calcul produit un nombre positif x qui est plus petit que $\tilde{x}^{+}_{min} * eps$, on dit qu'il y a *underflow*. Néanmoins, le calcul ne s'arrête pas dans ce cas, il continue avec la valeur de x mise à zéro.

1.3.3 Matlab

Sous Matlab, les calculs réels sont en double précision par défaut. La fonction `single` peut être utilisée pour convertir les nombres en simple précision. Pour voir la représentation des nombres réels sous Matlab, on peut les afficher au format hexadécimal avec la commande *format hex*.

Le système hexadécimal est celui en base 16 et utilise 16 symboles : 0 à 9 pour représenter les valeurs de 0 à 9, et A, B, C, D, E, F pour représenter les valeurs de 10 à 15. La lecture s'effectue de droite à gauche. La valeur vaut la somme des chiffres affectés de poids correspondant aux puissances successives du nombre 16. Par exemple, $5EB52_{16}$ vaut $2 * 16^0 + 5 * 16^1 + 11 * 16^2 + 14 * 16^3 + 5 * 16^4 = 387922$. Pour passer du binaire au format hexadecimal, c'est facile en regardant la chaîne binaire en groupe de 4 chiffres, et en représentant chaque groupe en un chiffre hexadécimal. Par exemple,

$$\begin{aligned} 387922 = 01011110101101010010_2 &= 0101\quad 1110\quad 1011\quad 0101\quad 0010_2 \\ &= 5\quad E\quad B\quad 5\quad 2_{16} \\ &= 5EB52_{16} \end{aligned}$$

La conversion de l'hexadécimal au binaire est le processus inverse.

1.4 Calculs sur les nombres flottants

1.4.1 Erreurs d'arrondi

Si $\tilde{x}$ et $\tilde{y}$ sont deux nombres machine, alors $z = \tilde{x} \times \tilde{y}$ ne correspondra pas en général à un nombre machine puisque le produit demande une quantité double de chiffres. Le résultat sera un nombre machine $\tilde{z}$ proche de z.

On définit l'*erreur absolue* entre un nombre réel x et le nombre machine correspondant $\tilde{x}$ par

$$r_a = |x - \tilde{x}|.$$

L'*erreur relative* entre ces nombres (si $x \neq 0$) est définie par

$$r = \frac{|x - \tilde{x}|}{|x|}.$$

Opérations machine: On désigne par *flop* (de l'anglais *floating operation*) une opération élémentaire à virgule flottante (addition, soustraction, multiplication ou division) de l'ordinateur. Sur les calculateurs actuels on peut s'attendre à la précision suivante, obtenue dans les opérations basiques:

$$\tilde{x} \tilde{\oplus} \tilde{y} = (\tilde{x} \oplus \tilde{y})(1 + r)$$

où $|r| < eps$, la précision machine, et $\oplus$ représente l'opération exacte, $\oplus \in \{+, -, *, /\}$ et $\tilde{\oplus}$ représente l'opération de l'ordinateur (*flop*).

1.4.2 Associativité

L'associativité des opérations élémentaires comme par exemple l'addition:

$$(x + y) + z = x + (y + z),$$

n'est plus valide en arithmétique finie. Par exemple, avec 6 chiffres de précision, si on prend les trois nombres

$$x = 1.23456e - 3, \quad y = 1.00000e0, \quad z = -y,$$

on obtient $(x + y) + z = (0.00123 + 1.00000e0) - 1.00000e0 = 1.23000e - 3$ alors que $x + (y + z) = x = 1.23456e - 3$. Il est donc essentiel de considérer l'ordre des opérations et faire attention où l'on met les parenthèses.

1.4.3 Monotonicité

Supposons que l'on a une fonction f strictement croissante sur un intervalle $[a, b]$. Peut-on assurer en arithmétique finie que

$$\tilde{x} < \tilde{y} \Rightarrow f(\tilde{x}) < f(\tilde{y})?$$

En général non. Dans la norme IEEE les *fonctions standard* sont implémentées de façon à respecter la monotonicité (mais pas la stricte monotonicité).

1.4.4 Erreurs d'annulation

Ce sont les erreurs dûes à l'annulation numérique de chiffres significatifs, quand les nombres ne sont représentés qu'avec une quantité finie de chiffres, comme les nombres machine. Il est donc important en pratique d'être attentif aux signes dans les expressions, comme l'exemple suivant le montre :
Exemple : on cherche à évaluer sur l'ordinateur de façon précise, pour de petites valeurs de x, la fonction

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

Pour $|x| < \sqrt{eps}$, on risque d'avoir le nombre $\sqrt{1 - x^2}$ remplacé par 1, par la machine, et donc lors du calcul de $f(x)$, on risque d'effectuer une division par 0. Par exemple, pour $x = \frac{\sqrt{eps}}{2}$, on obtient :

```
» f=@(x) 1./(1-sqrt(1-x.^2));
» f(0.5*sqrt(eps))

ans =

   Inf
```

On ne peut donc pas évaluer précisément $f(x)$ sur l'ordinateur dans ce cas. Maintenant, si on multiplie dans $f(x)$ le numérateur et le dénominateur par $1+\sqrt{1-x^2}$, on obtient

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

Cette fois, on peut évaluer $f(x)$ de façon précise :

```
>> f=@(x) (1+sqrt(1-x.^2))./x.^2;
>> f(0.5*sqrt(eps))

ans =

   3.6029e+16
```

C'est ce qui se passe dans la cas du calcul de π avec l'algorithme naïf 1.1. En utilisant la formule

$$\sin\frac{\alpha_n}{2} = \sqrt{\frac{1-\cos\alpha_n}{2}} = \sqrt{\frac{1-\sqrt{1-\sin^2\alpha_n}}{2}},$$

comme $\sin\alpha_n \to 0$, le numérateur à droite est de la forme

$$1-\sqrt{1-\varepsilon^2}, \quad \text{avec } \varepsilon = \sin\alpha_n \text{ petit },$$

donc sujet aux erreurs d'annulation. Pour y palier, il faut reformuler les équations de façon à s'affranchir des erreurs d'annulations, par exemple en multipliant le numérateur et le dénominateur par $(1+\sqrt{1-\sin^2\alpha_n})$.

$$\begin{aligned}\sin\frac{\alpha_n}{2} &= \sqrt{\frac{1-\sqrt{1-\sin^2\alpha_n}}{2}} = \sqrt{\frac{(1-\sqrt{1-\sin^2\alpha_n})(1+\sqrt{1-\sin^2\alpha_n})}{2(1+\sqrt{1-\sin^2\alpha_n})}} \\ &= \sqrt{\frac{1-(1-\sin^2\alpha_n)}{2(1+\sqrt{1-\sin^2\alpha_n})}} = \frac{\sin\alpha_n}{\sqrt{2(1+\sqrt{1-\sin^2\alpha_n})}}.\end{aligned}$$

On peut alors écrire l'Algorithme 1.2 correspondant au calcul de π avec cette nouvelle formule.

Algorithme 1.2 Algorithme de calcul de π, version stable

```
1: s ← 1, n ← 4,                                  ▷ Initialisations
2: Tantque s > 1e − 10 faire                      ▷ Arrêt si s = sin(α) est petit
3:     s ← s/sqrt(2 ∗ (1 + sqrt(1 − s ∗ s)))      ▷ nouvelle valeur de sin(α/2)
4:     n ← 2 ∗ n                                  ▷ nouvelle valeur de n
5:     A ← (n/2) ∗ s                              ▷ nouvelle valeur de l'aire du polygône
6: Fin Tantque
```

1.5 Quelques catastrophes dûes à l'arithmétique flottante

Il y a un petit nombre "connu" de catastrophes dans la vie réelle qui sont attribuables à une mauvaise gestion de l'arithmétique des ordinateurs (erreurs d'arondis, d'annulation), voir [6]. Dans le premier exemple ci-dessous cela c'est payé en vies humaines.

n	A_n	$\lvert A_n - \pi \rvert$	$\sin(\alpha_n)$
4	2.00000000000000	1.141593e+00	1.000000e+00
8	2.82842712474619	3.131655e-01	7.071068e-01
16	3.06146745892072	8.012519e-02	3.826834e-01
32	3.12144515225805	2.014750e-02	1.950903e-01
64	3.13654849054594	5.044163e-03	9.801714e-02
128	3.14033115695475	1.261497e-03	4.906767e-02
256	3.14127725093277	3.154027e-04	2.454123e-02
512	3.14151380114430	7.885245e-05	1.227154e-02
1024	3.14157294036709	1.971322e-05	6.135885e-03
2048	3.14158772527716	4.928313e-06	3.067957e-03
4096	3.14159142151120	1.232079e-06	1.533980e-03
8192	3.14159234557012	3.080197e-07	7.669903e-04
16384	3.14159257658487	7.700492e-08	3.834952e-04
32768	3.14159263433856	1.925123e-08	1.917476e-04
65536	3.14159264877699	4.812807e-09	9.587380e-05
131072	3.14159265238659	1.203202e-09	4.793690e-05
262144	3.14159265328899	3.008003e-10	2.396845e-05
524288	3.14159265351459	7.519985e-11	1.198422e-05
1048576	3.14159265357099	1.879963e-11	5.992112e-06
2097152	3.14159265358509	4.699352e-12	2.996056e-06
4194304	3.14159265358862	1.174172e-12	1.498028e-06
8388608	3.14159265358950	2.926548e-13	7.490141e-07
16777216	3.14159265358972	7.238654e-14	3.745070e-07
33554432	3.14159265358978	1.731948e-14	1.872535e-07
67108864	3.14159265358979	3.552714e-15	9.362676e-08
134217728	3.14159265358979	0.000000e+00	4.681338e-08
268435456	3.14159265358979	8.881784e-16	2.340669e-08
536870912	3.14159265358979	1.332268e-15	1.170334e-08
1073741824	3.14159265358979	1.332268e-15	5.851672e-09
2147483648	3.14159265358979	1.332268e-15	2.925836e-09
4294967296	3.14159265358979	1.332268e-15	1.462918e-09
8589934592	3.14159265358979	1.332268e-15	7.314590e-10
17179869184	3.14159265358979	1.332268e-15	3.657295e-10
34359738368	3.14159265358979	1.332268e-15	1.828648e-10
68719476736	3.14159265358979	1.332268e-15	9.143238e-11

Table 1.7: Calcul de π avec l'algorithme stable

Missile Patriot

En février 1991, pendant la Guerre du Golfe, une batterie américaine de missiles Patriot, à Dharan (Arabie Saoudite), a échoué dans l'interception d'un missile Scud irakien. Le Scud a frappé un baraquement de l'armée américaine et a tué 28 soldats. La commission d'enquête a conclu à un calcul incorrect du temps de parcours, dû à un problème d'arrondi. Les nombres étaient représentés en virgule fixe sur 24 bits, donc 24 chiffres binaires. Le temps était compté par l'horloge interne du système en 1/10 de seconde. Malheureusement, 1/10 n'a pas d'écriture finie dans le système binaire : 1/10 = 0,1 (dans le système décimal) = 0,0001100110011001100110011... (dans le système binaire). L'ordinateur de bord arrondissait 1/10 à 24 chiffres, d'où une petite erreur dans le décompte du temps pour chaque 1/10 de seconde. Au moment de l'attaque, la batterie de missile Patriot était allumée depuis environ 100 heures, ce qui avait entraîné une accumulation des erreurs d'arrondi de 0,34 s. Pendant ce temps, un missile Scud parcourt environ 500 m, ce qui explique que le Patriot soit passé à côté de sa cible. Ce qu'il aurait fallu faire c'était redémarrer régulièrement le système de guidage du missile.

Explosion d'Ariane 5

Le 4 juin 1996, une fusée Ariane 5, à son premier lancement, a explosé 40 secondes après l'allumage. La fusée et son chargement avaient coûté 500 millions de dollars. La commission d'enquête a rendu son

rapport au bout de deux semaines. Il s'agissait d'une erreur de programmation dans le système inertiel de référence. À un moment donné, un nombre codé en virgule flottante sur 64 bits (qui représentait la vitesse horizontale de la fusée par rapport à la plate-forme de tir) était converti en un entier sur 16 bits. Malheureusement, le nombre en question était plus grand que 32768 (overflow), le plus grand entier que l'on peut coder sur 16 bits, et la conversion a été incorrecte.

Bourse de Vancouver

Un autre exemple où les erreurs de calcul ont conduit à une erreur notable est le cas de l'indice de la Bourse de Vancouver. En 1982, elle a crée un nouvel indice avec une valeur nominale de 1000. Après chaque transaction boursière, cet indice était recalculé et tronqué après le troisième chiffre décimal et, au bout de 22 mois, la valeur obtenue était 524,881, alors que la valeur correcte était 1098.811. Cette différence s'explique par le fait que toutes les erreurs d'arrondi étaient dans le même sens : l'opération de troncature diminuait à chaque fois la valeur de l'indice.

Chapitre 2

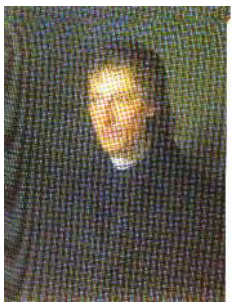
Résolution de systèmes non linéaires

(a) *Scipione del Ferro* 1465-1526, mathématicien italien

(b) *Niccolo Fontana* 1499-1557, mathématicien italien

(c) *Ludovico Ferrari* 1522-1565, mathématicien italien

(d) *Paolo Ruffini* 1765-1822, mathématicien italien



(e) *Bernard Bolzano* 1781-1848, mathématicien, logicien, philosophe et théologien de l'empire d'Autriche

(f) *Niels Henrick Abel* 1802-1829, mathématicien norvégien

Un problème *simple* comme la recherche des zéros/racines d'un polynôme n'est pas ... *simple*. Depuis tout petit, on sait trouver les racines d'un polynôme de degré 2 : $a_2x^2 + a_1x + a_0 = 0$. Quid des racines de polynômes de degré plus élevé?

- **degré 2** : Babylonniens en 1600 avant J.-C.
- **degré 3** : *Scipione del Ferro* (1465-1526, mathématicien italien) et *Niccolo Fontana* (1499-1557, mathématicien italien)
- **degré 4** : *Lodovico Ferrari* (1522-1565, mathématicien italien)
- **degré 5** : *Paolo Ruffini* (1765-1822, mathématicien italien) en 1799, *Niels Henrick Abel* (1802-1829, mathématicien norvégien) en 1824, montrent qu'il n'existe **pas de solution analytique**.

L'objectif de ce chapitre est de rechercher **numériquement** les *zéros/racines* d'une fonction ou d'un systéme d'équations lorsqu'ils existent :

- Soit $f : [a, b] \subset \mathbb{R} \longrightarrow \mathbb{R}$, trouver $x \in [a, b]$ tel que $f(x) = 0$, étudier en section 2.1
- Soit $\boldsymbol{f} : \Omega \subset \mathbb{K}^n \longrightarrow \mathbb{K}^n$, trouver $\boldsymbol{x} \in \Omega$ tel que $\boldsymbol{f}(\boldsymbol{x}) = 0$ ($\mathbb{K} = \mathbb{R}$ ou $\mathbb{K} = \mathbb{C}$). étudier en section 2.3

Avant de commencer une petite piqure de rappels par la lecture de l'Annexe B.1

2.1 Recherche des zéros d'une fonction

Soit $f : [a, b] \subset \mathbb{R} \longrightarrow \mathbb{R}$ une fonction **continue**. On cherche à déterminer les zéros de f ou plus précisément l'ensemble des $x \in [a, b]$ tels que $f(x) = 0$.

Dans un premier temps, on étudiera la **méthode de dichotomie** qui est *assez naturelle*. Puis on étudiera plusieurs algorithmes liés à la méthode du point fixe.

2.1.1 Méthode de dichotomie ou de bissection

Principe et résultats

 **principe de la méthode de dichotomie :** Soit I un intervalle contenant un unique zéro de la fonction f, on le divise par son milieu en deux intervalles et on détermine lequel des deux contient le zéro. On itére ce processus sur le nouvel intervalle.

Plus précisement, on suppose que la fonction f vérifie $f(a)f(b) < 0$. D'après le théorème des valeurs intermédiaires ou le théorème de Bolzano, il existe alors un $\xi \in]a, b[$ tel que $f(x) = 0$.
On note $I^{(0)} =]a, b[$ et $x_0 = (a+b)/2$. Si $f(x_0) = 0$ alors on a fini! Supposons $f(x_0) \neq 0$, alors ξ appartient à $]a, x_0[$ si $f(a)f(x_0) < 0$, sinon il appartient à $]x_0, b[$. On vient donc de déterminer une méthode permettant de diviser par 2 la longueur de l'intervalle de recherche de ξ. On peut bien évidemment itérer ce principe en définissant les trois suites $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$ et $(x_k)_{k\in\mathbb{N}}$ par

- $a_0 = a$, $b_0 = b$ et $x_0 = \frac{a+b}{2}$,
- $\forall k \in \mathbb{N}$,

$$\begin{cases} a_{k+1} = b_{k+1} = x_k & \text{si } f(x_k) = 0, \\ a_{k+1} = x_k,\ b_{k+1} = b_k & \text{si } f(b_k)f(x_k) < 0, \\ a_{k+1} = a_k,\ b_{k+1} = x_k & \text{sinon (i.e. } f(a_k)f(x_k) < 0.) \end{cases}$$

et

$$x_{k+1} = (a_{k+1} + b_{k+1})/2.$$

Par construction on a $a_k \leqslant x_k \leqslant b_k$.

En Figure 2.2, on représente les intervalles successifs obtenus par la méthode de dichotomie pour $a = 0.1$, $b = 2.4$ et $f : x \mapsto (x + 2)(x + 1)(x - 1)$.

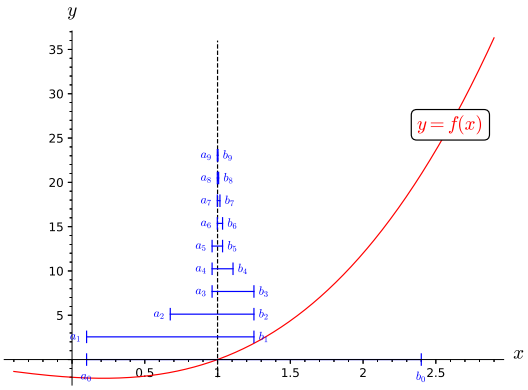


Figure 2.2: Méthode de dichotomie: $f(x) = (x+2)(x+1)(x-1)$,

Exercice 2.1.1

On suppose que la fonction f est continue sur $[a,b]$, vérifie $f(a)f(b) < 0$ et qu'il existe un unique $\alpha \in]a,b[$ tel que $f(\alpha) = 0$.

Q. 1 1. *Montrer que les suites (a_k) et (b_k) convergent vers α.*

2. *En déduire que la suite (x_k) converge vers α.*

Q. 2 1. *Montrer que pour tout $k \in \mathbb{N}$, $|x_k - \alpha| \leqslant \frac{b-a}{2^{k+1}}$.*

2. *Soit $\epsilon > 0$. En déduire que si $k \geqslant \frac{\log(\frac{b-a}{\epsilon})}{\log(2)} - 1$ alors $|x_k - \alpha| \leqslant \epsilon$.*

Correction Exercice 2.1.1

Q. 1 1. Supposons qu'il existe $k \in \mathbb{N}$ tel que $f(x_k) = 0$, (i.e. $x_k = \alpha$ car $x_k \in [a,b]$) alors par construction $a_{k+i} = b_{k+i} = x_{k+i} = \alpha$ pour tout $i \in \mathbb{N}^*$. Ceci assure la convergence des 3 suites vers α.

Supposons maintenant que $\forall k \in \mathbb{N}$, $f(x_k) \neq 0$. Par construction, nous avons $a_k \leqslant a_{k+1} \leqslant b$, $a \leqslant b_{k+1} \leqslant b_k$ et $a_k \leqslant b_k$. La suite (a_k) est convergente car elle est croissante et majorée. La suite (b_k) est décroissante et minorée : elle est donc convergente. De plus $0 \leqslant b_k - a_k \leqslant \frac{b_{k-1}-a_{k-1}}{2}$ et donc $0 \leqslant b_k - a_k \leqslant \frac{b-a}{2^k}$. On en déduit que les suites (a_k) et (b_k) ont même limite. Comme par construction, $\forall k \in \mathbb{N}$, $\alpha \in [a_k, b_k]$ ceci entraine que α est la limite de ces 2 suites.

2. Par construction, $\forall k \in \mathbb{N},$, $a_k \leqslant x_k \leqslant b_k$. D'après le théorème des gendarmes, les suites (a_k) et (b_k) convergeant vers α, on obtient la convergence de la suite (x_k) vers α.

Q. 2 1. On a $\forall k \in \mathbb{N}$, $x_k = \frac{a_k+b_k}{2}$ et $a_k \leqslant \alpha \leqslant b_k$ d'où $|x_k - \alpha| \leqslant \frac{b_k-a_k}{2}$. Ce qui donne

$$|x_k - \alpha| \leqslant \frac{b-a}{2^{k+1}}.$$

2. Pour avoir $|x_k - \alpha| \leqslant \epsilon$, il suffit d'avoir $\frac{b-a}{2^{k+1}} \leqslant \epsilon$, et la fonction log étant croissante on obtient

$$k \geqslant \frac{\log(\frac{b-a}{\epsilon})}{\log(2)} - 1.$$

◇

On peut alors écrire la proposition suivante :

Proposition 2.1: Méthode de dichotomie/bissection

Soit $f : [a,b] \subset \mathbb{R} \longrightarrow \mathbb{R}$ une fonction continue vérifiant $f(a)f(b) < 0$ et admettant $\alpha \in]a,b[$ comme

unique solution de $f(x) = 0$. Alors la suite $(x_k)_{k\in\mathbb{N}}$ définie par la méthode de dichotomie converge vers α et

$$|x_k - \alpha| \leqslant \frac{b-a}{2^{k+1}}, \quad \forall k \in \mathbb{N}.$$

On a alors $\forall \epsilon > 0$, $\forall k \geqslant \frac{\log(\frac{b-a}{\epsilon})}{\log(2)} - 1$

$$|x_k - \alpha| \leqslant \epsilon.$$

A partir de ce résultat, on va proposer plusieurs variantes de l'algorithme de dichotomie.

On note tout d'abord que pour déterminer α il faut calculer la limite d'une suite et donc une infinité de termes! Numériquement et algorithmiquement, on se limite donc à déterminer une approximation de α. La proposition 2.1 permet d'obtenir une approximation α_ϵ de α en un nombre fini d'itérations avec une précision ϵ donnée: on choisi $\alpha_\epsilon = x_{k_{\min}}$ avec $k_{\min} = E(\frac{\log(\frac{b-a}{\epsilon})}{\log(2)})$ où $E(.)$ est la fonction *partie entière*.

Algorithmique

Tout d'abord nous allons poser "correctement" le problème pour lever toute ambiguité. En effet, si le problème est *rechercher une racine de f sur l'intervalle $[a,b]$ par la méthode de dichotomie*, dois-je uniquement rechercher une approximation d'une racine ou calculer la suite (x_k) ou l'ensemble des suites (x_k), (a_k) et (b_k)? C'est ce que nous appelerons le(s) **résultat**(s)/**objectif**(s) de l'algorithme. L'écriture d'un algorithme est fortement corrélée aux objectifs souhaités.
Sauf note contraire, on choisira par la suite comme objectif de déterminer une approximation de la racine α de f :

Résultat : α_ϵ : un réel tel que $|\alpha_\epsilon - \alpha| \leqslant \epsilon$.

Ensuite, pour aboutir à cet objectif on détermine les données (avec hypothèses) nécessaires et suffisantes :

Données :
- a, b : deux réels $a < b$,
- f : $f : [a,b] \subset \mathbb{R} \longrightarrow \mathbb{R}$ vérifiant les hypothèses de la proposition 2.1,
- ϵ : un réel strictement positif.

On peut alors noter (formellement pour le moment) DICHOTOMIE la fonction permettant, à partir des données f, a, b, et ϵ, de trouver α_ϵ. Cette fonction algorithmique aura la syntaxe suivante :

$$\alpha_\epsilon \leftarrow \text{DICHOTOMIE}\ (f, a, b, \epsilon)$$

Nous allons maintenant proposer une manière d'aborder l'écriture de cette fonction dans un langage algorithmique présenté au chapitre A. On doit s'affranchir de tout (ou presque) symbolismes mathématiques pour s'approcher au plus près des langages de programmation. Par exemple, la notation mathématique $(x_k)_{k=0}^{10}$ pour noter les 11 premiers itérés d'une suite n'est pas directement disponible dans les langages de programmations : on peut par exemple passer par un tableau `X` de dimension 11 (au moins) pour stocker les valeurs de la suite. Suivant les langages, l'accès aux composantes d'un tableau diffère : sous Matlab/Octave l'accès au 1er élément d'un tableau `X` se fait par `X(1)` et en C/C++/Python par `X[0]`. Lors de l'écriture d'un algorithme, nous utiliserons l'opérateur `()` ou(exclusif) `[]` pour accéder aux différentes composantes d'un tableau : `()` si le 1er élément est d'indice 1 ou `[]` si le 1er élément est d'indice 0.

Pour écrire un algorithme non trivial, nous utiliserons une *technique de raffinements d'algorithme* (voir chapitre A) : par raffinements succéssifs, nous allons écrire des algorithmes **équivalents** pour aboutir au final à un algorithme n'utilisant que

- des instructions élémentaires (affectation, addition, somme, ...)
- des instructions composées (conditionnelle, boucles pour, tantque, ...)
- des fonctions usuelles, mathématiques (COS , EXP , E partie entière, ...), entrées/sorties , ...

présentent dans tous les langages.
L'idée est donc de partir d'un algorithme formel facile à comprendre puis de le détailler de plus en plus au fur et à mesure des raffinements. Le passage entre deux raffinements successifs doit être simple.

Un petit rappel pour les algorithmes qui suivent : les données sont supposées ... données!

Algorithme 2.1 $\mathcal{R}_0$

1: $k_{\min} \leftarrow \mathrm{E}(\frac{\log(\frac{b-a}{\epsilon})}{\log(2)})$ ▷ E , partie entière
2: Calcul de la suite $(x_k)_{k=0}^{k_{\min}}$ par dichotomie
3: $\alpha_\epsilon \leftarrow x_{k_{\min}}$

Algorithme 2.1 $\mathcal{R}_1$

1: $k_{\min} \leftarrow \mathrm{E}(\frac{\log(\frac{b-a}{\epsilon})}{\log(2)})$ ▷ E , partie entière
2: Initialisation de x_0
3: **Pour** $k \leftarrow 0$ **à** $k_{\min} - 1$ **faire**
4: Calcul de la suite (x_{k+1}) par dichotomie
5: **Fin Pour**
6: $\alpha_\epsilon \leftarrow x_{k_{\min}}$

Entre les raffinements $\mathcal{R}_0$ et $\mathcal{R}_1$, nous avons juste décrit le principe de calcul de toute suite récurrente d'ordre 1.
Pour faciliter la lecture, nous rappelons la méthode de dichotomie :

- $a_0 = a$, $b_0 = b$ et $x_0 = \frac{a+b}{2}$,
- $\forall k \in [\![0, k_{\min} - 1]\!]$,

$$\begin{cases} a_{k+1} = b_{k+1} = x_k & \text{si } f(x_k) = 0, \\ a_{k+1} = x_k,\ b_{k+1} = b_k & \text{si } f(b_k)f(x_k) < 0, \\ a_{k+1} = a_k,\ b_{k+1} = x_k & \text{sinon (i.e. } f(a_k)f(x_k) < 0.) \end{cases}$$

et

$$x_{k+1} = \frac{a_{k+1} + b_{k+1}}{2}$$

Ecrit sous cette forme, le calcul de la suite (x_k) nécessite le calcul simultané des suites (a_k) et (b_k).

Algorithme 2.1 $\mathcal{R}_1$

1: $k_{\min} \leftarrow \mathrm{E}(\frac{\log(\frac{b-a}{\epsilon})}{\log(2)})$ ▷ E , partie entière
2: Initialisation de x_0
3: **Pour** $k \leftarrow 0$ **à** $k_{\min} - 1$ **faire**
4: Calcul de la suite (x_{k+1}) par dichotomie
5: **Fin Pour**
6: $\alpha_\epsilon \leftarrow x_{k_{\min}}$

Algorithme 2.1 $\mathcal{R}_2$

1: $k_{\min} \leftarrow \mathrm{E}(\frac{\log(\frac{b-a}{\epsilon})}{\log(2)})$
2: $a_0 \leftarrow a,\ b_0 \leftarrow b$
3: $x_0 \leftarrow \frac{a_0+b_0}{2}$
4: **Pour** $k \leftarrow 0$ **à** $k_{\min} - 1$ **faire** ▷ Calcul de x_{k+1}
5: **Si** $f(x_k) == 0$ **alors**
6: $a_{k+1} \leftarrow x_k,\ b_{k+1} \leftarrow x_k$
7: **Sinon Si** $f(x_k)f(b_k) < 0$) **alors**
8: $a_{k+1} \leftarrow x_k,\ b_{k+1} \leftarrow b_k$
9: **Sinon**
10: $a_{k+1} \leftarrow a_k,\ b_{k+1} \leftarrow x_k$
11: **Fin Si**
12: $x_{k+1} \leftarrow \frac{a_{k+1}+b_{k+1}}{2}$
13: **Fin Pour**
14: $\alpha_\epsilon \leftarrow x_{k_{\min}}$

La transcription de ce raffinement dans un langage de programmation n'est pas forcément triviale pour tout le monde. Quid de a_k, ϵ, ... qui sont syntaxiquement incorrectes dans un langage de programmation? Le cas du ϵ est très simple : les lettres grecques étant prohibées, on la remplacera par exemple par eps. Pour les suites (a_k), (b_k) et (x_k), tronquées à $k_{\min}$, on pourra utiliser des tableaux (vecteurs) de $k_{\min} + 1$ réels notés $\boldsymbol{A}$, $\boldsymbol{B}$ et $\boldsymbol{X}$ qui contiendront respectivement l'ensemble des valeurs a_k, b_k et x_k pour $0 \leqslant k \leqslant k_{\min}$. Plus précisément nous pourrons utiliser les opérateurs () (indice des tableaux commence à 1) ou [] (indice des tableaux commence à 0) pour accéder aux différents éléments des tableaux et nous aurons par convention $\boldsymbol{A}(k+1) = \boldsymbol{A}[k] = a_k$, $\forall k \in [\![0, k_{\min}]\!]$. Par la suite nous utiliserons les opérateurs () et nous aurons alors les relations suivantes entre les notations *mathématiques* et *algorithmiques*.

$$\forall k \in [\![0, k_{\min}]\!],\ \boldsymbol{A}(k+1) = a_k,\ \boldsymbol{B}(k+1) = b_k \text{ et } \boldsymbol{X}(k+1) = x_k.$$

En utilisant ces changements de notations nous obtenons (enfin) l'Algorithme 2.1.

Algorithme 2.1 Méthode de dichotomie : version 1

Données : $a,\ b$: deux réels $a < b$,
f : $f : [a, b] \subset \mathbb{R} \longrightarrow \mathbb{R}$ vérifiant les hypothèses de la proposition 2.1,
eps : un réel strictement positif.
Résultat : x : un réel tel que $|x - \alpha| \leqslant$ eps.

Fonction $x \leftarrow$ Dichotomie1 ($f, a, b,$ eps)
  kmin $\leftarrow$ E(log((b − a)/eps)/ log(2))
  $\boldsymbol{A},\ \boldsymbol{B},\ \boldsymbol{X} \in \mathbb{R}^{\text{kmin}+1}$ ▷ $\boldsymbol{A}(k+1)$ contiendra a_k, ...
  $\boldsymbol{A}(1) \leftarrow a,\ \boldsymbol{B}(1) \leftarrow b,\ \boldsymbol{X}(1) \leftarrow (a+b)/2$
  Pour $k \leftarrow 1$ **à** kmin **faire**
    Si $f(\boldsymbol{X}(k)) == 0$ **alors**
      $\boldsymbol{A}(k+1) \leftarrow \boldsymbol{X}(k),\ \boldsymbol{B}(k+1) \leftarrow \boldsymbol{X}(k)$
    Sinon Si $f(\boldsymbol{B}(k))f(\boldsymbol{X}(k)) < 0$ **alors**
      $\boldsymbol{A}(k+1) \leftarrow \boldsymbol{X}(k),\ \boldsymbol{B}(k+1) \leftarrow \boldsymbol{B}(k)$
    Sinon
      $\boldsymbol{A}(k+1) \leftarrow \boldsymbol{A}(k),\ \boldsymbol{B}(k+1) \leftarrow \boldsymbol{X}(k)$
    Fin Si
    $\boldsymbol{X}(k+1) \leftarrow (\boldsymbol{A}(k+1) + \boldsymbol{B}(k+1))/2$
  Fin Pour
  $x \leftarrow \boldsymbol{X}(\text{kmin} + 1)$
Fin Fonction

Des codes Matlab/Octave et C correspondant de cette fonction sont données en Annexe B.4.1, respectivement en Listing 18 et 28. Les Listings 8 et 25 correspondent respectivement à un *script* Matlab et un *main* C utilisant cette fonction.

On peut noter qu'une réécriture de la méthode de dichotomie permet d'éviter d'utiliser les deux suites (a_k) et (b_k). En effet, on a

- $A = a$, $B = b$ et $x_0 = \frac{A+B}{2}$,
- $\forall k \in [\![0, k_{\min} - 1]\!]$,

$$\begin{cases} A = B = x_k & \text{si } f(x_k) = 0, \\ A = x_k,\ B \text{ inchangé} & \text{si } f(B)f(x_k) < 0, \\ B = x_k,\ A \text{ inchangé} & \text{sinon (i.e. } f(A)f(x_k) < 0.) \end{cases}$$

et

$$x_{k+1} = \frac{A+B}{2}$$

On utilise, ces formules dans l'Algorithme 2.2.

Algorithme 2.2 Méthode de dichotomie : version 2

Données : $a,\ b$: deux réels $a < b$,
f : $f : [a, b] \subset \mathbb{R} \longrightarrow \mathbb{R}$ vérifiant les hypothèses de la proposition 2.1,
eps : un réel strictement positif.
Résultat : x : un réel tel que $|x - \alpha| \leqslant$ eps.

Fonction $x \leftarrow$ DICHOTOMIE2 ($f, a, b,$ eps)
  kmin $\leftarrow$ E(log((b − a)/eps)/ log(2))
  $\boldsymbol{X} \in \mathbb{R}^{\text{kmin}+1}$ ▷ $\boldsymbol{X}(k+1)$ contiendra x_k, ...
  A $\leftarrow a$, B $\leftarrow b$, $\boldsymbol{X}(1) \leftarrow$ (A + B)/2
  Pour $k \leftarrow 1$ **à** kmin **faire**
    Si $f(\boldsymbol{X}(k)) == 0$ **alors**
      A $\leftarrow \boldsymbol{X}(k)$, B $\leftarrow \boldsymbol{X}(k)$
    Sinon Si $f(\text{B})f(\boldsymbol{X}(k)) < 0$ **alors**
      A $\leftarrow \boldsymbol{X}(k)$ ▷ B inchangé
    Sinon
      B $\leftarrow \boldsymbol{X}(k)$ ▷ A inchangé
    Fin Si
    $\boldsymbol{X}(k+1) \leftarrow$ (A + B)/2
  Fin Pour
  $x \leftarrow \boldsymbol{X}(\text{kmin} + 1)$
Fin Fonction

Si notre objectif n'est que de calculer α_ϵ, on peut, sur le même principe, s'affranchir d'utiliser un tableau pour stocker tous les termes de la suite x_k : seul le dernier nous interesse. On propose dans l'Algorithme 2.3, une version n'utilisant pas de tableaux.

Algorithme 2.3 Méthode de dichotomie : version 3

Données : $a,\ b$: deux réels $a < b$,
f : $f : [a, b] \subset \mathbb{R} \longrightarrow \mathbb{R}$ vérifiant les hypothèses de la proposition 2.1,
eps : un réel strictement positif.
Résultat : x : un réel tel que $|x - \alpha| \leqslant$ eps.

Fonction $x \leftarrow$ DICHOTOMIE3 ($f, a, b,$ eps)
  kmin $\leftarrow$ E(log((b − a)/eps)/ log(2))
  $A,\ B \in \mathbb{R}$
  $A \leftarrow a,\ B \leftarrow b,\ x \leftarrow (a+b)/2$
  Pour $k \leftarrow 1$ **à** kmin **faire**
    Si $f(x) == 0$ **alors**
      $A \leftarrow x,\ B \leftarrow x$
    Sinon Si $f(B)f(x) < 0$ **alors**
      $A \leftarrow x$ ▷ B inchangé
    Sinon
      $B \leftarrow x$ ▷ A inchangé
    Fin Si
    $x \leftarrow (A + B)/2$
  Fin Pour
Fin Fonction

Une autre écriture est possible sans utiliser le nombre $k_{\min}$ et donc en utilisant une boucle **Tantque** (pour les anglophobes!) ou **While** (pour les anglophiles!). Celle-ci est présentée dans l'Algorithme 2.4

Algorithme 2.4 Méthode de dichotomie : version 4

Données : $a,\ b$: deux réels $a < b$,
f : $f \in \mathcal{C}^0([a,b];\mathbb{R})$ et $f(a)f(b) < 0$
eps : un réel strictement positif.
Résultat : x : un réel tel que $|x - \alpha| \leqslant$ eps.

```
1: Fonction x ← Dichotomie4 ( f, a, b, eps )
2:   A, B ∈ ℝ
3:   A ← a, B ← b, x ← (a + b)/2
4:   Tantque |x − A| > eps faire
5:     Si f(x) == 0 alors
6:       A ← x, B ← x
7:     Sinon Si f(B)f(x) < 0 alors
8:       A ← x                                   ▷ B inchangé
9:     Sinon
10:      B ← x                                   ▷ A inchangé
11:    Fin Si
12:    x ← (A + B)/2
13:  Fin Tantque
14: Fin Fonction
```

Que pensez-vous de l'algorithme suivant ?

Algorithme 2.5 Méthode de dichotomie : version 5

Données : $a,\ b$: deux réels $a < b$,
f : $f \in \mathcal{C}^0([a,b];\mathbb{R})$ et $f(a)f(b) < 0$.
Résultat : x : un réel tel que $f(x) = 0$.

```
1: Fonction x ← Dichotomie5 ( f, a, b )
2:   A, B ∈ ℝ
3:   A ← a, B ← b, x ← (a + b)/2, xp ← a
4:   Tantque x ∼= xp faire
5:     Si f(B)f(x) < 0 alors
6:       A ← x                                   ▷ B inchangé
7:     Sinon
8:       B ← x                                   ▷ A inchangé
9:     Fin Si
10:    xp ← x
11:    x ← (A + B)/2
12:  Fin Tantque
13: Fin Fonction
```

Des codes Matlab/Octave et C correspondant à cette fonction sont données en Annexe B.4.1, respectivement en Listing 15 et 16. Les Listings 8 et 17 correspondent respectivement à un *script* Matlab et un *main* C utilisant cette fonction.

2.2 Points fixes d'une fonction (dimension 1)

Soit $\Phi : [a,b] \subset \mathbb{R} \longrightarrow \mathbb{R}$ une fonction donnée. Rechercher un **point fixe** de Φ revient à

> Trouver $\alpha \in [a,b]$ tel que
>
> $$\alpha = \Phi(\alpha).$$

L'algorithme de la **méthode du point fixe** consiste en la construction, si elle existe, de la suite

$$x_{k+1} = \Phi(x_k),\ \forall k \in \mathbb{N} \tag{2.1}$$

avec $x_0 \in [a,b]$ donné.

Avant de formuler le théorème du point fixe, un petit rappel peut être utile:

Definition 2.2

On dit qu'une suite $(x_k)_{k\in\mathbb{N}}$ obtenue par une méthode numérique, **converge vers α avec un ordre** $p \geqslant 1$ si

$$\exists k_0 \in \mathbb{N},\ \exists C > 0 \ \text{ tels que } \ |x_{k+1} - \alpha| \leqslant C|x_k - \alpha|^p,\ \forall k \geqslant k_0. \tag{2.2}$$

où $C < 1$ si $p = 1$.

Voici une version du théorème du point fixe dans $\mathbb{R}$. Pour la résolution des systèmes non linéaires un théorème plus général sera proposé (Théorème 2.13): le théorème du point fixe dans un espace de Banach.

Théorème 2.3: Théorème du point fixe dans $\mathbb{R}$

Soient $[a,b]$ un intervalle non vide de $\mathbb{R}$ et Φ une application continue de $[a,b]$ dans lui-même. Alors, il existe **au moins** un point $\alpha \in [a,b]$ vérifiant $\Phi(\alpha) = \alpha$. Le point α est appelé **point fixe de la fonction** Φ.
De plus, si Φ est contractante (lipschitzienne de rapport $L \in [0,1[$), c'est à dire

$$\exists L < 1 \text{ t.q. } |\Phi(x) - \Phi(y)| \leqslant L|x - y|\ \forall (x,y) \in [a,b]^2, \tag{2.3}$$

alors Φ admet un **unique** point fixe $\alpha \in [a,b]$.
Pour tout $x_0 \in [a,b]$, la suite

$$x_{k+1} = \Phi(x_k),\ \forall k \in \mathbb{N} \tag{2.4}$$

est bien définie et elle converge vers α avec un ordre 1 au moins.
On a les deux estimations suivantes :

$$|x_k - \alpha| \leqslant L^k|x_0 - \alpha|,\ \forall k \geqslant 0, \tag{2.5}$$

$$|x_k - \alpha| \leqslant \frac{L}{1-L}|x_k - x_{k-1}|,\ \forall k \geqslant 0, \tag{2.6}$$

Preuve. **1ère approche** :

- On montre tout d'abord l'existence du point fixe. Pour celà, on note $f(x) = \Phi(x) - x$. f est donc une application continue de $[a,b]$ à valeurs réelles. On a $f(a) = \Phi(a) - a \geqslant 0$ et $f(b) = \Phi(b) - b \leqslant 0$ car $a \leqslant \Phi(x) \leqslant b$, pour tout $x \in [a,b]$. Si $f(a) = 0$ ou $f(b) = 0$, alors l'existence est immédiate. Sinon (i.e. $f(a) \neq 0$ et $f(b) \neq 0$), on a $f(a)f(b) < 0$ et par application directe du Théorème B.1 de Bolzano (ou TVI) on obtient l'existence.

- On montre ensuite l'unicité sous l'hypothèse de contraction (2.3). On suppose qu'il existe α_1 et α_2 dans $[a,b]$ tels que $\Phi(\alpha_1) = \alpha_1$ et $\Phi(\alpha_2) = \alpha_2$. Dans ce cas on a

  $$|\alpha_1 - \alpha_2| = |\Phi(\alpha_1) - \Phi(\alpha_2)| \leqslant L|\alpha_1 - \alpha_2|.$$

  On en déduit

  $$(1-L)|\alpha_1 - \alpha_2| \leqslant 0$$

  Comme $L < 1$ on a $1 - L > 0$ et donc $|\alpha_1 - \alpha_2| \leqslant 0$ ce qui entraine $\alpha_1 = \alpha_2$.

- On a $\Phi([a,b]) \subset [a,b]$ et comme $x_0 \in [a,b]$, la suite $(x_k)_{k\in\mathbb{N}}$ est bien définie. On va démontrer la convergence de la suite x_k vers l'unique point fixe α de Φ. Pour celà, en utilisant la définition de la suite et du point fixe, on a

  $$|x_{k+1} - \alpha| = |\Phi(x_k) - \Phi(\alpha)|$$

  Comme Φ est contractante, $x_k \in [a,b]$ et $\alpha \in [a,b]$ on obtient

  $$|x_{k+1} - \alpha| \leqslant L|x_k - \alpha|$$

et une simple récurrence donne alors $\forall k \in \mathbb{N}$

$$|x_k - \alpha| \leqslant L^k |x_0 - \alpha|$$

ce qui démontre la formule 2.5. En faisant tendre k vers $+\infty$ on obtient la convergence de x_k vers α.
Pour la dernière estimation, on a:

$$\begin{aligned} |x_k - \alpha| &\leqslant L|x_{k-1} - \alpha| && \text{car } \Phi \text{ contractante} \\ &\leqslant L(|x_{k-1} - x_k| + |x_k - \alpha| && \text{inégalité triangulaire} \end{aligned}$$

ce qui donne

$$(1 - L)|x_k - \alpha| \leqslant L|x_{k-1} - x_k|$$

Comme $1 - L > 0$, on en déduit immédiatement l'estimation (2.6).

2ème approche :
L'objectif de cette approche est de proposer une démonstration très proche de celle utilisée pour des espaces plus *complexes* (par exemple $\mathbb{C}$, $\mathbb{R}^n$, ...). Dans la 1ère approche le théorème de Bolzano et des inégalités ont été utilisées pour démontrer l'existence d'un point fixe: ceci n'est plus possible dans un cadre plus générale.
On va donc supposer, dès le départ, que l'application Φ est contractante de $[a, b]$ dans lui-même. Ceci va permettre d'établir que la suite x_k est de Cauchy.

- Comme $x_0 \in [a, b]$ et Φ est une application continue de $[a, b]$ dans lui-même, la suite x_k est bien définie $\forall k \in \mathbb{N}$.
- On démontre ensuite que x_k est une suite de Cauchy. Soit $k > 0$. On a

  $$|x_{k+1} - x_k| = |\Phi(x_k) - \Phi(x_{k-1})| \leqslant L|x_k - x_{k-1}|,$$

  et on obtient par récurrence

  $$|x_{k+1} - x_k| \leqslant L^k |x_1 - x_0|$$

  et

  $$\forall l \geqslant 0,\ |x_{k+l} - x_{k+l-1}| \leqslant L^l |x_k - x_{k-1}|.$$

  Soit $p > 2$. On en déduit par application répétée de l'inégalité triangulaire que

  $$\begin{aligned} |x_{k+p} - x_k| &= |(x_{k+p} - x_{k+p-1}) + (x_{k+p-1} - x_{k+p-2}) + \ldots + (x_{k+1} - x_k)| = |\sum_{l=0}^{p-1} (x_{k+l+1} - x_{k+l})| \\ &\leqslant \sum_{l=0}^{p-1} |x_{k+l+1} - x_{k+l})| \\ &\leqslant \sum_{l=0}^{p-1} L^l |x_{k+1} - x_k| = \frac{1 - L^p}{1 - L} |x_{k+1} - x_k| \quad \textit{(voir somme partielle d'une série géométrique)} \\ &\leqslant \frac{1 - L^p}{1 - L} L^k |x_1 - x_0|. \end{aligned}$$

  Comme $L^k \to 0$ quand $k \to +\infty$, on conclu que (x_k) est une suite de Cauchy.
- La suite (x_k) est une suite de Cauchy dans $\mathbb{R}$ espace complet donc elle converge vers β dans $\mathbb{R}$. De plus pour tout k, x_k appartient à $[a, b]$ fermé borné, donc sa limite β aussi.
- Φ étant contractante sur $[a, b]$, elle est donc continue. On a alors par continuité de Φ

  $$\lim_{k \to +\infty} \Phi(x_k) = \Phi(\beta).$$

  Comme $x_{k+1} = \Phi(x_k)$ on aussi

  $$\lim_{k \to +\infty} \Phi(x_k) = \lim_{k \to +\infty} x_{k+1} = \beta$$

  et donc β est un point fixe de Φ. L'existence d'un point fixe est donc établi.
  La suite de la démonstration est inchangée.

□

Théorème 2.4: Convergence globale de la méthode du point fixe

Soit $\Phi \in \mathcal{C}^1([a,b])$ vérifiant $\Phi([a,b]) \subset [a,b]$ et

$$\exists L < 1 \text{ tel que } \forall x \in [a,b],\ |\Phi'(x)| \leqslant L, \tag{2.7}$$

Soit $x_0 \in [a,b]$ et $(x_k)_{k\in\mathbb{N}}$ la suite définie par $x_{k+1} = \Phi(x_k)$. On a alors

1. la fonction Φ admet un unique point fixe $\alpha \in [a,b]$,
2. $\forall k \in \mathbb{N},\ x_k \in [a,b]$,
3. la suite (x_k) converge vers α avec un ordre 1 au moins.
4. Si $x_0 \neq \alpha$, alors
$$\lim_{k\to+\infty} \frac{x_{k+1}-\alpha}{x_k-\alpha} = \Phi'(\alpha). \tag{2.8}$$

Preuve. Pour démontrer les trois premiers points il suffit de montrer que (2.7) entraine que Φ est contractante sur $[a,b]$ pour pouvoir appliquer le théorème 2.3.

En effet, soit $(x,y) \in [a,b]^2$, $x \neq y$. D'après le théorème B.2 des accroissements finis il existe $\xi \in]\min(x,y), \max(x,y)[$ tel que

$$\frac{\Phi(x)-\Phi(y)}{x-y} = \Phi'(\xi).$$

Ce résultat s'obtient aussi par un développement de Taylor. On obtient alors

$$|\Phi(x)-\Phi(y)| = |x-y||\Phi'(\xi)| \leqslant L|x-y|.$$

L'application Φ est donc contractante sur $[a,b]$ et le théorème 2.3 s'applique.

Pour le dernier point, on utilise la définition de la suite et du point fixe α:

$$x_{k+1}-\alpha = \Phi(x_k)-\Phi(\alpha).$$

On utilise le théorème B.2 des accroissements finis pour obtenir: $\exists \xi_k \in]\min(x_k,\alpha), \max(x_k,\alpha)[$ tel que

$$\frac{\Phi(x_k)-\Phi(\alpha)}{x_k-\alpha} = \Phi'(\xi_k).$$

Quand $k \to +\infty$, on a $x_k \to \alpha$ et donc $\xi_k \to \alpha$. Par continuité de la fonction Φ' on obtient (2.8).

□

Théorème 2.5: Convergence locale de la méthode du point fixe

Soit α un point fixe d'une fonction Φ de classe $\mathcal{C}^1$ au voisinage de α.
Si $|\Phi'(\alpha)| < 1$, alors il existe $\delta > 0$ pour lequel x_k converge vers α pour tout x_0 tel que $|x_0-\alpha| \leqslant \delta$.
De plus, si $x_0 \neq \alpha$, on a

$$\lim_{k\to+\infty} \frac{x_{k+1}-\alpha}{x_k-\alpha} = \Phi'(\alpha). \tag{2.9}$$

Preuve. On va construire un intervalle fermé borné, noté $\mathcal{V}$, pour lequel les hypothèses du Théorème 2.4 sont vérifiées.
Comme Φ' est continue dans un voisinage de α avec $|\Phi'(\alpha)| < 1$, alors il existe $\beta > 0$ tel que

$$\forall x \in]\alpha-\beta, \alpha+\beta[, \quad |\Phi'(x)| < 1.$$

En posant $\delta = \beta/2$ (par ex.), on a en posant $\mathcal{V} = [\alpha-\delta, \alpha+\delta]$,

$$\forall x \in \mathcal{V}, \quad |\Phi'(x)| < 1.$$

Comme l'intervalle $\mathcal{V}$ est un fermé et que l'application $|\Phi'|$ est continue sa borne supérieure est atteinte dans $\mathcal{V}$:

$$\exists \bar{x} \in \mathcal{V} \text{ tel que } |\Phi'(\bar{x})| = \sup_{x\in\mathcal{V}} |\Phi'(x)|.$$

On a donc $L = |\Phi'(\bar{x})| < 1$.

Montrons que $\Phi(\mathcal{V}) \subset \mathcal{V}$. En effet, d'après la formule de Taylor-Lagrange:
$\forall x \in \mathcal{V}$, $\exists \xi \in]\min(x,\alpha), \max(x,\alpha)[\subset \mathcal{V}$ tel que

$$\Phi(x) = \Phi(\alpha) + (x-\alpha)\Phi'(\xi).$$

On en déduit

$$|\Phi(x) - \Phi(\alpha)| = |x-\alpha||\Phi'(\xi)| \leqslant \delta|\Phi'(\xi)| \leqslant \delta$$

Comme $\Phi(\alpha) = \alpha$, on obtient $|\Phi(x) - \alpha| \leqslant \delta$ i.e. $\Phi(x) \in \mathcal{V}$.

On peut donc appliquer le Théorème 2.4 avec $[a,b] = \mathcal{V} = [\alpha - \delta, \alpha + \delta]$) ce qui permet de conclure. □

Exercice 2.2.1

Soit α un point fixe d'une fonction Φ de classe $\mathcal{C}^1$ au voisinage de α et vérifiant $\Phi'(\alpha) = 0$.

Q. 1 *Montrer qu'il existe $\delta > 0$ tel que $\forall x_0 \in]\alpha-\delta, \alpha+\delta[$ la suite définie par $x_{k+1} = \Phi(x_k)$ converge vers α.*

On suppose de plus que Φ' est dérivable sur $]\alpha - \delta, \alpha + \delta[$ et qu'il existe $M \in \mathbb{R}^+$ tel que

$$\forall x \in [\alpha - \delta, \alpha + \delta] |\Phi''(x)| \leqslant M$$

Q. 2 *1. Montrer que*

$$\forall x_0 \in [\alpha - \delta, \alpha + \delta],\ |x_k - \alpha| \leqslant \frac{2}{M}\left(\frac{1}{2}M|x_0 - \alpha|\right)^{2^k}$$

2. Quel est l'ordre de convergence dans ce cas.

Q. 3 *A quelle condition a-t'on*

$$|x_k - \alpha| \leqslant \frac{2}{M}10^{-2^k}.$$

Correction Exercice 2.2.1

Q. 1 C'est juste une application du Théorème 2.5.

Q. 2 1. D'après la formule de Taylor-Lagrange rappelée au Théorème B.3, on a $\exists \eta \in]\min(\alpha,x), \max(\alpha,x)[$ tel que

$$\begin{aligned}\Phi(x) &= \Phi(\alpha) + (x-\alpha)\Phi'(\alpha) + \frac{(x-\alpha)^2}{2!}\Phi''(\eta)\\ &= \alpha + \frac{1}{2}\Phi''(\eta)(x-\alpha)^2.\end{aligned}$$

On en déduit que $|\Phi(x) - \alpha| \leqslant \frac{M}{2}|x-\alpha|^2$ ce qui s'écrit encore $\frac{M}{2}|\Phi(x) - \alpha| \leqslant (\frac{M}{2}|x-\alpha|)^2$. Or si $x_0 \in [\alpha - \delta, \alpha + \delta]$, alors $x_{k-1} \in [\alpha - \delta, \alpha + \delta]$, $\forall k \in \mathbb{N}^*$. On a alors

$$\frac{M}{2}|\Phi(x_{k-1}) - \alpha| = \frac{M}{2}|x_k - \alpha| \leqslant (\frac{M}{2}|x_{k-1} - \alpha|)^2$$

et donc par récurrence

$$\frac{M}{2}|x_k - \alpha| \leqslant (\frac{M}{2}|x_0 - \alpha|)^{2^k}.$$

2. On a

$$|x_k - \alpha| \leqslant \frac{2}{M}(\frac{M}{2}|x_{k-1} - \alpha|)^2$$

c'est à dire

$$\frac{|x_k - \alpha|}{|x_{k-1} - \alpha|^2} \leqslant \frac{M}{2}$$

et donc la méthode est d'ordre 2.

Q. 3 En supposant de plus que $|x_0 - \alpha| \leqslant \frac{1}{5M}$, on obtient immédiatement le résultat.

◇

On va maintenant généraliser le résultat de cet exercice:

Proposition 2.6

Soit $p \in \mathbb{N}^*$, et $\Phi \in \mathcal{C}^{p+1}(\mathcal{V})$ pour un certain voisinage $\mathcal{V}$ de α point fixe de Φ. Si $\Phi^{(i)}(\alpha) = 0$, pour $1 \leqslant i \leqslant p$ et si $\Phi^{(p+1)}(\alpha) \neq 0$, alors la méthode de point fixe associée à la fonction Φ est d'ordre $p+1$ et

$$\lim_{k \to +\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^{p+1}} = \frac{\Phi^{(p+1)}(\alpha)}{(p+1)!}. \tag{2.10}$$

Preuve. Les hypothèses du théorème 2.5 étant vérifiées ($\Phi'(\alpha) = 0$), la suite (x_k) converge vers α. D'après un développement de Taylor-Lagrange (voir le Théorème B.3) de Φ en α, il existe η_k entre x_k et α vérifiant

$$\Phi(x_k) = \sum_{i=0}^{p} \frac{(x_k - \alpha)^i}{i!} \Phi^{(i)}(\alpha) + \frac{(x_k - \alpha)^{p+1}}{(p+1)!} \Phi^{(p+1)}(\eta_k).$$

Comme $\alpha = \Phi(\alpha)$ et $\Phi^{(i)}(\alpha) = 0$, pour $1 \leqslant i \leqslant p$ on obtient

$$\begin{aligned} x_{k+1} &= \Phi(\alpha) + \sum_{i=1}^{p} \frac{(x_k - \alpha)^i}{i!} \Phi^{(i)}(\alpha) + \frac{(x_k - \alpha)^{p+1}}{(p+1)!} \Phi^{(p+1)}(\eta_k) \\ &= \alpha + \frac{(x_k - \alpha)^{p+1}}{(p+1)!} \Phi^{(p+1)}(\eta_k) \end{aligned}$$

De plus (η_k) converge vers α car η_k est entre x_k et α et donc comme $\Phi^{(p+1)}$ est continue au voisinage de α on obtient

$$\lim_{k \to +\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^{p+1}} = \lim_{k \to +\infty} \frac{\Phi^{(p+1)}(\eta_k)}{(p+1)!} = \frac{\Phi^{(p+1)}(\alpha)}{(p+1)!}.$$

□

2.2.1 Points fixes attractifs et répulsifs

Soit $\Phi : [a, b] \longrightarrow [a, b]$ une application de classe $\mathcal{C}^1$ admettant un point fixe $\alpha \in [a, b]$.

Points fixes attractifs

On suppose $|\Phi'(\alpha)| < 1$. D'après le théorème 2.5, il existe $\delta > 0$ tel que

$$\forall x_0 \in [\alpha - \delta, \alpha + \delta], \quad \lim_{k \to +\infty} x_k = \alpha.$$

Dans ce cas on dit que α est un **point fixe attractif** pour Φ.

Points fixes répulsifs

Cas $|\Phi'(\alpha)| > 1$. Par définition de la dérivée en α on a

$$\lim_{\substack{x \to \alpha \\ x \neq \alpha}} \left| \frac{\Phi(x) - \Phi(\alpha)}{x - \alpha} \right| = |\Phi'(\alpha)| > 1.$$

Il existe alors $\delta > 0$ tel que

$$\forall x \in [\alpha - h, \alpha + \delta] \backslash \{\alpha\}, \quad |\Phi(x) - \alpha| > |x - \alpha|$$

et donc la suite (x_k) ne peut converger vers α et ceci même si x_0 est très proche de α. Dans ce cas on dit que α est un **point fixe répulsif** pour Φ.

Toutefois, on *peut rattrapper le coup*. En effet, comme Φ' est non nulle et de signe constant au voisinage de α : il existe $h > 0$ tel que la fonction Φ soit strictement monotone sur $I = [\alpha - h, \alpha + h]$.. D'après le corollaire B.4 Φ est une bijection de I dans l'intervalle fermé $J = \Phi^{-1}(I)$. Comme $\alpha = \Phi(\alpha)$, on a $\alpha \in J$ et $\Phi^{-1}(\alpha) = \alpha$. Le problème de point fixe trouver $x \in I$ tel que $\Phi(x) = x$ revient alors à trouver $x \in J$ tel que $\Phi^{-1}(x) = x$. Or $\Phi'(\alpha) \neq 0$ et $\Phi(\alpha) = \alpha$, la proposition B.5 donne alors $(\Phi^{-1})'(\alpha) = 1/\Phi'(\alpha)$ et donc $|(\Phi^{-1})'(\alpha)| < 1$. Le point α est un **point fixe attractif** pour Φ^{-1}.

Points fixes indéterminés

Dans le cas $|\Phi'(\alpha)| = 1$, on ne peut conclure dans le cas général.

2.2.2 Interprétations graphiques de la méthode du point fixe

Exemple 1 : point fixe de la fonction $x \mapsto x^2$.

On s'interesse ici au point fixe $\alpha = 1$ de la fonction $\Phi : x \mapsto x^2$.

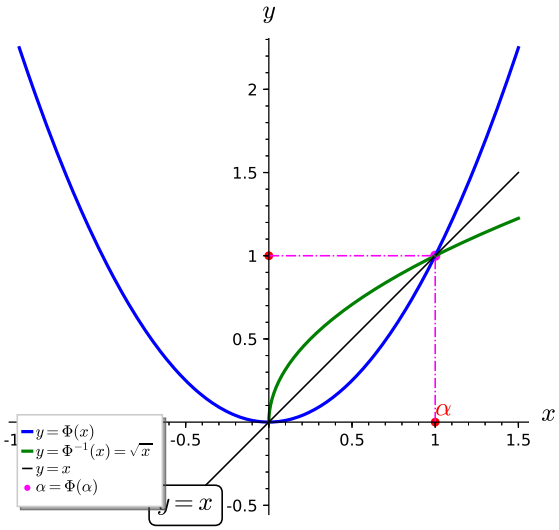


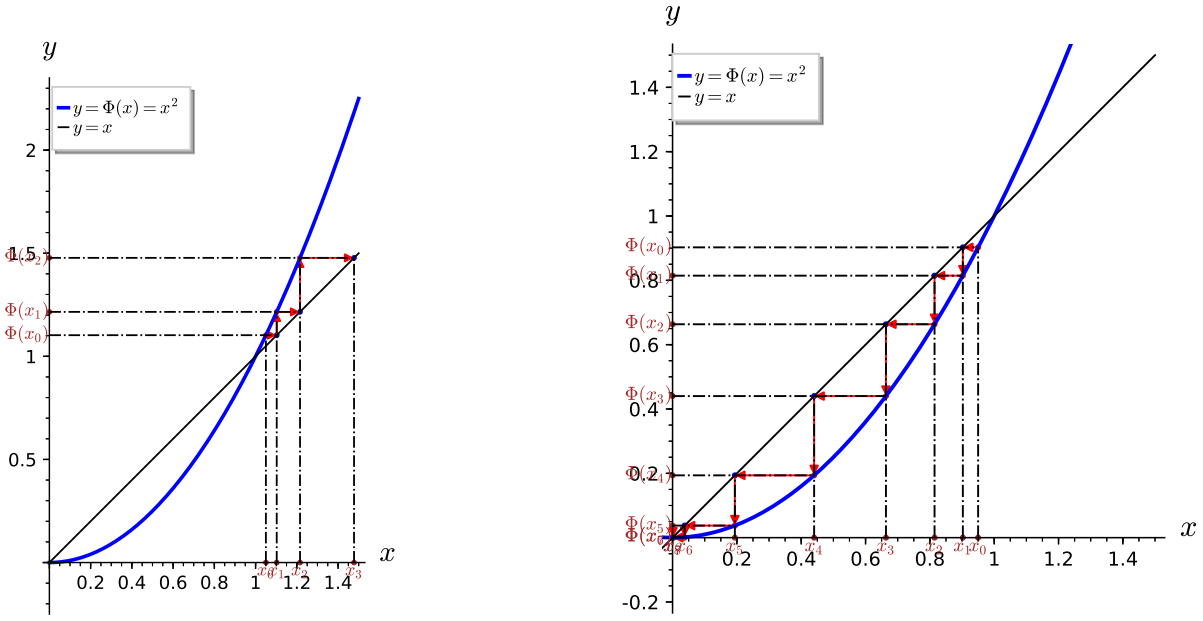
Figure 2.3: fonction x^2 et son fonction inverse $\sqrt{x}$ sur $[0, +\infty[$

Comme $\Phi'(\alpha) = 2$ le point fixe α est répulsif. On donne dans le tableau suivant les premiers termes de deux suites : l'une avec $x_0 = 1.05$ et l'autre avec $x_0 = 0.95$. Dans ce dernier cas, la suite va converger ... vers un autre point fixe.

k	x_k	x_k
0	1.05000	0.950000
1	1.10250	0.902500
2	1.21551	0.814506
3	1.47746	0.663420
⋮	⋮	⋮
8	265742.	1.98264×10^{-6}

Les premières itérations de ces deux suites sont représentées en Figure 2.4.

Sur l'intervalle $[0, +\infty[$, la fonction Φ admet comme fonction inverse $\Phi^{-1}(x) = \sqrt{x}$ et on a aussi $\Phi^{-1}(1) = 1$ (i.e. $\alpha = 1$ est un point fixe de Φ^{-1}). De plus $(\Phi^{-1})'(1) = 1/2 < 1$, ce qui permet d'affirmer que $\alpha = 1$ est un point fixe **attractif** de Φ^{-1}). On donne dans le tableau suivant les premiers termes de



(a) $x_0 = 1.05$ (b) $x_0 = 0.950$

Figure 2.4: $\alpha = 1$, point fixe répulsif de $x \mapsto x^2$

deux suites : l'une avec $x_0 = 1.40$ et l'autre avec $x_0 = 0.200$.

k	x_k	x_k
0	1.40000	0.200000
1	1.18322	0.447214
2	1.08776	0.668740
3	1.04296	0.817765
⋮	⋮	⋮
8	1.00132	0.993733

Les premières itérations de ces deux suites sont représentées en Figure 2.5.

Exemple 2 : point fixe de la fonction $x \mapsto x^2 - x + 1$.

On s'interesse ici au point fixe $\alpha = 1$ de la fonction $f : x \mapsto x^2 - x + 1$ représenté en Figure 2.6. On note que $f'(\alpha) = 1$.

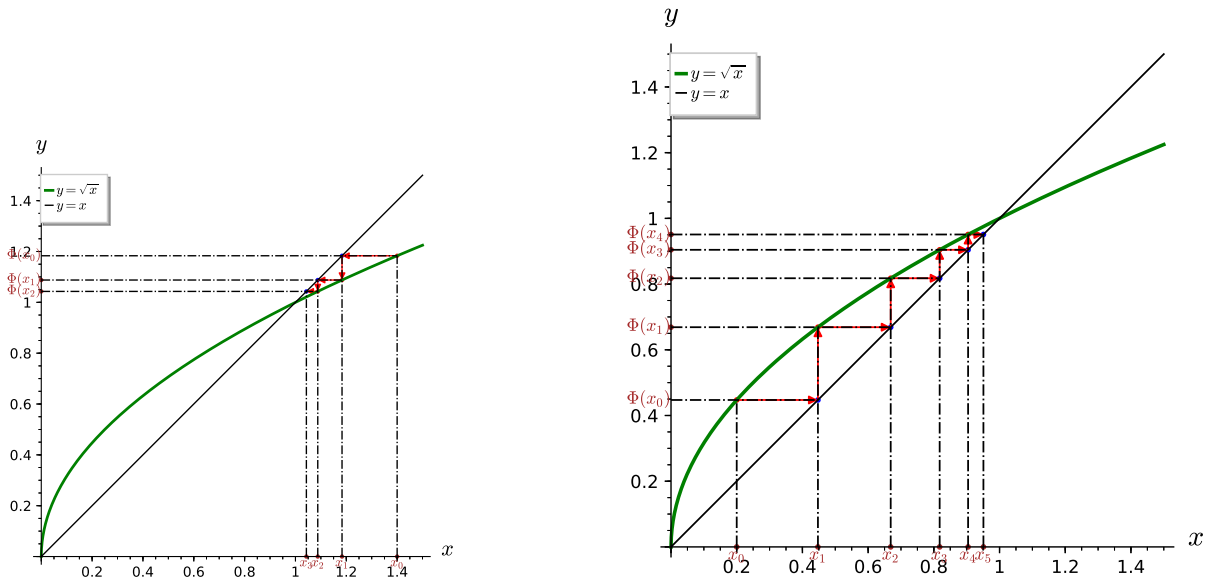
On donne dans le tableau suivant les premiers termes de deux suites : l'une avec $x_0 = 1.10$ diverge et l'autre avec $x_0 = 0.500$ converge vers α.

k	x_k	x_k
0	1.10000	0.500000
1	1.11000	0.750000
2	1.12210	0.812500
3	1.13701	0.847656
⋮	⋮	⋮
8	1.32397	0.918223

Les premières itérations de ces deux suites sont représentées en Figure 2.7.

Exemple 3 : point fixe de fonctions affines particulières

On va étudier les points fixes des trois fonctions affines vérifiant $f_1(1) = f_2(1) = f_3(1) = 1$ et $f_1'(1) = -1$, $f_2'(1) = -3/4$ et $f_3'(1) = -4/3$. Ces trois fonctions sont données par $f_1(x) = -x + 2$, $f_2(x) = -\frac{3}{4}x + \frac{7}{4}$ et $f_3(x) = -\frac{4}{3}x + \frac{7}{3}$.



(a) $x_0 = 1.40$ (b) $x_0 = 0.200$

Figure 2.5: $\alpha = 1$, point fixe attractif de $x \mapsto \sqrt{x}$

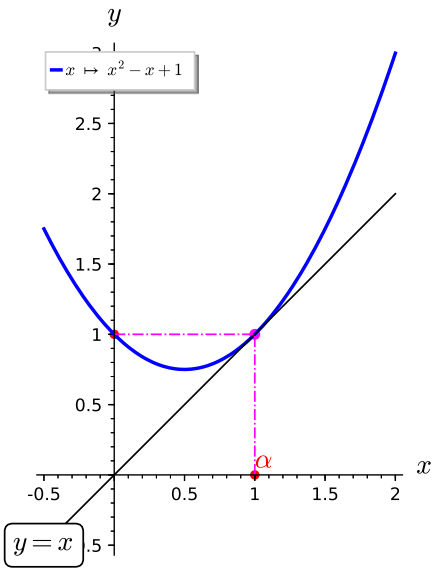
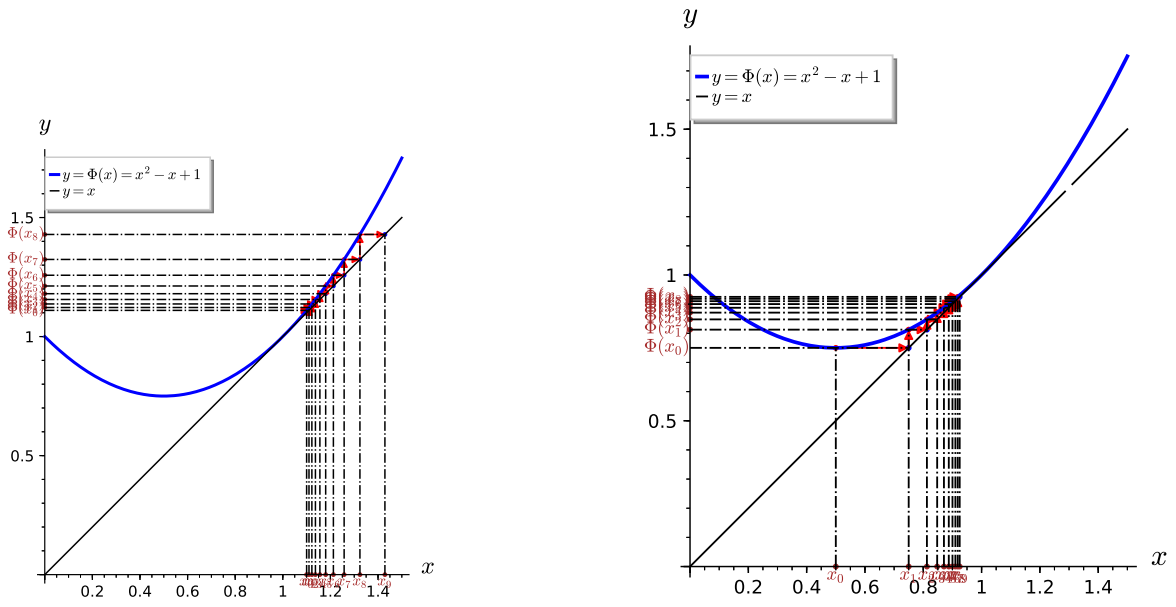


Figure 2.6: fonction $x^2 - x + 1$ et son point fixe $\alpha = 1$.



(a) $x_0 = 1.1$, α point répulsif

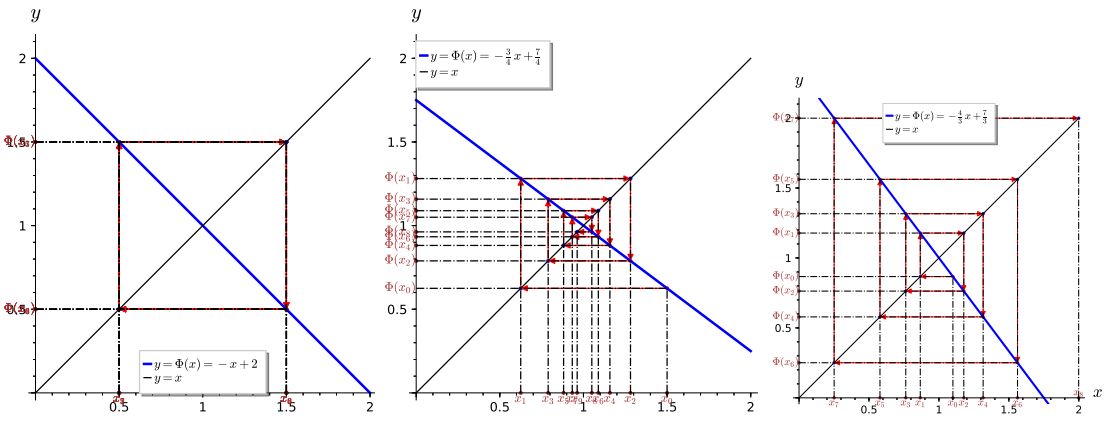
(b) $x_0 = 0.50$, α point attractif

Figure 2.7: $\alpha = 1$, point fixe attractif ou répulsif de $x \mapsto x^2 - x + 1$

Comme $\Phi'(\alpha) = 2$ le point fixe α est répulsif. On donne dans le tableau suivant les premiers termes de trois suites obtenues avec $x_0 = 1.5$ pour les suites associées à f_1 et f_2, et avec $x_0 = 1.1$ pour celle associée à f_2.

k	x_k (f_1)	x_k (f_2)	x_k (f_3)
0	1.50000	1.50000	1.10000
1	0.500000	0.625000	0.866667
2	1.50000	1.28125	1.17778
3	0.500000	0.789062	0.762963
⋮	⋮	⋮	
19	0.500000	0.997886	−22.6503
20	1.50000	1.00159	32.5337

Les premières itérations des trois suites obtenues sont représentées en Figure 2.8.



(a) $x_0 = 1.50$ (b) $x_0 = 1.50$ (c) $x_0 = 1.10$

Figure 2.8: $\alpha = 1$, point fixe de fonctions affines particulières

2.2.3 Algorithme générique du point fixe

Le principe de base est très simple et donné par les Algorithmes 2.6 et 2.7, écrit respectivement avec une boucle **Tantque** et une boucle **Répéter**.

Algorithme 2.6 Méthode de point fixe : version **Tantque** *formel*

$k \leftarrow 0$
Tantque non convergence **faire**
  $x_{k+1} \leftarrow \Phi(x_k)$
  $k \leftarrow k+1$
Fin Tantque
$\alpha_\epsilon \leftarrow x_k$ ▷ le dernier calculé.

Algorithme 2.7 Méthode de point fixe : version **Répéter** *formel*

$k \leftarrow 0$
Répéter
  $k \leftarrow k+1$
  $x_k \leftarrow \Phi(x_{k-1})$
jusqu'à convergence
$\alpha_\epsilon \leftarrow x_k$ ▷ le dernier calculé.

Nous allons plus particulièrement étudier les critères d'arrêt des boucles **Tantque** et **Répéter** (négation l'un de l'autre) qui sont bien trop flous : on s'arrête quand on converge!

Tout d'abord, on n'est pas sur de converger : il faudra donc n'autoriser qu'un nombre maximum d'itérations kmax. De plus, si l'on converge vers une certaine valeur α celle-ci n'étant pas connue à l'avance (sinon l'algo n'a aucun intérêt), on ne peut utiliser, comme condition du **Tantque**, $|x_k - \alpha| > \epsilon$ ($|x_k - \alpha| \leqslant \epsilon$ pour la condition du **Répéter**) où ϵ serait la précision souhaitée. L'idée serait de comparer x_{k+1} et x_k et donc de tester s'ils sont suffisament proches : la condition serait alors $|\Phi(x_k) - x_k| = |x_{k+1} - x_k| >$ tol, (boucle **Tantque**) ou $|\Phi(x_k) - x_k| \leqslant$ tol, (boucle **Répéter**) avec tol la tolérance souhaitée. Il faut noter que dans ce cas la valeur tol doit être choisie correctement et *dépend* de l'ordre de grandeur de α. Si α est de l'ordre de 10^8, tol $= 1$ comme valeur est raisonnable. Toutefois on peut lui préférer une condition *relative* $\frac{|\Phi(x_k)-x_k|}{|x_k|+1} >$ tol (boucle **Tantque**) ou $\frac{|\Phi(x_k)-x_k|}{|x_k|+1} \leqslant$ tol (boucle **Répéter**) qui permet à tol de correspondre à une tolérance *relative* souhaitée que ce soit pour de grandes valeurs de α ou pour des petites. Les algorithmes du point fixe (recherche de α solution de $\Phi(x) = x$) avec notations mathématiques sont donnés par Algorithme 2.8 et 2.9, respectivement avec des boucles **Tantque** et **Répéter**.

Algorithme 2.8 Méthode de point fixe : version **Tantque** *formel* avec critères d'arrêt

$k \leftarrow 0$
err $\leftarrow |\Phi(x_0) - x_0|$ ▷ ou $\frac{|\Phi(x_0)-x_0|}{|x_0|+1}$
Tantque err $>\epsilon$ et $k \leqslant$ kmax **faire**
  $k \leftarrow k+1$
  $x_k \leftarrow \Phi(x_{k-1})$
  err $\leftarrow |\Phi(x_k) - x_k|$ ▷ ou $\frac{|\Phi(x_k)-x_k|}{|x_k|+1}$
Fin Tantque
Si err $\leqslant$ tol **alors** ▷ Convergence
  $\alpha_{\text{tol}} \leftarrow x_k$ ▷ $|\Phi(\alpha_{\text{tol}}) - \alpha_{\text{tol}}| \leqslant$ tol
Fin Si

Algorithme 2.9 Méthode de point fixe : version **Répéter** *formel* avec critères d'arrêt

$k \leftarrow 0$
Répéter
  err $\leftarrow |\Phi(x_k) - x_k|$ ▷ ou $\frac{|\Phi(x_k)-x_k|}{|x_k|+1}$
  $x_{k+1} \leftarrow \Phi(x_k)$
  $k \leftarrow k+1$
jusqu'à err $\leqslant$ tol ou $k >$ kmax
Si err $\leqslant$ tol **alors** ▷ Convergence
  $\alpha_{\text{tol}} \leftarrow x_k$ ▷ $|\Phi(\alpha_{\text{tol}}) - \alpha_{\text{tol}}| \leqslant$ tol
Fin Si

Des versions, sous forme de fonction, plus proches de la programmation sont proposées en Algorithme 2.10 et 2.11. Bien évidemment, suivant l'usage que l'on souhaite de ces fonctions, il est facile de les adapter pour retourner plus d'informations (nombre d'itération effectives, statut de convergence, ...)

Algorithme 2.10 Méthode de point fixe : version **Tantque** avec critères d'arrêt

Données :

Φ : $\Phi : \mathbb{R} \longrightarrow \mathbb{R}$,
x_0 : donnée initiale, $x_0 \in \mathbb{R}$,
tol : la tolérence, tol $\in \mathbb{R}^+$,
kmax : nombre maximum d'itérations, kmax $\in \mathbb{N}^*$

Résultat :

α_{tol} : un réel tel que $|\Phi(\alpha_{\text{tol}}) - \alpha_{\text{tol}}| \leqslant$ tol (ou $\frac{|\Phi(\alpha_{\text{tol}})-\alpha_{\text{tol}}|}{|\alpha_{\text{tol}}|+1} \leqslant$ tol)

1: **Fonction** $\alpha_{\text{tol}} \leftarrow$ PtFixe ($\Phi, x_0, \text{tol}, \text{kmax}$)
2: $k \leftarrow 0$, $\alpha_{\text{tol}} \leftarrow \varnothing$
3: x $\leftarrow x_0$, fx $\leftarrow \Phi(x_0)$,
4: err $\leftarrow |\text{fx} - \text{x}|$ ▷ ou $\frac{|\text{fx}-\text{x}|}{|\text{x}|+1}$
5: **Tantque** err $>$ tol et $k \leqslant$ kmax **faire**
6: $k \leftarrow k + 1$
7: x $\leftarrow$ fx
8: fx $\leftarrow \Phi$(x)
9: err $\leftarrow |\text{fx} - \text{x}|$ ▷ ou $\frac{|\text{fx}-\text{x}|}{|\text{x}|+1}$
10: **Fin Tantque**
11: **Si** err $\leqslant$ tol **alors** ▷ Convergence
12: $\alpha_{\text{tol}} \leftarrow$ x
13: **Fin Si**
14: **Fin Fonction**

Algorithme 2.11 Méthode de point fixe : version **Répéter** avec critères d'arrêt

Données :

Φ : $\Phi : \mathbb{R} \longrightarrow \mathbb{R}$,
x_0 : donnée initiale, $x_0 \in \mathbb{R}$,
tol : la tolérence, tol $\in \mathbb{R}^+$,
kmax : nombre maximum d'itérations, kmax $\in \mathbb{N}^*$

Résultat :

α_{tol} : un réel tel que $|\Phi(\alpha_{\text{tol}}) - \alpha_{\text{tol}}| \leqslant$ tol (ou $\frac{|\Phi(\alpha_{\text{tol}})-\alpha_{\text{tol}}|}{|\alpha_{\text{tol}}|+1} \leqslant$ tol)

1: **Fonction** $\alpha_{\text{tol}} \leftarrow$ PtFixe ($\Phi, x_0, \text{tol}, \text{kmax}$)
2: $k \leftarrow 0$, $\alpha_{\text{tol}} \leftarrow \varnothing$
3: x $\leftarrow x_0$
4: **Répéter**
5: xp $\leftarrow$ x
6: x $\leftarrow \Phi$(xp)
7: err $\leftarrow |\text{x} - \text{xp}|$ ▷ ou $\frac{|\text{x}-\text{xp}|}{|\text{xp}|+1}$
8: $k \leftarrow k + 1$
9: **jusqu'à** err $\leqslant$ tol ou $k >$ kmax
10: **Si** err $\leqslant$ tol **alors** ▷ Convergence
11: $\alpha_{\text{tol}} \leftarrow$ x
12: **Fin Si**
13: **Fin Fonction**

2.2.4 Méthodes de points fixes pour la recherche de racines

On peut noter que résoudre $f(x) = 0$ revient, par exemple, à résoudre $\Phi(x) := x + f(x) = x$. De manière plus générale, si $\mathcal{F}$ est une fonction continue vérifiant $\mathcal{F}(0) = 0$ alors $\Phi(x) = x + \mathcal{F}(f(x))$ est un autre choix possible.

Soit f une fonction de classe $\mathcal{C}^1$ dans un voisinage d'une de ses racines simple α.

Selectionner une version :

Version 1 En appliquant la formule de Taylor pour tout x dans ce voisinage, il existe $\xi \in]\min(x, \alpha), \max(x, \alpha)[$ tel que

$$f(\alpha) = 0 = f(x) + (\alpha - x)f'(\xi).$$

Ceci conduit à la méthode itérative suivante : x_0 étant donné dans le voisinage de α, on doit, $\forall k \in N$, déterminer x_{k+1} vérifiant $f(x_k) + (x_{k+1} - x_k)q_k = 0$ sachant que q_k est une approximation de $f'(\xi)$.

Version 2 On cherche à déterminer une méthode itérative permettant de calculer x_{k+1} en fonction des valeurs précédentes en espérant que $|x_{k+1} - \alpha| \leqslant |x_k - \alpha|$. Pour celà, on écrit une formule de taylor en x_k supposé proche de α et on note $h = \alpha - x_k$

$$f(\alpha) = 0 = f(x_k) + hf'(x_k) + o(h)$$

L'idéal serait de trouver la valeur exacte de h et de prendre $x_{k+1} = x_k + h$ mais ceci n'est pas possible dans un cadre général. C'est pourquoi on cherche une approximation $\tilde{h}$ de h. De plus, on ne connait pas forcément explicitement la dérivée de f. On note donc q_k une approximation de $f'(x_k)$. On choisi alors $\tilde{h}$ comme solution de

$$f(x_k) + \tilde{h}q_k = 0$$

Si $q_k \neq 0$, on obtient la suite itérative

$$x_{k+1} = x_k - \frac{f(x_k)}{q_k}, \ \forall k \in \mathbb{N} \tag{2.11}$$

La valeur x_{k+1} est de fait l'intersection de la droite passant par le point $((x_k), f(x_k))$ et de pente q_k avec l'axe des x.

Par la suite, on va écrire un algorithme du point fixe et étudier différentes méthodes :

- la **méthode de la corde** :
$$q_k = q = \frac{f(b) - f(a)}{b - a}$$
- la **méthode de la sécante** :
$$q_k = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$$
où x_{-1} et x_0 sont données,
- la **méthode de Newton** : en supposant f' connu, on prend
$$q_k = f'(x_k).$$

La méthode de la corde

Soit f une fonction de classe $\mathcal{C}^1$ sur $[a, b]$ tel que $f(a) \neq f(b)$. Soit $x_0 \in [a, b]$ donné. La suite obtenue par la méthode de la corde est donnée par

$$x_{k+1} = x_k - \frac{b - a}{f(b) - f(a)} f(x_k), \ \forall k \in \mathbb{N}. \tag{2.12}$$

On peut voir cette relation comme un cas particulier de la méthode du point fixe. En effet, en prenant $\Phi(x) = x - \frac{b-a}{f(b)-f(a)} f(x)$, la suite définie en (2.12) s'écrit $x_{k+1} = \Phi(x_k)$.

On a alors le résultat suivant

Proposition 2.7: convergence de la méthode de la corde

Soit $f \in \mathcal{C}^1([a, b])$ tel que $f(b) \neq f(a)$ et $\lambda = \frac{f(b)-f(a)}{b-a}$. On note $(x_k)_{k\in\mathbb{N}}$ la suite définie par $x_0 \in [a, b]$ et pour tout $k \geqslant 0$

$$x_{k+1} = x_k - \frac{f(x_k)}{\lambda}. \tag{2.13}$$

On suppose de plus que $\forall x \in [a, b]$

$$\min(\lambda(x - a), \lambda(x - b)) \leqslant f(x) \leqslant \max(\lambda(x - a), \lambda(x - b)) \tag{2.14}$$
$$\min(0, 2\lambda) < f'(x) < \max(0, 2\lambda) \tag{2.15}$$

alors la suite (x_k) converge vers l'unique racine $\alpha \in [a, b]$ de f.

Preuve. voir correction Exercice 2.2.2 □

Exercice 2.2.2

Soit f une fonction de classe $\mathcal{C}^1$ sur $[a, b]$ vérifiant $f(a)f(b) < 0$. et $\lambda = \frac{f(b)-f(a)}{b-a}$. Soit $x_0 \in [a, b]$ donné. La suite obtenue par la méthode de la corde est donnée par

$$x_{k+1} = x_k - \frac{f(x_k)}{\lambda}, \ \forall k \in \mathbb{N}.$$

On note $\Phi(x) = x - \frac{f(x)}{\lambda}$.

Q. 1 *Montrer que si pour tout $x \in [a, b]$ on a*

$$\min(\lambda(x - a), \lambda(x - b)) \leqslant f(x) \leqslant \max(\lambda(x - a), \lambda(x - b)) \tag{2.16}$$

alors $\Phi([a, b]) \subset [a, b]$.

Q. 2 *Montrer que si pour tout $x \in [a, b]$ on a*

$$\min(0, 2\lambda) < f'(x) < \max(0, 2\lambda) \tag{2.17}$$

alors $|\Phi'(x)| < 1$.

Q. 3 *En déduire que sous les deux conditions précédentes la méthode de la corde converge vers l'unique solution $\alpha \in [a,b]$ de $f(x)=0$.*

Correction Exercice 2.2.2

Q. 1 Si $\lambda > 0$, l'inéquation (2.16) devient

$$\begin{aligned}\lambda(x-b) \leqslant f(x) \leqslant \lambda(x-a) &\Leftrightarrow a \leqslant x - \frac{f(x)}{\lambda} \leqslant b \\ &\Leftrightarrow a \leqslant \Phi(x) \leqslant b.\end{aligned}$$

Si $\lambda < 0$, l'inéquation (2.16) devient

$$\begin{aligned}\lambda(x-a) \leqslant f(x) \leqslant \lambda(x-b) &\Leftrightarrow a \leqslant x - \frac{f(x)}{\lambda} \leqslant b \\ &\Leftrightarrow a \leqslant \Phi(x) \leqslant b.\end{aligned}$$

Q. 2 Si $\lambda > 0$, l'inéquation (2.17) devient

$$\begin{aligned}0 < f'(x) < 2\lambda &\Leftrightarrow 0 < \frac{f'(x)}{\lambda} < 2 \\ &\Leftrightarrow -1 < 1 - \frac{f'(x)}{\lambda} < 1 \\ &\Leftrightarrow -1 < \Phi'(x) < 1.\end{aligned}$$

Si $\lambda < 0$, l'inéquation (2.17) devient

$$\begin{aligned}2\lambda < f'(x) < 0 &\Leftrightarrow 0 < \frac{f'(x)}{\lambda} < 2 \\ &\Leftrightarrow -1 < 1 - \frac{f'(x)}{\lambda} < 1 \\ &\Leftrightarrow -1 < \Phi'(x) < 1.\end{aligned}$$

Q. 3 Sous les hypothèses (2.16) et (2.17) on a $\Phi([a,b]) \subset [a,b]$ et $\forall x \in [a,b]$, $|\Phi'(x)| < 1$. Comme f est de classe $\mathcal{C}^1$ sur $[a,b]$, la fonction Φ l'est aussi. La suite (x_k) est définie par $x_{k+1} = \Phi(x_k)$. Ainsi les hypothèses du théorème 2.4 sont vérifiées ce qui assure l'unicité du point fixe ainsi que la convergence de la suite (x_k) vers ce point fixe.

◇

Proposition 2.8: ordre de convergence de la méthode de la corde

Soit $f \in \mathcal{C}^1([a,b])$ tel que $f(b) \neq f(a)$. Si la suite (x_k) définie par la méthode de la corde en (2.13) converge vers $\alpha \in]a,b[$ alors la convergence est au moins d'ordre 1.
De plus, si f est de classe $\mathcal{C}^2$ sur un certain voisinage $\mathcal{V}$ de α et si $f'(\alpha) = \frac{f(b)-f(a)}{b-a}$ alors la convergence est au moins d'ordre 2.

Preuve. • **Order 1 :** On note $\lambda = \frac{f(b)-f(a)}{b-a}$. On a par définition $x_{k+1} = x_k - \frac{f(x_k)}{\lambda}$ (ce qui suppose que $f(x_k)$ soit bien définie, i.e. $x_k \in [a,b]$). Comme $\lambda \neq 0$ et f continue, l'hypothèse (x_k) converge vers α entraine que $f(\alpha) = 0$.
Pour définir l'ordre de convergence, on suppose de plus que $\forall k \in \mathbb{N}$, $x_k \neq \alpha$. On peut alors appliquer la formule de Taylor-Lagrange : il existe ξ_k compris entre x_k et α tel que

$$f(x_k) = \underbrace{f(\alpha)}_{=0} + (x_k - \alpha) f'(\xi_k) = (x_k - \alpha) f'(\xi_k).$$

On a alors en utilisant cette expression dans la définition de la suite x_k

$$x_{k+1} = x_k - (x_k - \alpha)\frac{f'(\xi_k)}{\lambda}.$$

En soustrayant α à cette équation on obtient

$$x_{k+1} - \alpha = x_k - \alpha - (x_k - \alpha)\frac{f'(\xi_k)}{\lambda} = (x_k - \alpha)(1 - \frac{f'(\xi_k)}{\lambda}).$$

Comme $x_k \neq \alpha$, on a alors

$$\frac{x_{k+1} - \alpha}{x_k - \alpha} = 1 - \frac{f'(\xi_k)}{\lambda}.$$

Or x_k converge vers α et ξ_k compris entre x_k et α, ce qui entraine que ξ_k converge vers α. La fonction f' étant continue, on en déduit que $f'(\xi_k)$ converge vers $f'(\alpha)$. Ceci donne donc

$$\lim_{k\to+\infty} \frac{x_{k+1} - \alpha}{x_k - \alpha} = 1 - \frac{f'(\alpha)}{\lambda}.$$

La convergence est donc (au moins) d'ordre 1.

- **Order 2 :** La suite étant convergente, il existe $k_0 \in N$ tel que $\forall k \geqslant k_0$, $x_k \in \mathcal{V}$. Soit $k \geqslant k_0$, comme $f \in \mathcal{C}^2(\mathcal{V})$, on peut appliquer la formule de Taylor-Lagrange : il existe $\eta_k \in \mathcal{V}$ compris entre x_k et α que

$$f(x_k) = \underbrace{f(\alpha)}_{=0} + (x_k - \alpha)f'(\alpha) + \frac{(x_k - \alpha)^2}{2!}f^{(2)}(\eta_k).$$

On a alors en utilisant cette expression dans la définition de la suite x_k

$$x_{k+1} = x_k - \frac{1}{\lambda}\left((x_k - \alpha)f'(\alpha) + \frac{(x_k - \alpha)^2}{2!}f^{(2)}(\eta_k)\right).$$

En soustrayant α à cette équation on obtient

$$x_{k+1} - \alpha = (x_k - \alpha)(\underbrace{1 - \frac{f'(\alpha)}{\lambda}}_{=0 \text{ par hyp.}}) + \frac{1}{\lambda}\frac{(x_k - \alpha)^2}{2!}f^{(2)}(\eta_k).$$

Comme $\eta_k \in \mathcal{V}$ converge vers α (car compris entre x_k et α) et $f^{(2)}$ continue sur $\mathcal{V}$, on en déduit

$$\lim_{k\to+\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \frac{f^{(2)}(\alpha)}{2\lambda}.$$

La convergence est donc (au moins) d'ordre 2.

□

On présente en Algorithme 2.12 l'implémentation usuelle de la méthode de la corde. Une autre version utilisant la fonction PtFixe est donnée en Algorithme 2.13.

Algorithme 2.12 Méthode de la corde

Données :

f	:	$f : \mathbb{R} \longrightarrow \mathbb{R}$,
$a,\ b$	:	deux réels tels que $f(a) \neq f(b)$,
x_0	:	donnée initiale, $x_0 \in \mathbb{R}$,
tol	:	la tolérence, tol $\in \mathbb{R}^+$,
kmax	:	nombre maximum d'itérations, kmax $\in \mathbb{N}^*$

Résultat :

α_{tol}	:	un réel tel que $\lvert f(\alpha_{\text{tol}})\rvert \leqslant$ tol

1: **Fonction** $\alpha_{\text{tol}} \leftarrow$ CORDE ($f, a, b, x_0, \text{tol}, \text{kmax}$)
2: $k \leftarrow 0$, $\alpha_{\text{tol}} \leftarrow \varnothing$
3: $q \leftarrow \frac{b-a}{f(b)-f(a)}$
4: x $\leftarrow x_0$,
5: err $\leftarrow$ tol $+ 1$
6: **Tantque** err $>$ tol et $k \leqslant$ kmax **faire**
7: $k \leftarrow k + 1$
8: xp $\leftarrow$ x
9: x $\leftarrow$ xp $- q * f$(xp)
10: err $\leftarrow$ |x $-$ xp|
11: **Fin Tantque**
12: **Si** err $\leqslant$ tol **alors** ▷ Convergence
13: $\alpha_{\text{tol}} \leftarrow$ x
14: **Fin Si**
15: **Fin Fonction**

Algorithme 2.13 Méthode de la corde utilisant la fonction PTFIXE

Données :

f	:	$f : \mathbb{R} \longrightarrow \mathbb{R}$,
$a,\ b$	:	deux réels tels que $f(a) \neq f(b)$,
x_0	:	donnée initiale, $x_0 \in \mathbb{R}$,
tol	:	la tolérence, tol $\in \mathbb{R}^+$,
kmax	:	nombre maximum d'itérations, kmax $\in \mathbb{N}^*$

Résultat :

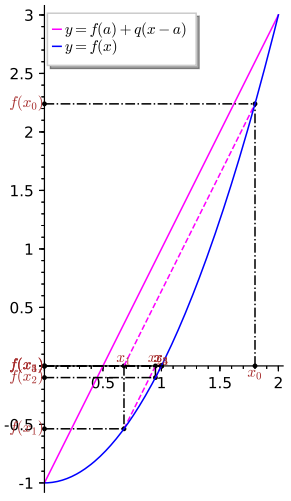
α_{tol}	:	un réel tel que $\lvert f(\alpha_{\text{tol}})\rvert \leqslant$ tol

1: **Fonction** $\alpha_{\text{tol}} \leftarrow$ CORDE ($f, a, b, x_0, \text{tol}, \text{kmax}$)
2: $q \leftarrow \frac{b-a}{f(b)-f(a)}$
3: $\Phi \leftarrow \big(x \mapsto x - q * f(x)\big)$ ▷ définition de fonction
4: $\alpha_{\text{tol}} \leftarrow$ PTFIXE$(\Phi, x_0, \text{tol}, \text{kmax})$
5: **Fin Fonction**

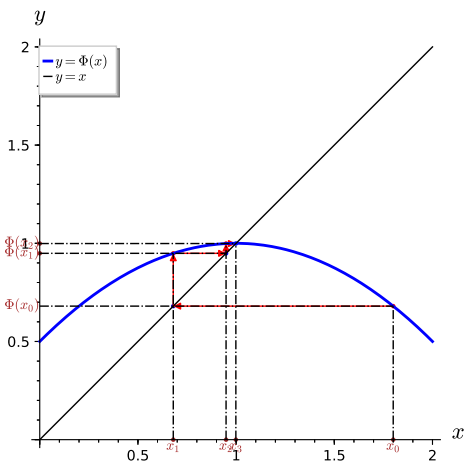
Pour illustrer ces résultats de convergence, on va rechercher la racine positive de $f(x) = x^2 - 1$ par la méthode de la corde avec comme données

- exemple 1 : $a = 0.000$, $b = 2.000$, $x_0 = 1.800$,
- exemple 2 : $a = 0.5000$, $b = 1.900$, $x_0 = 1.800$.

On représente en Figure 2.9 les itérations de la méthode de la corde pour l'exemple 1 à partir du graphe de la fonction f (figure de gauche) et à partir du graphe de la fonction Φ (figure de droite). Même chose pour l'exemple 2 avec la Figure 2.10.

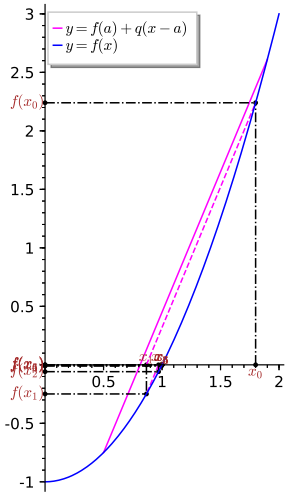


(a) représentation usuelle

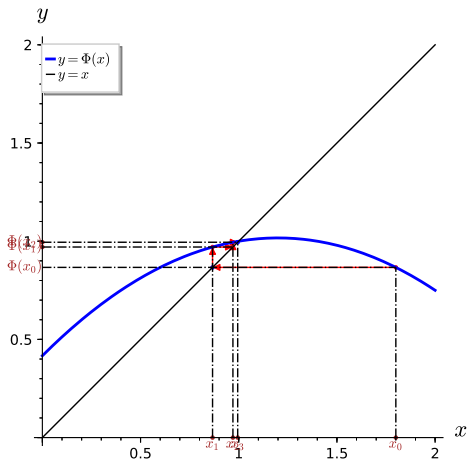


(b) Représentation point fixe

Figure 2.9: Exemple 1, méthode de la corde, $\alpha = 1$, racine de $f\ :\ x \mapsto x^2 - 1$ avec $a = 0.00$, $b = 2.00$, $x_0 = 1.80$,



(a) représentation usuelle



(b) Représentation point fixe

Figure 2.10: Exemple 2, méthode de la corde, $\alpha = 1$, racine de f : $x \mapsto x^2 - 1$ avec $a = 0.50$, $b = 1.90$, $x_0 = 1.80$,

Dans le tableau suivant on donne l'erreur commise par les suites dérivées des exemples 1 et 2 et ceci pour quelques itérations.

	exemple 1	**exemple 2**
k	$\lvert x_k - \alpha \rvert$	$\lvert x_k - \alpha \rvert$
0	8.0000e-01	8.0000e-01
1	3.2000e-01	1.3333e-01
2	5.1200e-02	2.9630e-02
3	1.3107e-03	5.3041e-03
4	8.5899e-07	8.9573e-04
5	3.6893e-13	1.4962e-04
6	0.0000e+00	2.4947e-05
⋮	⋮	⋮
15	0.0000e+00	2.4756e-12

On remarque qu'avec les données de l'exemple 1 la convergence est beaucoup plus rapide. Ceci est

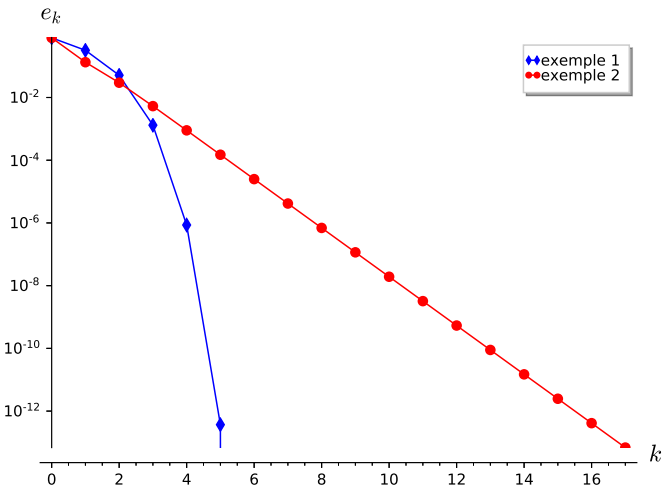


Figure 2.11: Erreurs en fonctions des itérations

illustré en Figure 2.11. En effet dans ce cas on a

$$\frac{f(b)-f(a)}{b-a} = 2 \quad \text{et} \quad f'(\alpha) = 2.$$

D'après la proposition 2.8, la convergence est d'ordre 2 pour l'exemple 1. Par contre, on a pour l'exemple 2

$$\frac{f(b)-f(a)}{b-a} = 2.400 \neq f'(\alpha) = 2$$

et donc la convergence est d'ordre 1. On retrouve ces résultats sur la Figure 2.12 ou l'on a représenté en échelle logarithmique $e_{k+1} = |x_{k+1} - \alpha|$ en fonction de $e_k = |x_k - \alpha|$. En effet si une méthode est convergente d'ordre p exactement on aura pour k suffisament grand $e_{k+1} \approx Ce_k^p$.

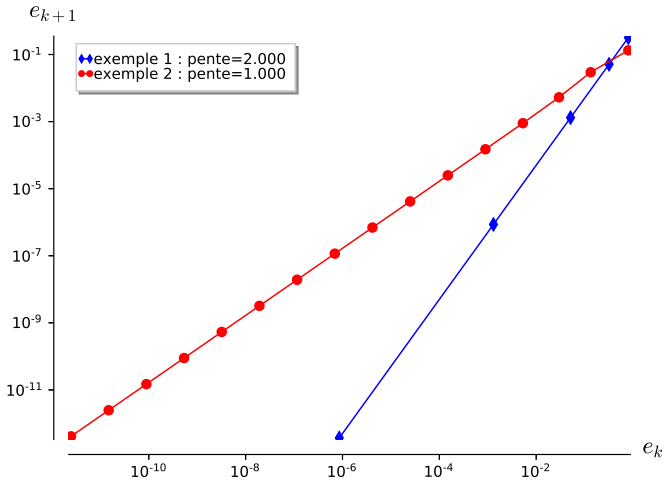


Figure 2.12: Représentation en échelle logarithmique de e_{k+1} en fonction de e_k. Les pentes sont calculées numériquement

Ajouter un autre exemple

La méthode de Newton

Soient f une fonction de classe $\mathcal{C}^1$ et x_0 un réel donné. La suite obtenue par la méthode de Newton est donnée par

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \ \forall k \in \mathbb{N}. \tag{2.18}$$

Bien évidemment, il faudra s'assurer que $f'(x_k) \neq 0$. On peut voir cette relation comme un cas particulier de la méthode du point fixe. En effet, en prenant $\Phi(x) = x - \frac{f(x)}{f'(x)}$, la suite définie en (2.20) s'écrit $x_{k+1} = \Phi(x_k)$.

Proposition 2.9: convergence de la méthode de Newton

Soit f une fonction de classe $\mathcal{C}^2$ sur un certain voisinage d'une racine simple α de f. Soit x_0 donné dans ce voisinage, la suite $(x_k)_{k\in\mathbb{N}}$ définie par la méthode de Newton

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \ \forall k \in \mathbb{N}. \tag{2.19}$$

est localement convergente d'ordre 2.

Preuve. Comme α est racine simple de f (i.e. $f(\alpha) = 0$ et $f'(\alpha) \neq 0$) et f' continue, il existe un voisinage $\mathcal{V}$ de α tel que pour tout $x \in \mathcal{V}$, $f'(x) \neq 0$. On peut alors définir la fonction Φ sur $\mathcal{V}$ par

$$\forall x \in \mathcal{V}, \quad \Phi(x) = x - \frac{f(x)}{f'(x)}.$$

On a alors $x_{k+1} = \Phi(x_k)$. La fonction Φ est de classe $\mathcal{C}^1$ sur $\mathcal{V}$ et

$$\Phi'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}.$$

On a donc $\Phi'(\alpha) = 0$ car $f(\alpha) = 0$. D'après le théorème de convergence locale du point fixe (théorème 2.5), on en déduit que la suite (x_k) converge vers α (et que la convergence est au moins d'ordre 1. Pour démontrer qu'elle est d'ordre 2, on ne peut utiliser le théorème 2.6 car Φ n'est pas de classe $\mathcal{C}^2$. Toutefois comme f est de classe $\mathcal{C}^2$, on applique la formule de Taylor-Lagrange aux points x_k, α (en supposant $x_k \neq \alpha$) : il existe ξ_k compris entre x_k et α tel que

$$\underbrace{f(\alpha)}_{=0} = f(x_k) + (\alpha - x_k)f'(x_k) + \frac{(\alpha - x_k)^2}{2!}f^{(2)}(\xi_k).$$

Comme $f'(x_k) \neq 0$, l'équation précédente s'écrit aussi

$$\begin{aligned}\alpha &= x_k - \frac{f(x_k)}{f'(x_k)} - (\alpha - x_k)^2\frac{f^{(2)}(\xi_k)}{2f'(x_k)} \\ &= x_{k+1} - (\alpha - x_k)^2\frac{f^{(2)}(\xi_k)}{2f'(x_k)}.\end{aligned}$$

On obtient alors

$$\frac{\alpha - x_{k+1}}{(\alpha - x_k)^2} = -\frac{f^{(2)}(\xi_k)}{2f'(x_k)}.$$

La fonction f est de classe $\mathcal{C}^2$ et la suite ξ_k converge vers α (car ξ_k compris entre x_k et α). Ceci entraine par passage à la limite que

$$\lim_{k\to+\infty} \frac{x_{k+1} - \alpha}{(x_k - \alpha)^2} = \frac{f^{(2)}(\alpha)}{2f'(\alpha)}.$$

La convergence est donc d'ordre 2. □

Soit f une fonction de classe $\mathcal{C}^2$ au voisinage de α, racine de f. On suppose que $f'(x) \neq 0$ pour tout $x \in \mathcal{V}$ (i.e. α racine simple de f). Soit $x_0 \in \mathcal{V}$ donné. La suite obtenue par la méthode de Newton est donnée par

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \ \forall k \in \mathbb{N}. \tag{2.20}$$

On peut voir cette relation comme un cas particulier de la méthode du point fixe. En effet, en prenant $\Phi(x) = x - \frac{f(x)}{f'(x)}$, la suite définie en (2.20) s'écrit $x_{k+1} = \Phi(x_k)$ et on note que $\Phi(\alpha) = \alpha$ (i.e. α point fixe de Φ).
De plus f étant de classe $\mathcal{C}^3$ et sa dérivée non nulle sur $\mathcal{V}$, on obtient que Φ est de classe $\mathcal{C}^2$ sur $\mathcal{V}$, et $\forall x \in \mathcal{V}$,

$$\Phi'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}$$

et

$$\Phi''(x) = \frac{(f'(x)f''(x) + f(x)f^{(3)}(x))(f'(x))^2 - 2f(x)f'(x)(f''(x))^2}{(f'(x))^4}$$

On a alors

$$\Phi'(\alpha) = 0, \quad \Phi''(\alpha) = \frac{f''(\alpha)}{f'(\alpha)}.$$

D'après la proposition 2.6, si $f''(\alpha) \neq 0$ alors la **méthode de Newton est d'ordre** 2.
Faire méthode de Newton modifiés dans le cas de racine de multiplicité $m > 1$.???

Exercice 2.2.3

En -1700 av. J.-C., les babyloniens ne connaissaient que les nombres rationnels (fractions) et ils utilisaient le système sexagésimal (base 60). Pour approcher la valeur $\sqrt{2}$, ils utilisaient comme approximation (voir tablette YBC 7289)

$$\alpha = 1 + \frac{24}{60} + \frac{51}{60^2} + \frac{10}{60^3} = \frac{30547}{21600}$$

L'erreur commise est $|\alpha - \sqrt{2}| \approx 5.994e-7$.

Q. 1 *Comment feriez-vous pour trouver* ***à la main*** *une méthode permettant de trouver des nombres rationnels approchant* $\sqrt{2}$.

Q. 2 *Généraliser la méthode pour trouver une approximation rationnelle de* $\sqrt{a}$ *où* a *est un réel positif.*

Q. 3 *Généraliser la méthode pour trouver une approximation rationnelle de* $\sqrt[n]{a}$ *où* a *est un réel positif et* $n \in \mathbb{N}^*$.

Correction Exercice 2.2.3

Q. 1 Il suffit de voir que $\sqrt{2}$ est la racine positive de $f(x) = x^2 - 2$ et d'appliquer la méthode de Newton par exemple. La suite des itérés de Newton s'écrit alors

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^2 - 2}{2x_k} = \frac{x_k^2 + 2}{2x_k}$$

Avec $x_0 = 1$, on obtient

k	x_k	$\lvert\sqrt{2} - x_k\rvert$
1	$\frac{3}{2}$	8.57864e-02
2	$\frac{17}{12}$	2.45310e-03
3	$\frac{577}{408}$	2.12390e-06

Avec $x_0 = \frac{5}{4}$, on obtient

k	x_k	$\lvert\sqrt{2} - x_k\rvert$
1	$\frac{57}{40}$	1.07864e-02
2	$\frac{6449}{4560}$	4.08236e-05
3	$\frac{83176801}{58814880}$	5.89203e-10

Q. 2 Il suffit de voir que $\sqrt{a}$ est la racine positive de $f(x) = x^2 - a$ et d'appliquer la méthode de Newton par exemple. La suite des itérés de Newton s'écrit alors

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^2 - a}{2x_k} = \frac{x_k^2 + a}{2x_k}$$

Avec $a = 3$ et $x_0 = 1$, on obtient

k	x_k	$\lvert\sqrt{3} - x_k\rvert$
1	2	2.67949e-01
2	$\frac{7}{4}$	1.79492e-02
3	$\frac{97}{56}$	9.20496e-05

Q. 3 Il suffit de voir que $\sqrt[n]{a}$ est la racine positive de $f(x) = x^n - a$ et d'appliquer la méthode de Newton par exemple. La suite des itérés de Newton s'écrit alors

$$x_{k+1} == x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k^n - a}{nx_k^{n-1}} = \frac{(n-1)x_k^n - a}{nx_k^{n-1}}$$

Avec $a = 3$, $n = 4$ et $x_0 = 1$, on obtient

k	x_k	$\lvert\sqrt[4]{3} - x_k\rvert$
1	$\frac{3}{2}$	1.83926e-01
2	$\frac{97}{72}$	3.11482e-02
3	$\frac{115403137}{87616608}$	1.06368e-03
4	$\frac{2362972972710088378167380851522 57}{1795469431997009848644834162648 32}$	1.28780e-06

◇

On présente en Algorithme 2.14 l'implémentation standard de la méthode de Newton. Une autre version utilisant la fonction PtFixe est donnée en Algorithme 2.15.

Algorithme 2.14 Méthode de Newton

Données :
f : $f : \mathbb{R} \longrightarrow \mathbb{R}$,
df : la dérivée de f,
x_0 : donnée initiale, $x_0 \in \mathbb{R}$,
tol : la tolérence, tol $\in \mathbb{R}^+$,
kmax : nombre maximum d'itérations, kmax $\in \mathbb{N}^*$
Résultat :
α_{tol} : un réel tel que $|\Phi(\alpha_{\text{tol}}) - \alpha_{\text{tol}}| \leqslant$ tol

1: **Fonction** $\alpha_{\text{tol}} \leftarrow$ Newton (f, df, x_0, tol, kmax)
2: $k \leftarrow 0$, $\alpha_{\text{tol}} \leftarrow \varnothing$
3: x $\leftarrow x_0$,
4: err $\leftarrow$ tol $+ 1$
5: **Tantque** err $>$ tol et $k \leqslant$ kmax **faire**
6: $k \leftarrow k + 1$
7: xp $\leftarrow$ x
8: x $\leftarrow$ xp $- f$(xp)/df(xp) ▷ df(xp) $\neq 0$
9: err $\leftarrow |$x $-$ xp$|$
10: **Fin Tantque**
11: **Si** err $\leqslant$ tol **alors** ▷ Convergence
12: $\alpha_{\text{tol}} \leftarrow$ x
13: **Fin Si**
14: **Fin Fonction**

Algorithme 2.15 Méthode de Newton scalaire

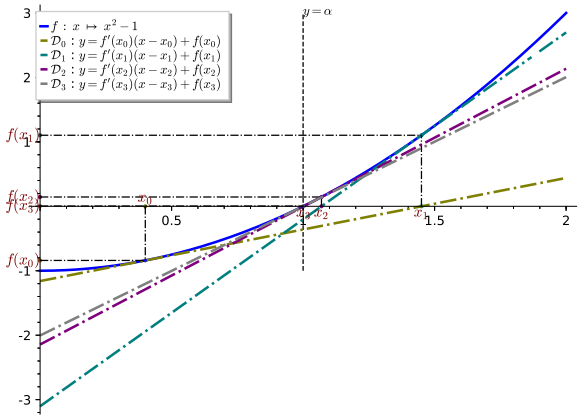
Données :
f : $f : \mathbb{R} \longrightarrow \mathbb{R}$,
df : la dérivée de f,
x_0 : donnée initiale, $x_0 \in \mathbb{R}$,
tol : la tolérence, tol $\in \mathbb{R}^+$,
kmax : nombre maximum d'itérations, kmax $\in \mathbb{N}^*$
Résultat :
α_{tol} : un réel tel que

1: **Fonction** $\alpha_{\text{tol}} \leftarrow$ Newton (f, df, x_0, tol, kmax)
2: $\Phi \leftarrow x \mapsto x - f(x)/\text{df}(x)$
3: $\alpha_{\text{tol}} \leftarrow$ PtFixe(Φ, x_0, tol, kmax)
4: **Fin Fonction**

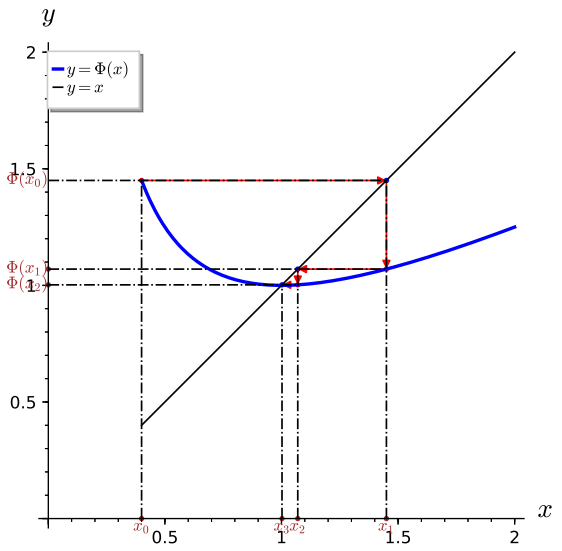
Comme premier exemple, on prend $f(x) = x^2 - 1$ avec $x_0 = 0.40$ et pour le second $f(x) = x^2 \cos(x)$ avec $x_0 = 2.00$.

On représente en Figures 2.13 and 2.14, respectivement pour les exemples 1 et 2, les itérations de la méthode de Newton à partir du graphe de la fonction f (figure de gauche) et à partir du graphe de la fonction Φ (figure de droite).

On illustre la convergence et l'ordre de convergence respectivement en Figures 2.15a et 2.15b. Pour cette dernière, on a représenté en échelle logarithmique $e_{k+1} = |x_{k+1} - \alpha|$ en fonction de $e_k = |x_k - \alpha|$. En effet si une méthode est convergente d'ordre p exactement on aura pour k suffisament grand $e_{k+1} \approx C e_k^p$.

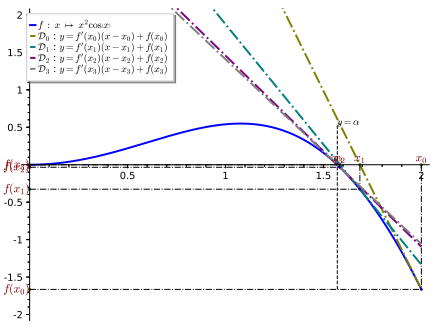


(a) représentation usuelle

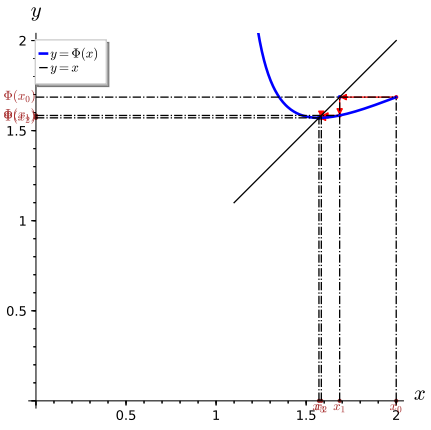


(b) Représentation point fixe, $\Phi \ : \ x \mapsto x - \frac{x^2-1}{2\,x}$

Figure 2.13: Exemple 1, méthode de Newton, $\alpha = 1$, racine de $f \ : \ x \mapsto x^2 - 1$ avec $x_0 = 0.40$,

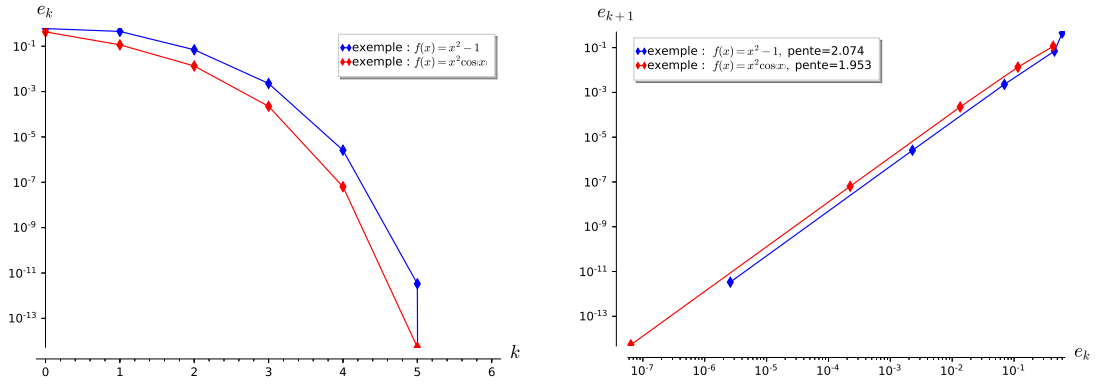


(a) représentation usuelle



(b) Représentation point fixe, $\Phi \ : \ x \mapsto \frac{x^2 \cos(x)}{x^2 \sin(x) - 2\,x \cos(x)} + x$

Figure 2.14: Exemple 2, méthode de Newton, $\alpha = \frac{1}{2}\,\pi$, racine de $f \ : \ x \mapsto x^2 \cos(x)$ avec $x_0 = 0.40$,



(a) Représentation de la convergence, e_k en fonction de k

(b) Représentation de l'ordre de convergence en échelle logarithmique, e_{k+1} en fonction de e_k . Ordre théorique 2

Figure 2.15: Méthode de Newton, convergence et ordre

2.2.5 La méthode de la sécante

Cette méthode est une alternative à la méthode de Newton lorsque l'on ne connait pas la dérivée de la fonction f. A l'itération k, de la méthode de Newton, on approche $f'(x_k)$ par $\frac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}}$. En effet, d'après la formule de Taylor-Lagrange, il existe ξ_k compris entre x_{k-1} et x_k vérifiant

$$f(x_{k-1}) = f(x_k) + (x_{k-1}-x_k)f'(x_k) + \frac{(x_{k-1}-x_k)^2}{2!}f^{(2)}(\xi_k)$$

ce qui donne

$$f'(x_k) = \frac{f(x_k)-f(x_{k-1})}{x_{k-1}-x_k} + \frac{x_{k-1}-x_k}{2!}f^{(2)}(\xi_k).$$

Si la suite est convergente, on a $\frac{x_{k-1}-x_k}{2!}f^{(2)}(\xi_k) \to 0$, ce qui justifie l'approximation précédente. On a alors la méthode de la sécante donnée en (2.21). Cette méthode est une **méthode à deux pas** : le calcul de x_{k+1} nécessite de connaitre x_k et x_{k-1}. Il faut donc deux valeurs d'initialisations x_{-1} et x_0 pour définir la suite (x_k).

Proposition 2.10: Convergence méthode de la sécante (Admis)

Soit f une fonction de classe $\mathcal{C}^2$ sur un certain voisinage d'une racine simple α de f. Soient x_{-1} et x_0 donnés dans ce voisinage tels que $f(x_{-1}) \neq f(x_0)$, la suite $(x_k)_{k\in\mathbb{N}}$ définie par la méthode de la sécante

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k)-f(x_{k-1})}f(x_k), \ \forall k \in \mathbb{N}. \tag{2.21}$$

est localement convergente d'ordre $\frac{1+\sqrt{5}}{2} \approx 1.618$.

Comme premier exemple, on recherche une racine de x^2-1 avec $x_{-1} = 0.000$ et $x_0 = 2.000$. Une représentation graphique des premières itérés de la suite est donnée en Figure 2.16. Sur cette figure, les droites $\mathcal{D}_k$ sont celles passant par les points $(x_{k-1}, f(x_{k-1}))$ et $(x_k, f(x_k))$. Pour deuxième exemple, on recherche une racine de $x^2\cos(x)$ avec $x_{-1} = 1.000$ et $x_0 = 3.000$. Une représentation graphique des premières itérés de la suite est donnée en Figure 2.17.

On illustre la convergence et l'ordre de convergence respectivement en Figures 2.18a et 2.18b. Pour cette dernière, on a représenté en échelle logarithmique $e_{k+1} = |x_{k+1}-\alpha|$ en fonction de $e_k = |x_k-\alpha|$. En effet si une méthode est convergente d'ordre p exactement on aura pour k suffisament grand $e_{k+1} \approx Ce_k^p$.

2.2.6 Méthode Regula-Falsi ou fausse position

Cette méthode diffère de la méthode de dichotomie par le choix de x_k à chaque itération. Pour la méthode de dichotomie on a pris $x_k = \frac{a_k+b_k}{2}$ point milieux du segment $[a_k, b_k]$. Pour la méthode Regula-Falsi, on

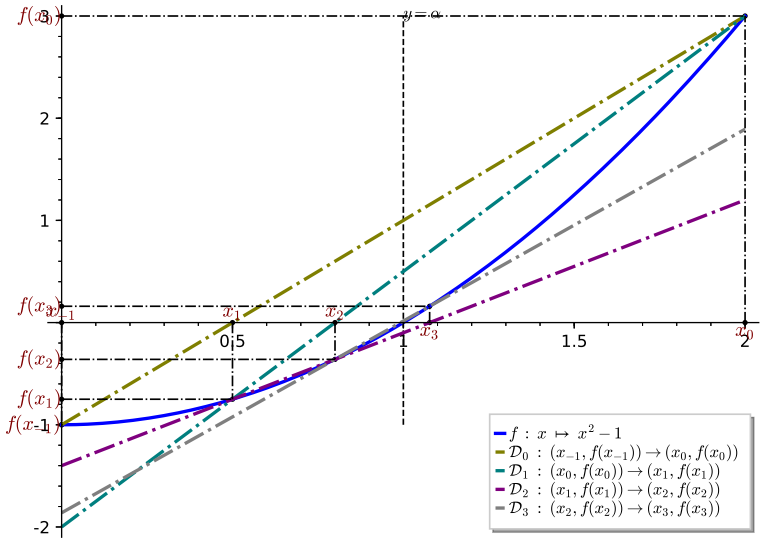


Figure 2.16: Méthode de la sécante pour $f(x) = x^2 - 1$, $x_{-1} = 0.000$ et $x_0 = 2.000$

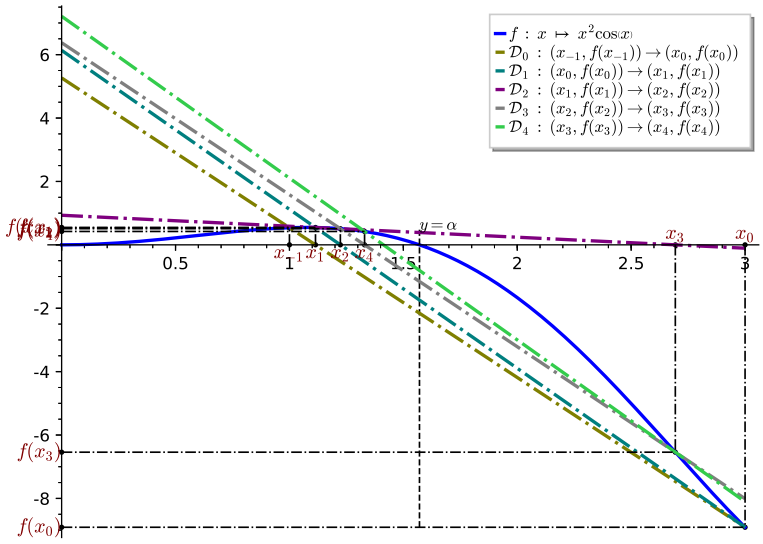
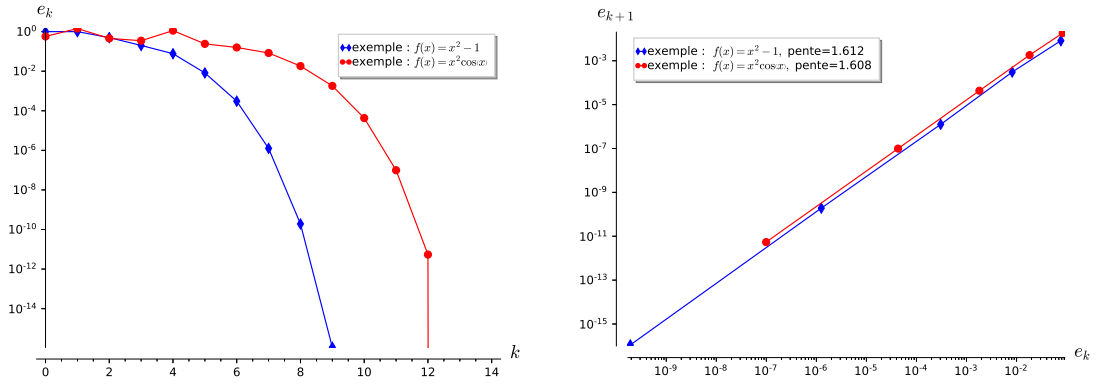


Figure 2.17: Méthode de la sécante pour $f(x) = x^2 \cos(x)$, $x_{-1} = 1.000$ et $x_0 = 3.000$



(a) Représentation de la convergence, e_k en fonction de k

(b) Représentation de l'ordre de convergence en échelle logarithmique, e_{k+1} en fonction de e_k . Ordre théorique $\frac{1+\sqrt{5}}{2} \approx 1.618$

Figure 2.18: Méthode de la sécante, convergence et ordre

prend pour x_k l'intersection de la droite passant par les points $(a_k, f(a_k))$ et $(b_k, f(b_k))$ avec l'axe des abscisses. Si $f(a_k)f(b_k) < 0$ cela nous assure que $x_k \in]a_k, b_k[$.
L'équation de la droite est donnée par

$$y = cx + d, \text{ avec } c = \frac{f(b_k) - f(a_k)}{b_k - a_k} \text{ et } d = \frac{b_k f(a_k) - a_k f(b_k)}{b_k - a_k}.$$

On a alors

$$x_k = -d/c = \frac{a_k f(b_k) - b_k f(a_k)}{f(b_k) - f(a_k)}.$$

En résumé, on définit les trois suites $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$ et $(x_k)_{k\in\mathbb{N}}$ par

- $a_0 = a$, $b_0 = b$ et $x_0 = \frac{a_0 f(b_0) - b_0 f(a_0)}{f(b_0) - f(a_0)}$,
- $\forall k \in \mathbb{N}$,

$$\begin{cases} a_{k+1} = b_{k+1} = x_{k+1} = x_k, & \text{si } f(x_k) = 0, \\ a_{k+1} = x_k,\ b_{k+1} = b_k, & \text{si } f(b_k)f(x_k) < 0, \\ a_{k+1} = a_k,\ b_{k+1} = x_k, & \text{si } f(a_k)f(x_k) < 0, \\ x_{k+1} = \frac{a_{k+1} f(b_{k+1}) - b_{k+1} f(a_{k+1})}{f(b_{k+1}) - f(a_{k+1})}, & \text{si } f(x_k) \neq 0. \end{cases}$$

Exercice 2.2.4

On suppose que la fonction f est continue sur $[a, b]$, vérifie $f(a)f(b) < 0$ et qu'il existe un unique $\xi \in]a, b[$ tel que $f(\xi) = 0$.

Q. 1 *Montrer que*

$$a \leqslant \frac{af(b) - bf(a)}{f(b) - f(a)} \leqslant b.$$

Q. 2 *Montrer que* $a_0 \leqslant a_1 \leqslant \ldots \leqslant a_k \leqslant x_k \leqslant b_k \leqslant \ldots \leqslant b_1 \leqslant b_0$ *pour tout* $k \in \mathbb{N}$. *et que si* $f(x_k) \neq 0$ *alors* $f(a_k)f(b_k) < 0$.

Q. 3 *En déduire la convergence de la suite* (x_k) *vers* ξ.

Correction Exercice 2.2.4

Q. 1 On pose $x = \frac{af(b) - bf(a)}{f(b) - f(a)}$ qui est bien défini car $f(a) \neq f(b)$. En effet si $f(a) = f(b)$ alors $f(a)f(b) = f(a)^2 \geqslant 0$ qui est en contradiction avec l'hypothèse $f(a)f(b) < 0$. On a

$$x - a = \frac{af(b) - bf(a) - a(f(b) - f(a))}{f(b) - f(a)} = (b - a)\frac{f(a)}{f(a) - f(b)}$$

Comme $f(a)f(b) < 0$, $f(a) - f(b)$ est du même signe que $f(a)$ et alors $\frac{f(a)}{f(a)-f(b)} \geqslant 0$. De plus $b - a > 0$ et donc on a $x - a \geqslant 0$.
De la même manière, on a

$$x - b = \frac{af(b) - bf(a) - b(f(b) - f(a))}{f(b) - f(a)} = (a - b)\frac{f(b)}{f(b) - f(a)}$$

Comme $f(a)f(b) < 0$, $f(b) - f(a)$ est du même signe que $f(b)$ et alors $\frac{f(b)}{f(b)-f(a)} \geqslant 0$. De plus $a - b < 0$ et donc on a $x - b \leqslant 0$.

Q. 2 On va démontrer par récurrence la validité de la proposition $(\mathcal{P}_k)$ suivante $\forall k \in \mathbb{N}$:

$$(\mathcal{P}_k) \quad \begin{cases} \text{(i)} & f(a_k)f(b_k) < 0 \text{ si } f(x_k) \neq 0 \\ \text{(ii)} & a_0 \leqslant a_1 \leqslant \ldots \leqslant a_k \leqslant x_k \leqslant b_k \leqslant \ldots \leqslant b_1 \leqslant b_0, \end{cases}$$

Initialisation On vérifie la proposition $(\mathcal{P}_k)$ pour $k = 0$. On a $f(a)f(b) < 0$ donc $(\mathcal{P}_0) - (i)$ est vérifiée. D'après **Q. 1**, on obtient $a_0 \leqslant x_0 \leqslant b_0$. La proposition $(\mathcal{P}_0)$ est donc vérifiée.

Hérédité Soit $k \geqslant 1$. On suppose la proposition $(\mathcal{P}_k)$ est vraie. Montrons que $(\mathcal{P}_{k+1})$ est vérifiée.
Si $f(x_k) = 0$ alors $a_{k+1} = b_{k+1} = x_{k+1} = x_k$ et la proposition $(\mathcal{P}_{k+1})$ est vérifiée.
On suppose maintenant que $f(x_k) \neq 0$.
De $(\mathcal{P}_k) - (i)$ on a $f(a_k) \neq 0$ et $f(b_k) \neq 0$. De plus $f(a_k) \neq f(b_k)$ car sinon $f(a_k)f(b_k) = f(a_k)^2 \geqslant 0$ ce qui est en contradiction avec $(\mathcal{P}_k) - (i)$. Par hypothèse $(\mathcal{P}_k)$, on a $a_k \leqslant x_k \leqslant b_k$ et $f(a_k)f(b_k) < 0$. Par continuité de f, on a alors soit $f(b_k)f(x_k) < 0$ (et donc $f(a_k)f(x_k) > 0$) soit $f(b_k)f(x_k) > 0$ (et donc $f(a_k)f(x_k) < 0$).

- Si $f(b_k)f(x_k) < 0$ on a $a_{k+1} = x_k$ et $b_{k+1} = b_k$. Comme $f(a_k) \neq f(b_k)$, x_{k+1} est bien défini. D'après **Q. 1** en prenant $[a_{k+1}, b_{k+1}]$ comme intervalle, et sachant que $f(a_{k+1})f(b_{k+1}) = f(x_k)f(b_k) < 0$ on obtient
$$a_{k+1} \leqslant x_{k+1} \leqslant b_{k+1}.$$
De plus par hypothèse $a_k \leqslant x_k = a_{k+1}$ et donc $a_k \leqslant a_{k+1}$ et $b_{k+1} \leqslant b_k$. La proposition $(\mathcal{P}_{k+1})$ est donc vérifiée.
- Si $f(a_k)f(x_k) < 0$ on a $a_{k+1} = a_k$ et $b_{k+1} = x_k$. Comme $f(a_k) \neq f(b_k)$, x_{k+1} est bien défini. D'après **Q. 1** en prenant $[a_{k+1}, b_{k+1}]$ comme intervalle, et sachant que $f(a_{k+1})f(b_{k+1}) = f(a_k)f(x_k) < 0$ on obtient
$$a_{k+1} \leqslant x_{k+1} \leqslant b_{k+1}.$$
La proposition $(\mathcal{P}_{k+1})$ est donc vérifiée.

Q. 3 Supposons qu'il existe $s \in \mathbb{N}$ tel que $f(x_s) = 0$. Alors, pour tout $i \leqslant 1$, on a $a_{s+i} = b_{s+i} = x_{s+i} = x_s$. Les trois suites convergent donc vers x_s. D'après la question précédente, $x_s \in [a, b]$. Comme $f(a) \neq 0$ et $f(b) \neq 0$, on en déduit $x_s \in]a, b[$. Par hypothèse il existe un unique $\xi \in]a, b[$ tel que $f(\xi) = 0$, on a alors $x_s = \xi$.
Supposons que $\forall k \in \mathbb{N}$, $f(x_k) \neq 0$. D'après **Q. 2**, la suite (a_k) est croissante majorée par b et la suite (b_k) est décroissante minorée par a. Elles sont donc convergentes et l'on note respectivement l et L les limites de (a_k) et (b_k). Comme $a \leqslant a_k \leqslant b_k \leqslant b$, on a $a \leqslant l \leqslant L \leqslant b$.

- Supposons $f(l) = f(L)$. On a $f(a_k)f(b_k) < 0$. Comme f est continue, à la limite on obtient $f(l)f(L) = f(l)^2 = f(L)^2 \leqslant 0$ et donc $f(l) = f(L) = 0$. On a necessairement l et L dans $]a, b[$ car $f(a)f(b) < 0$ et donc $f(a)$ et $f(b)$ non nuls. Par unicité du zéro de f dans $]a, b[$ on obtient $l = L = \xi$. Comme $a_k \leqslant x_k \leqslant b_k$, on en déduit que la suite (x_k) converge aussi vers ξ.
- Supposons $f(l) \neq f(L)$. Par continuité de la fonction f la suite (x_k) converge alors vers $M = \frac{lf(L)-Lf(l)}{f(L)-f(l)}$. Comme $a_k \leqslant x_k \leqslant b_k$ on a aussi

$$l \leqslant M = \frac{lf(L) - Lf(l)}{f(L) - f(l)} \leqslant L. \tag{2.22}$$

De plus ayant $f(x_k) \neq 0$ $\forall k \in \mathbb{N}$, on a $f(a_k)f(b_k) < 0$ $\forall k \in \mathbb{N}$. En passant à la limite et par continuité de f on obtient $f(l)f(L) \leqslant 0$.
Montrons que $f(l) = 0$ ou $f(L) = 0$.

– Si $f(l) < f(L)$, alors de (2.22) on obtient

$$l(f(L) - f(l)) \leqslant lf(L) - Lf(l) \leqslant L(f(L) - f(l))$$

ce qui donne $lf(l) \geqslant Lf(l)$ et $lf(L) \leqslant Lf(L)$ i.e. $(l - L)f(l) \geqslant 0$ et $(L - l)f(L) \leqslant 0$. Comme $L - l \geqslant 0$, on en déduit $f(l) \leqslant 0$ et $f(L) \leqslant 0$. Or $f(l)f(L) \leqslant 0$, ce qui donne $f(l) = 0$ ou $f(L) = 0$.

– Si $f(l) > f(L)$, alors de (2.22) on obtient

$$l(f(L) - f(l)) \geqslant lf(L) - Lf(l) \geqslant L(f(L) - f(l))$$

ce qui donne $lf(l) \leqslant Lf(l)$ et $lf(L) \geqslant Lf(L)$ i.e. $(L - l)f(l) \geqslant 0$ et $(L - l)f(L) \leqslant 0$. Comme $L - l \geqslant 0$, on en déduit $f(l) \geqslant 0$ et $f(L) \geqslant 0$. Or $f(l)f(L) \leqslant 0$, ce qui donne $f(l) = 0$ ou $f(L) = 0$.

On a donc démontré que si $f(l) \neq f(L)$, alors $f(l) = 0$ ou $f(L) = 0$ et donc

– si $f(l) = 0$ alors $M = \frac{lf(L)-Lf(l)}{f(L)-f(l)} = l$ et donc $f(M) = 0$.

– si $f(L) = 0$ alors $M = \frac{lf(L)-Lf(l)}{f(L)-f(l)} = L$ et donc $f(M) = 0$.

Puisque l et L appartiennent à $]a, b[$, on a $M \in]a, b[$. Par unicité du zéro de f sur $]a, b[$ on en déduit que $M = \xi$.

◇

On vient de démontrer le théorème suivant

Théorème 2.11

Soit $f : [a, b] \subset \mathbb{R} \longrightarrow \mathbb{R}$ une fonction continue vérifiant $f(a)f(b) < 0$ et admettant $\alpha \in]a, b[$ comme **unique** solution de $f(x) = 0$. Alors la suite $(x_k)_{k\in\mathbb{N}}$ définie par la **méthode Regula-Falsi** converge vers α

On a aussi la proposition suivante

Proposition 2.12: ordre de la méthode de Regula-Falsi

Soit $f : [a, b] \subset \mathbb{R} \longrightarrow \mathbb{R}$ une fonction continue vérifiant $f(a)f(b) < 0$ et admettant $\alpha \in]a, b[$ comme **unique** solution de $f(x) = 0$. Si f est deux fois dérivables sur $[a, b]$ et si f'' est monotone sur $]a, b[$ alors il existe $C \in \mathbb{R}$ tel que

$$\lim_{k\to+\infty} \frac{|x_{k+1} - \alpha|}{|x_k - \alpha|} \leqslant C \tag{2.23}$$

La méthode de Regula-Falsi est alors à convergence linéaire (d'ordre 1).

2.3 Résolution de systèmes non linéaires

Nous allons (très rapidement) introduire les premières notions permettant la résolution de systèmes d'équations non linéaires. Par exemple, dans $\mathbb{R}^2$ nous allons regarder le problème suivant avec c une constante réelle

$$\begin{cases} f_1(x_1, x_2) = -x_1^3 + x_2 - \frac{1}{2} & = 0 \\ f_2(x_1, x_2) = \frac{1}{25}\left(10\, x_2 + 1\right)^2 + c - x_1 & = 0. \end{cases} \tag{2.24}$$

En Figures 2.19 et 2.21, on représente pour différentes valeurs de c les courbes $f_1(x_1, x_2) = 0$ et $f_2(x_1, x_2) = 0$: les intersections de ces deux courbes sont les solutions du problème (2.24). Comme on le voit graphiquemment, il peut y avoir, suivant les valeurs de c, 0, 1, 2 ou 4 solutions.

De manière plus générale, soient $U \subset \mathbb{R}^N$ un ouvert et $\boldsymbol{f}$ une application continue de U dans $\mathbb{R}^N$. Le problème que l'on souhaite résoudre est le suivant

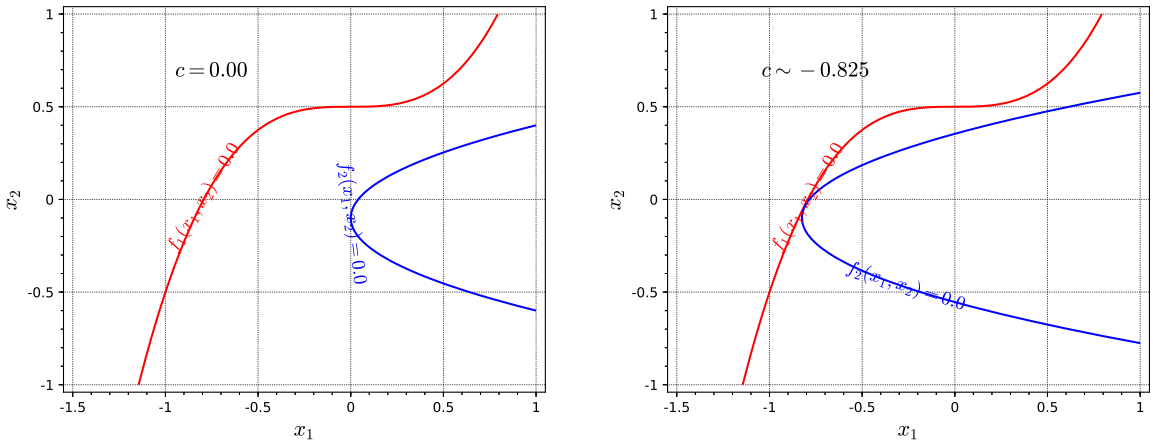


Figure 2.19: Résolution graphique de 2.24 avec $c = 0.00$ (gauche) et $c \sim -0.825$ (droite)

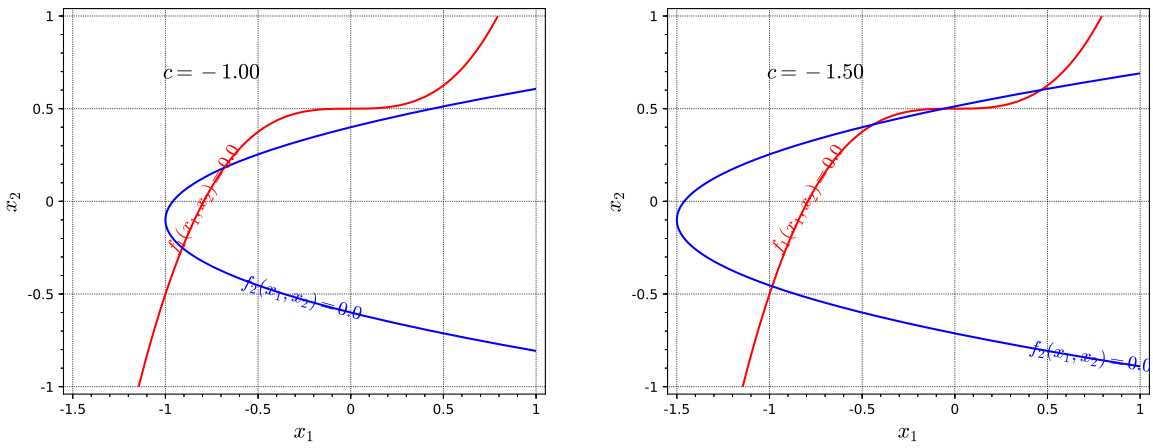


Figure 2.20: Résolution graphique de 2.24 avec $c = -1.00$ (gauche) et $c = -1.50$ (droite)

Trouver $\boldsymbol{\alpha} \in U \subset \mathbb{R}^N$ tel que

$$\boldsymbol{f}(\boldsymbol{\alpha}) = 0 \iff \begin{cases} \boldsymbol{f}_1(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N) &= 0 \\ \boldsymbol{f}_2(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N) &= 0 \\ \vdots & \\ \boldsymbol{f}_N(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N) &= 0 \end{cases}$$

Comme dans le cas scalaire, pour résoudre numériquement ce genre de problème on utilise des suites itératives et plus particulièrement celles basées sur les méthodes de points fixes. En effet, nous allons voir que le théorème du point fixe se généralise très facilement (voir Théorème 2.13).

En définissant, par exemple, la fonction $\boldsymbol{\Phi} \in \mathcal{C}^0(U; \mathbb{R}^N)$ par $\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{f}(\boldsymbol{x})$, on peut remarquer $\boldsymbol{f}(\boldsymbol{x}) = 0$ est équivalent à $\boldsymbol{\Phi}(\boldsymbol{x}) = \boldsymbol{x}$. On peut donc se ramener à la recherche d'un point fixe (s'il existe) de la fonction $\boldsymbol{\Phi}$.

Trouver $\boldsymbol{\alpha} \in U \subset \mathbb{R}^N$ tel que

$$\boldsymbol{\Phi}(\boldsymbol{\alpha}) = \boldsymbol{\alpha} \iff \begin{cases} \boldsymbol{\Phi}_1(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N) &= \boldsymbol{\alpha}_1 \\ \boldsymbol{\Phi}_2(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N) &= \boldsymbol{\alpha}_2 \\ \vdots & \\ \boldsymbol{\Phi}_N(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N) &= \boldsymbol{\alpha}_N \end{cases}$$

Les suites itératives sont donc de la forme

$$\boldsymbol{x}^{[k+1]} = \boldsymbol{\Phi}(\boldsymbol{x}^{[k]})$$

où $\boldsymbol{\Phi}$ est une fonction à déterminer et $\boldsymbol{x}^{[0]} \in \mathbb{R}^N$. Bien évidemment le choix d'une *bonne* fonction $\boldsymbol{\Phi}$ est primordiale pour espérer avoir convergence.

Ce type de problème peut s'avérer délicat à traiter : comment choisir $\boldsymbol{\Phi}$? $\boldsymbol{x}^{[0]}$? Si l'on converge vers quel point fixe?

2.3.1 Point fixe

Théorème 2.13

Soit $\mathcal{B}$ un espace de Banach et $U \subset \mathcal{B}$ un sous-ensemble fermé. On suppose que $\boldsymbol{\Phi} : U \longrightarrow U$ est une application strictement contractante, i.e.

$$\exists L \in]0,1[, \quad \|\boldsymbol{\Phi}(\boldsymbol{x}) - \boldsymbol{\Phi}(\boldsymbol{y})\| \leqslant L \|\boldsymbol{x} - \boldsymbol{y}\|, \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in U \times U. \tag{2.25}$$

Alors

1. $\boldsymbol{\Phi}$ admet un unique point fixe $\boldsymbol{\alpha} \in U$ (i.e. unique solution de $\boldsymbol{x} = \boldsymbol{\Phi}(\boldsymbol{x})$).
2. La suite des itérés $\boldsymbol{x}^{[k+1]} = \boldsymbol{\Phi}(\boldsymbol{x}^{[k]})$ converge vers $\boldsymbol{\alpha}$ pour toute valeur initiale $\boldsymbol{x}^{[0]} \in U$.
3. Pour tout $k \in \mathbb{N}$,
$$\left\|\boldsymbol{\alpha} - \boldsymbol{x}^{[k]}\right\| \leqslant \frac{L^{k-l}}{1-L} \left\|\boldsymbol{x}^{[l+1]} - \boldsymbol{x}^{[l]}\right\|, \quad 0 \leqslant l \leqslant k \tag{2.26}$$

Preuve. On démontre tout d'abord l'existence d'un point fixe. Pour celà on va démontrer que la suite $\boldsymbol{x}^{[k]}$ est de Cauchy dans U fermé d'un espace de Banach (donc elle converge dans U).
Comme $\boldsymbol{\Phi}$ est contractante, on a pour tout $k \in \mathbb{N}^*$

$$\left\|\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]}\right\| = \left\|\boldsymbol{\Phi}(\boldsymbol{x}^{[k]}) - \boldsymbol{\Phi}(\boldsymbol{x}^{[k-1]})\right\| \leqslant L \left\|\boldsymbol{x}^{[k]} - \boldsymbol{x}^{[k-1]}\right\|$$

ce qui donne par récurrence pour tout $0 \leqslant j \leqslant k$

$$\left\|\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]}\right\| \leqslant L^j \left\|\boldsymbol{x}^{[k+1-j]} - \boldsymbol{x}^{[k-j]}\right\| \tag{2.27}$$

ou encore pour tout $0 \leqslant l \leqslant k$, $(l = k - j)$

$$\left\|\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]}\right\| \leqslant L^{k-l} \left\|\boldsymbol{x}^{[l+1]} - \boldsymbol{x}^{[l]}\right\| \tag{2.28}$$

On obtient aussi par récurrence

$$\forall l \geqslant 0, \ \left\|\boldsymbol{x}^{[k+1+l]} - \boldsymbol{x}^{[k+l]}\right\| \leqslant L^{l} \left\|\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]}\right\|. \tag{2.29}$$

Soit $p \geqslant 1$. On en déduit par application répétée de l'inégalité triangulaire que

$$\begin{aligned}
\left\|\boldsymbol{x}^{[k+p]} - \boldsymbol{x}^{[k]}\right\| &= \left\|(\boldsymbol{x}^{[k+p]} - \boldsymbol{x}^{[k+p-1]}) + (\boldsymbol{x}^{[k+p-1]} - \boldsymbol{x}^{[k+p-2]}) + \ldots + (\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]})\right\| \\
&= \left\|\sum_{l=0}^{p-1} (\boldsymbol{x}^{[k+l+1]} - \boldsymbol{x}^{[k+l]})\right\| \\
&\leqslant \sum_{l=0}^{p-1} \left\|\boldsymbol{x}^{[k+l+1]} - \boldsymbol{x}^{[k+l]}\right\| \\
&\overset{(2.29)}{\leqslant} \sum_{l=0}^{p-1} L^{l} \left\|\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]}\right\| = \frac{1-L^p}{1-L} \left\|\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]}\right\| \\
&\leqslant \frac{1-L^p}{1-L} L^k \left\|\boldsymbol{x}^{[1]} - \boldsymbol{x}^{[0]}\right\|. \ \textit{(en utilisant (2.28), avec } l = 0\textit{)}
\end{aligned}$$

Comme $L^k \to 0$ quand $k \to +\infty$, on conclu que $(\boldsymbol{x}^{[k]})$ est une suite de Cauchy. De plus, par construction $\boldsymbol{x}^{[k]} \in U \subset \mathcal{B}$, pour tout $k \in \mathbb{N}$, et $\mathcal{B}$ étant un espace de Banach et U un fermé, la suite $(\boldsymbol{x}^{[k]})$ converge alors vers $\boldsymbol{\alpha} \in U$. Comme $\boldsymbol{\Phi}$ est contractante, elle est donc continue et en passant à la limite dans $\boldsymbol{x}^{[k+1]} = \boldsymbol{\Phi}(\boldsymbol{x}^{[k]})$, on abouti à $\boldsymbol{\alpha} = \boldsymbol{\Phi}(\boldsymbol{\alpha})$, i.e. $\boldsymbol{\alpha}$ est un point fixe de Φ dans U.
L'unicité se déduit immédiatement par la contraction de la fonction $\boldsymbol{\Phi}$. En effet, soit $\boldsymbol{\alpha}_1$ et $\boldsymbol{\alpha}_2$ deux points fixes de $\boldsymbol{\Phi}$, alors

$$\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\| = \|\boldsymbol{\Phi}(\boldsymbol{\alpha}_1) - \boldsymbol{\Phi}(\boldsymbol{\alpha}_2)\| \leqslant L \|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|$$

Or $L < 1$, et donc nécessairement on a $\boldsymbol{\alpha_1} = \boldsymbol{\alpha_2}$.

Il reste à démontrer l'inégalité (2.26). On a vu que pour $p \geqslant 1$

$$\left\|\boldsymbol{x}^{[k+p]} - \boldsymbol{x}^{[k]}\right\| \leqslant \frac{1-L^p}{1-L} \left\|\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]}\right\|$$

L'application norme étant continue et $L < 1$, on obtient à la limite quand $p \to +\infty$

$$\left\|\boldsymbol{\alpha} - \boldsymbol{x}^{[k]}\right\| \leqslant \frac{1}{1-L} \left\|\boldsymbol{x}^{[k+1]} - \boldsymbol{x}^{[k]}\right\|$$

On obtient l'inégalité en utilisant (2.28). □

2.3.2 Méthode de Newton

On commence par rappeler un résultat de calcul différentiel. Soit $\boldsymbol{f} : \mathbb{R}^N \longrightarrow \mathbb{R}^N$ une fonction suffisament régulière. On défini la **matrice Jacobienne de** $\boldsymbol{f}$, notée $\mathbb{J}_{\boldsymbol{f}}$, par

$$\mathbb{J}_{\boldsymbol{f}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_N} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_N}{\partial x_1} & \frac{\partial f_N}{\partial x_2} & \cdots & \frac{\partial f_N}{\partial x_N} \end{pmatrix}$$

On a alors $\forall \boldsymbol{h} \in \mathbb{R}^N$ à l'ordre 1

$$\boldsymbol{f}(\boldsymbol{x} + \boldsymbol{h}) \approx \boldsymbol{f}(\boldsymbol{x}) + \mathbb{J}_{\boldsymbol{f}}(\boldsymbol{x}).\boldsymbol{h}. \tag{2.30}$$

Nous voulons trouver $\boldsymbol{\alpha}$ tel que $\boldsymbol{f}(\boldsymbol{\alpha}) = 0$. Si $\boldsymbol{x}^{[k]}$ est proche de $\boldsymbol{\alpha}$, alors en utilisant (2.30) avec $\boldsymbol{x} = \boldsymbol{x}^{[k]}$ et $\boldsymbol{\alpha} = \boldsymbol{x}^{[k]} + \boldsymbol{h}$ (i.e. $\boldsymbol{h} = \boldsymbol{\alpha} - \boldsymbol{x}^{[k]}$) on obtient

$$\boldsymbol{f}(\boldsymbol{\alpha}) \approx \boldsymbol{f}(\boldsymbol{x}^{[k]}) + \mathbb{J}_{\boldsymbol{f}}(\boldsymbol{x}^{[k]}).\boldsymbol{h}$$

Au lieu de résoudre $\boldsymbol{f}(\boldsymbol{x}) = 0$, on résoud le système linéarisé

$$\boldsymbol{f}(\boldsymbol{x}^{[k]}) + \mathbb{J}_{\boldsymbol{f}}(\boldsymbol{x}^{[k]}).\tilde{\boldsymbol{h}} = 0$$

c'est à dire le système linéaire

$$\mathbb{J}_{\boldsymbol{f}}(\boldsymbol{x}^{[k]}).\tilde{\boldsymbol{h}} = -\boldsymbol{f}(\boldsymbol{x}^{[k]}). \tag{2.31}$$

On espère alors que $\tilde{\boldsymbol{h}}$ est une bonne approximation de $\boldsymbol{h}$ au sens ou $\boldsymbol{x}^{[k]}+\tilde{\boldsymbol{h}}$ est une meilleur approximation de $\boldsymbol{\alpha}$ que $\boldsymbol{x}^{[k]}$. On note alors $\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} + \tilde{\boldsymbol{h}}$. En posant $\Phi(\boldsymbol{x}) = \boldsymbol{x} - ((\mathbb{J}_{\boldsymbol{f}}(\boldsymbol{x}))^{-1}\boldsymbol{f}(\boldsymbol{x})$ la méthode de Newton s'écrit alors

$$\boldsymbol{x}^{[k+1]} = \Phi(\boldsymbol{x}^{[k]}) = \boldsymbol{x}^{[k]} - \left((\mathbb{J}_{\boldsymbol{f}}(\boldsymbol{x}^{[k]})\right)^{-1}\boldsymbol{f}(\boldsymbol{x}^{[k]}) \tag{2.32}$$

Cette formule est une généralisation de celle vue dans le cas scalaire (voir **??**). Il faut noter qu'a chaque itération la matrice Jacobienne est modifiée et qu'il faut calculer le vecteur $-\left((\mathbb{J}_{\boldsymbol{f}}(\boldsymbol{x}^{[k]})\right)^{-1}\boldsymbol{f}(\boldsymbol{x}^{[k]})$. Numeriquement, on ne calcule que très rarement l'inverse d'une matrice car celà est très couteux en temps mais on résoud le système linéaire (2.31) ce qui est bien plus efficace.

On admet dans ce cours le théorème suivant

Théorème 2.14

Soit $\boldsymbol{f} \in \mathcal{C}^3(\mathbb{R}^N;\mathbb{R}^N)$. On suppose que la matrice Jacobienne appliquée en $\boldsymbol{x}$, $\mathbb{J}_{\boldsymbol{f}}(\boldsymbol{x})$ est inversible dans un voisinage de $\boldsymbol{\alpha}$, avec $\boldsymbol{f}(\boldsymbol{\alpha}) = 0$. Alors pour tout $\boldsymbol{x}^{[0]}$ suffisament proche de $\boldsymbol{\alpha}$ la suite définie par

$$\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} - \left((\mathbb{J}_{\boldsymbol{f}}(\boldsymbol{x}^{[k]})\right)^{-1}\boldsymbol{f}(\boldsymbol{x}^{[k]})$$

converge vers $\boldsymbol{\alpha}$ et la convergence est d'ordre 2.

On donne ensuite l'algorithme 2.16 permettant de déterminer une approximation d'un point fixe d'une fonction $\boldsymbol{f}$. Dans cet algorithme on suppose donnée la fonction Solve permettant de résoudre un système linéaire.

Algorithme 2.16 Méthode de Newton

Données :

$\boldsymbol{f}$	:	$\boldsymbol{f} : \mathbb{R}^N \longrightarrow \mathbb{R}^N$,
Jf	:	la matrice Jacobienne de f,
x0	:	donnée initiale, x0 $\in \mathbb{R}^N$,
tol	:	la tolérence, tol $\in \mathbb{R}^+$,
kmax	:	nombre maximum d'itérations, kmax $\in \mathbb{N}^*$

Résultat :

$\boldsymbol{\alpha}_{\text{tol}}$: un élément de $\mathbb{R}^N$ proche de $\boldsymbol{\alpha}$.

Fonction $\alpha_{\text{tol}} \leftarrow$ Newton ($\boldsymbol{f}$, Jf, x0, tol, kmax)
  $k \leftarrow 0$, $\alpha_{\text{tol}} \leftarrow \varnothing$
  x $\leftarrow$ x0,
  err $\leftarrow$ tol + 1
  Tantque err $>$ tol et $k \leqslant$ kmax **faire**
    $k \leftarrow k + 1$
    xp $\leftarrow$ x
    h $\leftarrow$ Solve(Jf(xp), $-f$(xp)) ▷ $\boldsymbol{x} \leftarrow$ Solve($\mathbb{A}, \boldsymbol{b}$) : résoud le système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$
    x $\leftarrow$ xp + h
    err $\leftarrow$ Norm(x − xp)
  Fin Tantque
  Si err $\leqslant$ tol **alors** ▷ Convergence
    $\alpha_{\text{tol}} \leftarrow$ x
  Fin Si
Fin Fonction

Remarque 2.15 Si l'on ne connait pas explicitement la Jacobienne de f, il est possible de calculer une approximation de celle-ci en utilisant des formules de dérivation numérique.

2.3.3 Exemples

Exemple modèle

Comme premier exemple, nous reprenons le système 2.24 avec $c = -1.5$

$$\begin{cases} f_1(x_1, x_2) = -x_1^3 + x_2 - \frac{1}{2} & = 0 \\ f_2(x_1, x_2) = \frac{1}{25}\left(10\,x_2 + 1\right)^2 - x_1 - \frac{3}{2} & = 0. \end{cases} \tag{2.33}$$

On représente en Figure 2.21 les itérées successives pour 4 suites avec une initialisation différentes. On remarque qu'il est très difficile, si l'on n'est pas suffisament proche d'un point fixe, de prédire vers lequel on converge.

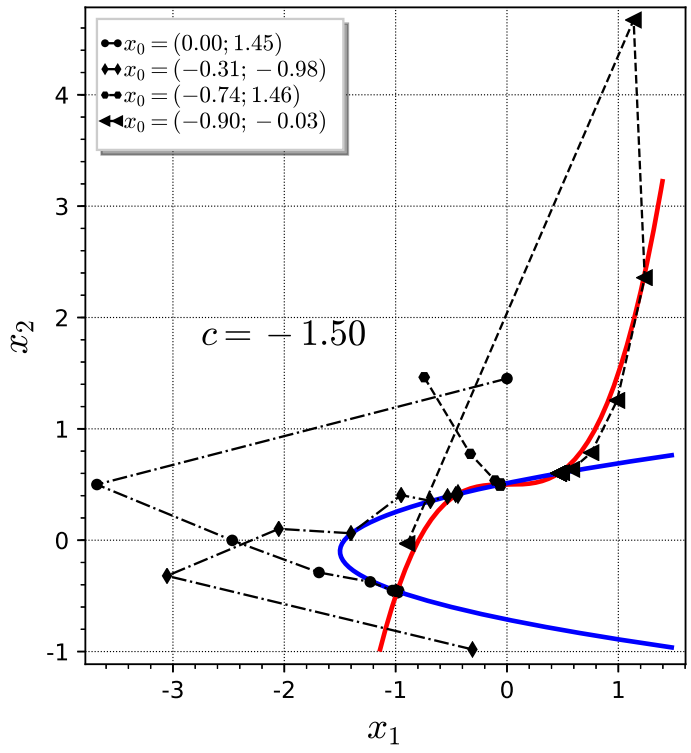


Figure 2.21: Représentation de 4 suites pour le système 2.33

En Figure 2.22a, on représente les bassins d'attraction pour les itérées de Newton associés au système 2.33 : à chaque point initial $x_0 = (x, y)$ on associe le point fixe vers lequel la suite de Newton converge et chaque point fixe correspond une couleur. En Figure 2.22b, on représente le nombre d'itérations assurant la convergence des itérées de Newton : à chaque point initial $x_0 = (x, y)$ on associe le nombre d'itérations nécessaire à la convergence et une échelle de couleur permet de visualer ces nombres.

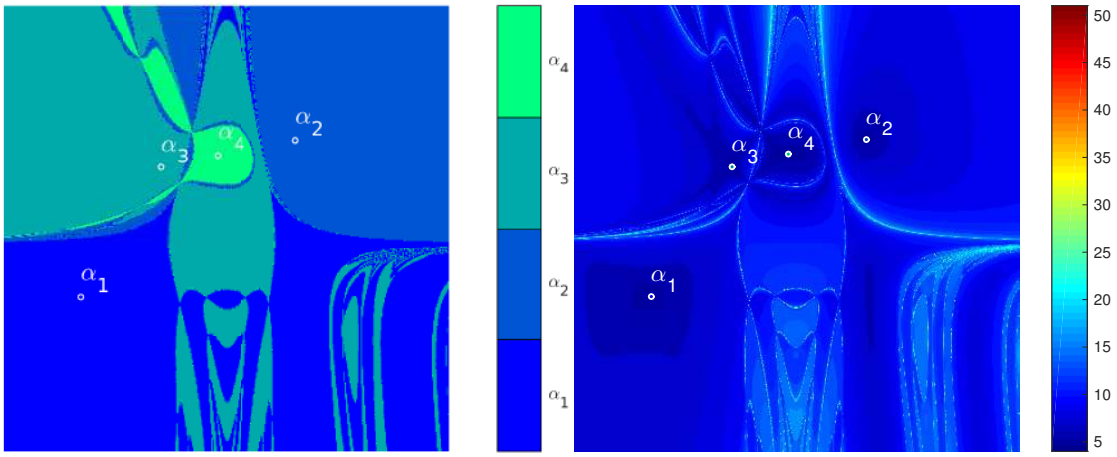
Exemple complexe : $z^3 - 1 = 0$

On souhaite trouver les racines complexes de $z^3 - 1$. Pour celà on peut poser $z = x + \imath y$, et le système équivalent devient

$$\begin{cases} f_1(x, y) = x^3 - 3xy^2 - 1 & = 0 \\ f_2(x, y) = 3x^2y - x^3 & = 0. \end{cases} \tag{2.34}$$

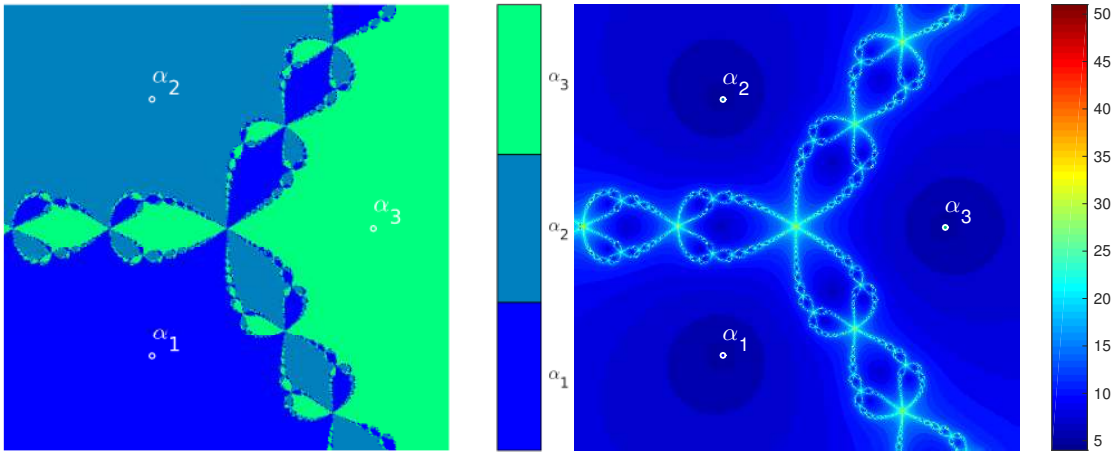
Bien évidemment en restant dans le corps des complexes, l'algorithme de Newton est le même (encore faut-il que le langage de programmation utilisé le supporte).

On représente en Figure 2.23 les bassins d'attraction et le nombre d'itérations de convergence associé à des données initiales dans $[-1.5, 1.5] \times [-1.5, 1.5]$. On obtient alors une *fractale de Newton*. Pour illustrer ce *caractère fractale* des représentations, on donne en Figures 2.24 et 2.25 des zooms successifs sur les graphes.



(a) Bassin d'attraction des racines (b) Nombre d'itérations de convergence

Figure 2.22: Méthode de Newton, système (2.33)



(a) Bassin d'attraction des racines (b) Nombre d'itérations de convergence

Figure 2.23: Méthode de Newton, système (2.34)

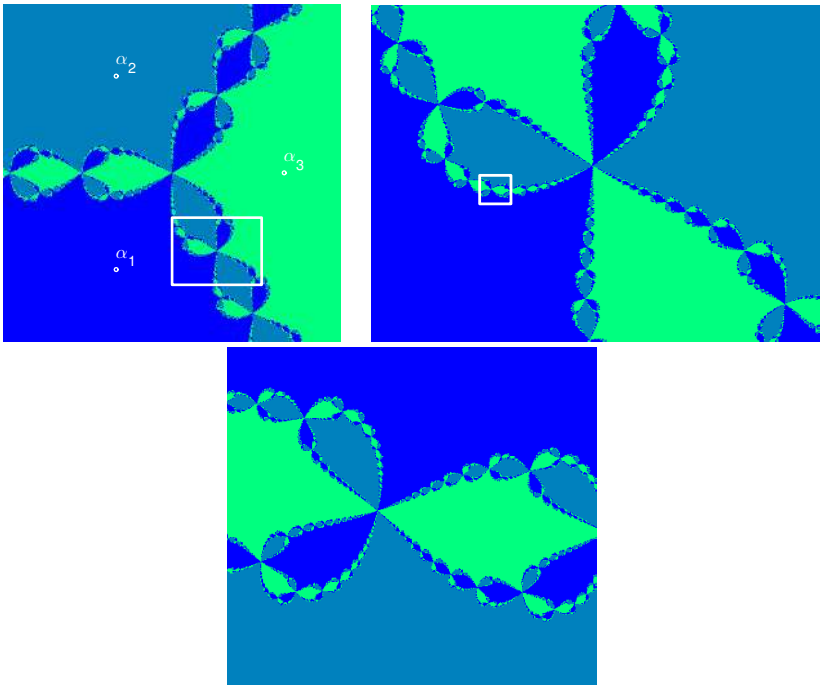


Figure 2.24: Méthode de Newton, système (2.34), zooms sur les bassins d'attraction

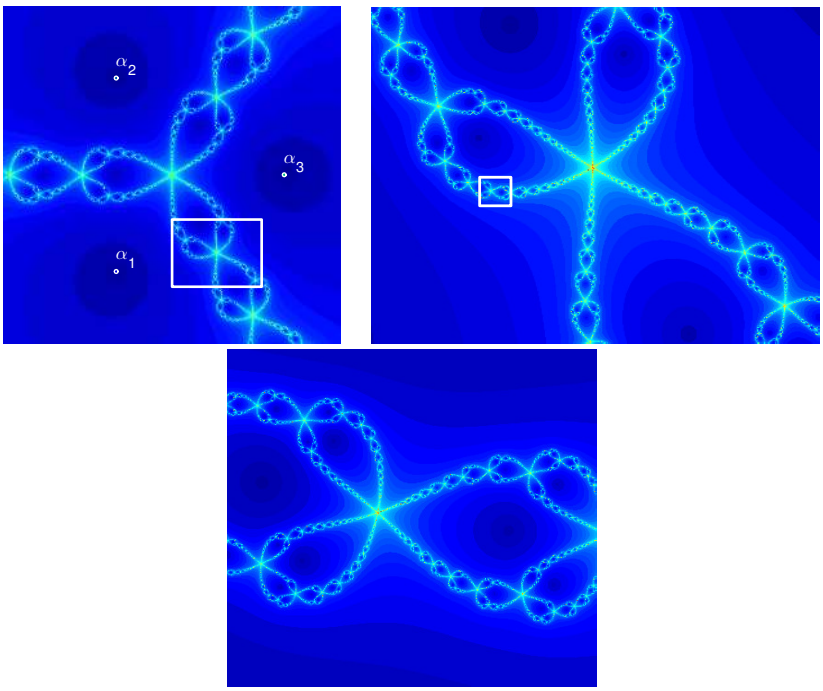
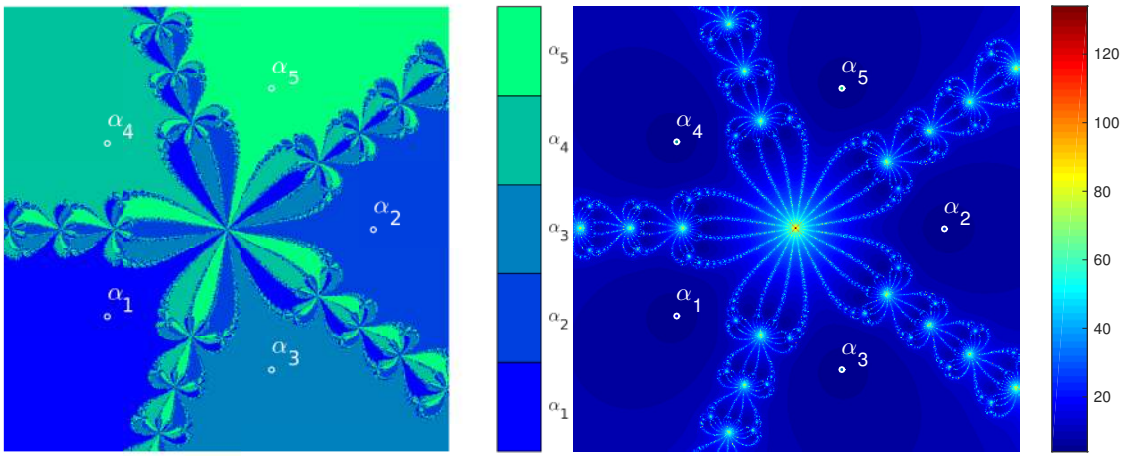


Figure 2.25: Méthode de Newton, système (2.34), zooms sur les nombres d'itérations

Exemple complexe : $z^5 - 1 = 0$

On représente en Figure 2.26 les bassins d'attraction et le nombre d'itérations de convergence associé à des données initiales dans $[-1.5, 1.5] \times [-1.5, 1.5]$.

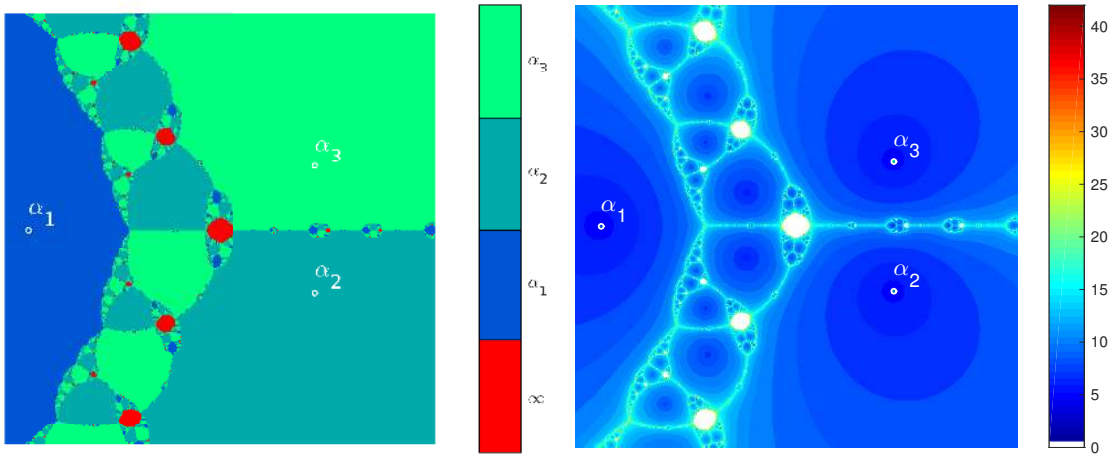


(a) Bassin d'attraction des racines (b) Nombre d'itérations de convergence

Figure 2.26: Méthode de Newton pour $z^5 - 1 = 0$

Exemple complexe : $z^3 - 2z + 2 = 0$

On représente en Figure 2.27 les bassins d'attraction et le nombre d'itérations de convergence associé à des données initiales dans $[-2, 2] \times [-2, 2]$.



(a) Bassin d'attraction des racines. En rouge zône de divergence (b) Nombre d'itérations de convergence. En blanc zône de divergence

Figure 2.27: Méthode de Newton pour $z^3 - 2z + 2 = 0$

Chapitre 3

Résolution de systèmes linéaires

Dans cette partie nous allons consisérer la résolution numérique d'un système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ dont la matrice $\mathbb{A}$ est inversible.

On pourrait penser que pour résoudre le système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$, $\mathbb{A}$ inversible, le plus simple serait de calculer la matrice $\mathbb{A}^{-1}$, inverse de $\mathbb{A}$, puis d'effectuer un produit matrice-vecteur pour obtenir $\boldsymbol{x} = \mathbb{A}^{-1}\boldsymbol{b}$. Or pour calculer l'inverse d'une matrice d'ordre n on doit résoudre n systèmes linéaires d'ordre n! En effet, déterminer l'inverse d'une matrice $\mathbb{A}$ revient à rechercher la matrice $\mathbb{X}$ solution de

$$\mathbb{A}\mathbb{X} = \mathbb{I} \iff \begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,n} \\ A_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{n-1,n} \\ A_{n,1} & \dots & A_{n,n-1} & A_{n,n} \end{pmatrix} \begin{pmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,n} \\ X_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & X_{n-1,n} \\ X_{n,1} & \dots & X_{n,n-1} & X_{n,n} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix}$$

Si on note $\boldsymbol{X}_{i,:}$ le i-ème vecteur colonne de la matrice $\mathbb{X}$ et $\boldsymbol{e}_i$ le i-ème de la base canonique de $\mathbb{R}^n$ alors le système précédant s'écrit

$$\begin{pmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,n} \\ A_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{n-1,n} \\ A_{n,1} & \dots & A_{n,n-1} & A_{n,n} \end{pmatrix} \left(\begin{array}{c|c|c|c} & & & \\ \boldsymbol{X}_1 & \dots & \boldsymbol{X}_n & \\ & & & \end{array} \right) = \left(\begin{array}{c|c|c|c} & & & \\ \boldsymbol{e}_1 & \dots & \boldsymbol{e}_n & \\ & & & \end{array} \right)$$

Ce dernier système est alors équivalent à résoudre les n systèmes linéaires

$$\mathbb{A}\boldsymbol{X}_j = \boldsymbol{e}_j, \quad \forall j \in [\![1, n]\!].$$

> ⚠ Pour résoudre un système linéaire, on ne calcule pas la matrice inverse associée.

Nous allons en section 3.1 étudier quelques **méthodes directes** pour la résolution d'un système linéaire basées sur la recherche d'une matrice $\mathbb{M}$ inversible telle quel la matrice $\mathbb{M}\mathbb{A}$ soit triangulaire supérieure. Ceci conduit à la résolution du système linéaire équivalent

$$\mathbb{M}\mathbb{A}\boldsymbol{x} = \mathbb{M}\boldsymbol{b}.$$

par la *méthode de la remontée* décrite en section 3.1.1).

En section 3.4, nous nous intéresserons aux **méthodes itératives** pour la résolution d'un système linéaire qui peuvent s'écrire sous la forme

$$\boldsymbol{x}^{[k+1]} = \mathbb{B}\boldsymbol{x}^{[k]} + \boldsymbol{c}, \quad k \geqslant 0, \ \boldsymbol{x}^{[0]} \text{ donné}$$

où la matrice $\mathbb{B}$ et le vecteur $\boldsymbol{c}$ sont construits à partir de la matrice $\mathbb{A}$ et du vecteur $\boldsymbol{b}$. On espère alors avoir $\lim_{k\to+\infty} \boldsymbol{x}^{[k]} = \boldsymbol{x}$.

3.1 Méthodes directes

Pour résoudre le système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$, nous allons le transformer en un système linéaire triangulaire supérieure équivalent

$$\mathbb{M}\mathbb{A}\boldsymbol{x} = \mathbb{M}\boldsymbol{b}$$

où $\mathbb{M}$ est une matrice inversible telle que $\mathbb{M}\mathbb{A}$ soit triangulaire supérieure. Nous allons voir que ce nouveau système est très facile à résoudre par la *méthode de la remontée*.

Nous étudierons la *méthode de Gauss-Jordan* que nous réécrions sous forme algébrique. Puis nous ferons le lien avec les méthodes utilisant la *factorisation* $\mathbb{L}\mathbb{U}$ et la *factorisation de Cholesky*. Nous finirons par une méthode utilisant la factorisation $\mathbb{Q}\mathbb{R}$ d'une matrice, factorisation qui sera réutilisée pour le calcul de valeurs propres et vecteurs propres en section **??**.

Nous allons tout d'abord regarder quelques cas particuliers : la matrice du système est diagonale, triangulaire inférieure ou triangulaire supérieure.

3.1.1 Matrices particulières

Matrices diagonales

Soit $\mathbb{A}$ une matrice de $\mathcal{M}_n(\mathbb{K})$ diagonale inversible et $\boldsymbol{b} \in \mathbb{K}^n$. Dans ce cas les coefficients diagonaux de $\mathbb{A}$ sont tous non nuls et l'on a

$$x_i = b_i/A_{i,i}, \quad \forall i \in [\![1, n]\!]. \tag{3.1}$$

On a immédiatement l'algorithme

Algorithme 3.1 Fonction RSLMatDiag permettant de résoudre le système linéaire à matrice diagonale inversible

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b}.$$

Données : $\mathbb{A}$: matrice diagonale de $\mathcal{M}_n(\mathbb{R})$ inversible.
$\boldsymbol{b}$: vecteur de $\mathbb{R}^n$.
Résultat : $\boldsymbol{x}$: vecteur de $\mathbb{R}^n$.

```
1: Fonction x ← RSLMatDiag ( A, b )
2:   Pour i ← 1 à n faire
3:     x(i) ← b(i)/A(i,i)
4:   Fin Pour
5: Fin Fonction
```

Matrices triangulaires inférieures

Soit $\mathbb{A}$ une matrice de $\mathcal{M}_n(\mathbb{K})$ triangulaire inférieure inversible et $\boldsymbol{b} \in \mathbb{K}^n$. On veut résoudre le système linéaire

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b} \iff \begin{pmatrix} A_{1,1} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ A_{n,1} & \dots & \dots & A_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ b_n \end{pmatrix}$$

Exercice 3.1.1

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$ une matrice **triangulaire**. Montrer que

$$\mathbb{A} \text{ inversible} \iff A_{i,i} \neq 0, \forall i \in [\![1,n]\!].$$

On remarque que l'on peut calculer successivement $x_1, x_1, \dots, x_n$, car il est possible de calculer x_i si on connait $x_1, \dots, x_{i-1}$: c'est la **méthode de descente**. En effet, on a

$$(\mathbb{A}\boldsymbol{x})_i = b_i, \ \forall i \in [\![1,n]\!].$$

et donc, par définition d'un produit matrice-vecteur,

$$\sum_{j=1}^{n} A_{i,j}x_j = b_i, \ \forall i \in [\![1,n]\!].$$

Comme $\mathbb{A}$ est une matrice triangulaire inférieure, on a (voir Définition B.44) $A_{i,j} = 0$ si $j > i$. Ceci donne alors pour tout $i \in [\![1,n]\!]$

$$\begin{aligned} b_i &= \sum_{j=1}^{i-1} A_{i,j}x_j + A_{i,i}x_i + \sum_{j=i+1}^{n} \underbrace{A_{i,j}}_{=0} x_j \\ &= \sum_{j=1}^{i-1} A_{i,j}x_j + A_{i,i}x_i \end{aligned}$$

De plus la matrice $\mathbb{A}$ étant triangulaire inversible ses éléments diagonaux sont tous non nuls. On obtient alors x_i en fonction des $x_1, \dots, x_{i-1}$:

$$x_i = \frac{1}{A_{i,i}} \left(b_i - \sum_{j=1}^{i-1} A_{i,j}x_j \right), \ \forall i \in [\![1,n]\!]. \tag{3.2}$$

On écrit en détail les raffinements successifs permettant d'aboutir à l'Algorithme 3.2 final ne comportant que des opérations élémentaires(.... à finaliser) de telle sorte que le passage entre deux raffinements successifs soit le plus compréhensible possible.

Algorithme 3.2 $\mathcal{R}_0$

1: Résoudre $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ en calculant successivement $x_1, \ x_2, \ \dots, x_n$.

Algorithme 3.2 $\mathcal{R}_1$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: $\quad x_i \leftarrow \frac{1}{A_{i,i}} \left(b_i - \sum_{j=1}^{i-1} A_{i,j}x_j \right)$
3: **Fin Pour**

Dans le raffinement $\mathcal{R}_1$, la seule difficulté restante est le calcul de la somme $\sum_{j=1}^{i-1} A_{i,j}x_j$. En effet, l'opérateur mathématique $\sum$ n'est pas défini dans notre langage algorithmique : il va donc falloir détailler un peu plus l'algorithme. Pour isoler le calcul de cette somme, on la note S. La ligne 2 peut donc s'écrire

$$x_i \leftarrow \frac{1}{A_{i,i}} (b_i - S).$$

Mais où calculer la valeur S? 4 choix possible : avant la ligne 1, entre les lignes 1 et 2, entre les lignes 2 et 3 ou après la ligne 3.
On ne peut pas calculer S après utilisation de sa valeur dans le calcul de x_i! ce qui élimine les 2 derniers choix. Ensuite, on ne peut sortir le calcul de S de la boucle puique la somme dépend de l'indice de boucle i : ce qui élimine le premier choix. On doit donc calculer S dans la boucle et avant le calcul de x_i.

Algorithme 3.2 $\mathcal{R}_1$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: $\quad x_i \leftarrow \dfrac{1}{A_{i,i}}\left(b_i - \sum_{j=1}^{i-1} A_{i,j}x_j\right)$
3: **Fin Pour**

Algorithme 3.2 $\mathcal{R}_2$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: $\quad S \leftarrow \sum_{j=1}^{i-1} A_{i,j}x_j$
3: $\quad x_i \leftarrow (b_i - S)/A_{i,i}$
4: **Fin Pour**

Maintenant que l'on a isolé la *difficulté*, il reste à détailler le calcul de S. Celui-ci se fait **intégralement** en lieu et place de la ligne 2.

Algorithme 3.2 $\mathcal{R}_2$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: $\quad S \leftarrow \sum_{j=1}^{i-1} A_{i,j}x_j$
3: $\quad x_i \leftarrow (b_i - S)/A_{i,i}$
4: **Fin Pour**

Algorithme 3.2 $\mathcal{R}_3$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: $\quad S \leftarrow 0$
3: $\quad$ **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
4: $\quad\quad S \leftarrow S + A(i,j) * x(j)$
5: $\quad$ **Fin Pour**
6: $\quad x_i \leftarrow (b_i - S)/A_{i,i}$
7: **Fin Pour**

Insister sur $S \leftarrow 0$ à l'intérieur de la boucle en i? (erreur trop courante des débutants)
On obtient alors l'algorithme final

Algorithme 3.2 Fonction RSLTriInf permettant de résoudre le système linéaire triangulaire inférieur inversible

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b}.$$

Données : $\mathbb{A}$: matrice triangulaire de $\mathcal{M}_n(\mathbb{K})$ inférieure inversible.
$\boldsymbol{b}$: vecteur de $\mathbb{K}^n$.
Résultat : $\boldsymbol{x}$: vecteur de $\mathbb{K}^n$.

1: **Fonction** $\boldsymbol{x} \leftarrow$ RSLTriInf ($\mathbb{A}, \boldsymbol{b}$)
2: $\quad$ **Pour** $i \leftarrow 1$ **à** n **faire**
3: $\quad\quad S \leftarrow 0$
4: $\quad\quad$ **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
5: $\quad\quad\quad S \leftarrow S + A(i,j) * x(j)$
6: $\quad\quad$ **Fin Pour**
7: $\quad\quad x(i) \leftarrow (b(i) - S)/A(i,i)$
8: $\quad$ **Fin Pour**
9: **Fin Fonction**

Matrices triangulaires supérieures

Soit $\mathbb{A}$ une matrice de $\mathcal{M}_n(\mathbb{R})$ triangulaire supérieure inversible et $\boldsymbol{b} \in \mathbb{R}^n$. On veut résoudre le système linéaire

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b} \iff \begin{pmatrix} A_{1,1} & \dots & \dots & A_{1,n} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & A_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ \vdots \\ b_n \end{pmatrix}$$

On remarque que l'on peut calculer successivement $x_n, x_{n-1}, \dots, x_1$, car il est possible de calculer x_i si on connait $x_{i+1}, \dots, x_n$: c'est la **méthode de remontée**. En effet, on a

$$(\mathbb{A}\boldsymbol{x})_i = b_i, \ \forall i \in [\![1, n]\!].$$

et donc, par définition d'un produit matrice-vecteur,

$$\sum_{j=1}^{n} A_{i,j}x_j = b_i, \ \forall i \in [\![1,n]\!].$$

Comme $\mathbb{A}$ est une matrice triangulaire supérieure, on a (voir Définition B.44) $A_{i,j} = 0$ si $j < i$. Ceci donne alors pour tout $i \in [\![1,n]\!]$

$$\begin{aligned} b_i &= \sum_{j=1}^{i-1} \underbrace{A_{i,j}}_{=0} x_j + A_{i,i}x_i + \sum_{j=i+1}^{n} A_{i,j}x_j \\ &= A_{i,i}x_i + \sum_{j=i+1}^{n} A_{i,j}x_j \end{aligned}$$

De plus la matrice $\mathbb{A}$ étant triangulaire inversible ses éléments diagonaux sont tous non nuls. On obtient donc x_i en fonction des $x_{i+1}, \dots, x_n$:

$$x_i = \frac{1}{A_{i,i}} \left(b_i - \sum_{j=i+1}^{n} A_{i,j}x_j \right), \ \forall i \in [\![1,n]\!]. \tag{3.3}$$

Algorithme 3.3 $\mathcal{R}_0$

1: Résoudre $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ en calculant successivement $x_n, \ x_{n-1}, \ \dots, x_1$.

Algorithme 3.3 $\mathcal{R}_1$

1: **Pour** $i \leftarrow n$ **à** 1 **faire**(**pas de** -1)
2: calculer x_i connaissant $x_{i+1}, \dots, x_n$
3: **Fin Pour**

Algorithme 3.3 $\mathcal{R}_1$

1: **Pour** $i \leftarrow n$ **à** 1 **faire**(**pas de** -1)
2: Calculer x_i connaissant $x_{i+1}, \dots, x_n$
3: **Fin Pour**

Algorithme 3.3 $\mathcal{R}_2$

1: **Pour** $i \leftarrow n$ **à** 1 **faire**(**pas de** -1)
2: $S \leftarrow \sum_{j=i+1}^{n} A_{i,j}x_j$
3: $x_i \leftarrow (b_i - S)/A_{i,i}$
4: **Fin Pour**

Algorithme 3.3 $\mathcal{R}_2$

1: **Pour** $i \leftarrow n$ **à** 1 **faire**(**pas de** -1)
2: $S \leftarrow \sum_{j=i+1}^{n} A_{i,j}x_j$
3: $x_i \leftarrow (b_i - S)/A_{i,i}$
4: **Fin Pour**

Algorithme 3.3 $\mathcal{R}_3$

1: **Pour** $i \leftarrow n$ **à** 1 **faire**(**pas de** -1)
2: $S \leftarrow 0$
3: **Pour** $j \leftarrow i+1$ **à** n **faire**
4: $S \leftarrow S + A(i,j) * x(j)$
5: **Fin Pour**
6: $x_i \leftarrow (b_i - S)/A_{i,i}$
7: **Fin Pour**

On obtient alors l'algorithme final

Algorithme 3.3 Fonction RSLTriSup permettant de résoudre le système linéaire triangulaire supérieur inversible

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b}.$$

Données : $\mathbb{A}$: matrice triangulaire de $\mathcal{M}_n(\mathbb{R})$ supérieure inversible.
$\boldsymbol{b}$: vecteur de $\mathbb{R}^n$.
Résultat : $\boldsymbol{x}$: vecteur de $\mathbb{R}^n$.

```
1: Fonction x ← RSLTriSup ( A, b )
2:   Pour i ← n à 1 faire(pas de −1)
3:     S ← 0
4:     Pour j ← i + 1 à n faire
5:       S ← S + A(i, j) ∗ x(j)
6:     Fin Pour
7:     x(i) ← (b(i) − S)/A(i, i)
8:   Fin Pour
9: Fin Fonction
```

3.1.2 Exercices et résultats préliminaires

Exercice 3.1.2: correction page 217

Soit $\mathbb{A} \in \mathcal{M}_{n,n}(\mathbb{C})$ une matrice et $(\lambda, \boldsymbol{u})$ un élément propre de $\mathbb{A}$ avec $\|\boldsymbol{u}\|_2 = 1$.

Q. 1 *En s'aidant de la base canonique* $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$, *construire une base orthonormée* $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ *telle que* $\boldsymbol{x}_1 = \boldsymbol{u}$.

Notons $\mathbb{P}$ la matrice de changement de base canonique $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ dans la base $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$:

$$\mathbb{P} = \left(\begin{array}{c:c:c} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{array} \right)$$

Soit $\mathbb{B}$ la matrice définie par $\mathbb{B} = \mathbb{P}^*\mathbb{A}\mathbb{P}$.

Q. 2 1. *Exprimer les coefficients de la matrice* $\mathbb{B}$ *en fonction de la matrice* $\mathbb{A}$ *et des vecteurs* $\boldsymbol{x}_i$, $i \in [\![1, n]\!]$.

$$\mathbb{B} = \mathbb{P}^*\mathbb{A}\mathbb{P}.$$

2. *En déduire que la première colonne de* $\mathbb{B}$ *est* $(\lambda, 0, \dots, 0)^t$.

Q. 3 *Montrer par récurrence sur l'ordre de la matrice que la matrice* $\mathbb{A}$ *s'écrit*

$$\mathbb{A} = \mathbb{U}\mathbb{T}\mathbb{U}^*$$

où $\mathbb{U}$ *est une matrice unitaire et* $\mathbb{T}$ *une matrice triangulaire supérieure.*

Q. 4 *En supposant* $\mathbb{A}$ *inversible et la décomposition* $\mathbb{A} = \mathbb{U}\mathbb{T}\mathbb{U}^*$ *connue, expliquer comment résoudre "simplement" le système linéaire* $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$.

Correction Exercice 3.1.2

Q. 1 La première chose à faire est de construire une base contenant $\boldsymbol{u}$ à partir de la base canonique $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$. Comme le vecteur propre $\boldsymbol{u}$ est non nul, il existe $j \in [\![1, n]\!]$ tel que $\langle \boldsymbol{u}, \boldsymbol{e}_j \rangle \neq 0$. La famille $\{\boldsymbol{u}, \boldsymbol{e}_1, \dots, \boldsymbol{e}_{j-1}, \boldsymbol{e}_{j+1}, \dots, \boldsymbol{e}_n\}$ forme alors une base de $\mathbb{C}^n$ car $\boldsymbol{u}$ n'est pas combinaison linéaire des $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_{j-1}, \boldsymbol{e}_{j+1}, \dots, \boldsymbol{e}_n\}$.
On note $\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\}$ la base dont le premier élément est $\boldsymbol{z}_1 = \boldsymbol{u}$:

$$\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\} = \{\boldsymbol{u}, \boldsymbol{e}_1, \dots, \boldsymbol{e}_{j-1}, \boldsymbol{e}_{j+1}, \dots, \boldsymbol{e}_n\}.$$

On peut ensuite utiliser le **procédé de Gram-Schmidt**, rappelé en Proposition B.19, pour construire une base orthonormée à partir de cette base.

On calcule successivement les vecteurs $\boldsymbol{x}_i$ à partir de la base $\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\}$ en construisant un vecteur $\boldsymbol{w}_i$ orthogonal aux vecteurs $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{i-1}$.

$$\boldsymbol{w}_i = \boldsymbol{z}_i - \sum_{k=1}^{i-1} \langle \boldsymbol{x}_k, \boldsymbol{z}_i \rangle \boldsymbol{x}_k$$

puis on obtient le vecteur $\boldsymbol{x}_i$ en normalisant

$$\boldsymbol{x}_i = \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}$$

Q. 2 1. En conservant l'écriture colonne de la matrice $\mathbb{P}$ on obtient

$$\mathbb{B} = \left(\begin{array}{c} \boldsymbol{x}_1^* \\ \hdashline \boldsymbol{x}_2^* \\ \hdashline \vdots \\ \hdashline \boldsymbol{x}_n^* \end{array}\right) \mathbb{A} \left(\begin{array}{c:c:c:c} \boldsymbol{x_1} & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n \end{array}\right) = \left(\begin{array}{c} \boldsymbol{x}_1^* \\ \hdashline \boldsymbol{x}_2^* \\ \hdashline \vdots \\ \hdashline \boldsymbol{x}_n^* \end{array}\right) \left(\begin{array}{c:c:c:c} \mathbb{A}\boldsymbol{x_1} & \mathbb{A}\boldsymbol{x}_2 & \dots & \mathbb{A}\boldsymbol{x}_n \end{array}\right)$$

Ce qui donne

$$\mathbb{B} = \begin{pmatrix} \boldsymbol{x}_1^*\mathbb{A}\boldsymbol{x_1} & \boldsymbol{x_1}^*\mathbb{A}\boldsymbol{x}_2 & \dots & \boldsymbol{x_1}^*\mathbb{A}\boldsymbol{x}_n \\ \boldsymbol{x}_2^*\mathbb{A}\boldsymbol{x_1} & \boldsymbol{x}_2^*\mathbb{A}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_2^*\mathbb{A}\boldsymbol{x}_n \\ \vdots & \vdots & & \vdots \\ \boldsymbol{x}_n^*\mathbb{A}\boldsymbol{x_1} & \boldsymbol{x}_n^*\mathbb{A}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_n^*\mathbb{A}\boldsymbol{x}_n \end{pmatrix}$$

On a donc

$$B_{i,j} = \boldsymbol{x}_i^*\mathbb{A}\boldsymbol{x_j}, \quad \forall (i,j) \in [\![1,n]\!]^2$$

.

2. On a $\mathbb{A}\boldsymbol{u} = \lambda\boldsymbol{u}$, $\|\boldsymbol{u}\| = 1$, la base $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ est orthonormée et $\boldsymbol{x}_1 = \boldsymbol{u}$. on obtient alors

$$\mathbb{B} = \left(\begin{array}{c:c:c:c} \lambda\boldsymbol{u}^*\boldsymbol{u} & \boldsymbol{u}^*\mathbb{A}\boldsymbol{x}_2 & \dots & \boldsymbol{u}^*\mathbb{A}\boldsymbol{x}_n \\ \hdashline \lambda\boldsymbol{x}_2^*\boldsymbol{u} & \boldsymbol{x}_2^*\mathbb{A}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_2^*\mathbb{A}\boldsymbol{x}_n \\ \hdashline \vdots & \vdots & & \vdots \\ \hdashline \lambda\boldsymbol{x}_n^*\boldsymbol{u} & \boldsymbol{x}_n^*\mathbb{A}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_n^*\mathbb{A}\boldsymbol{x}_n \end{array}\right) = \left(\begin{array}{c:c:c:c} \lambda & \boldsymbol{u}^*\mathbb{A}\boldsymbol{x}_2 & \dots & \boldsymbol{u}^*\mathbb{A}\boldsymbol{x}_n \\ \hdashline 0 & \boldsymbol{x}_2^*\mathbb{A}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_2^*\mathbb{A}\boldsymbol{x}_n \\ \hdashline \vdots & \vdots & & \vdots \\ \hdashline 0 & \boldsymbol{x}_n^*\mathbb{A}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_n^*\mathbb{A}\boldsymbol{x}_n \end{array}\right)$$

Q. 3 On veut démontrer, par récurrence faible, la proposition suivante pour $n \geqslant 2$

$(\mathcal{P}_n)$ $\forall \mathbb{A} \in \mathcal{M}_n(\mathbb{C})$, $\exists \mathbb{U} \in \mathcal{M}_n(\mathbb{C})$ unitaire, $\exists \mathbb{T} \in \mathcal{M}_n(\mathbb{C})$ triangulaire supérieure, telles que $\mathbb{A} = \mathbb{U}\mathbb{T}\mathbb{U}^*$.

Initialisation : Montrons que $(\mathcal{P}_2)$ est vérifié.
Soit $\mathbb{A}_2 \in \mathcal{M}_2(\mathbb{C})$. Elle admet au moins un élément propre $(\lambda, \boldsymbol{u})$ (voir Proposition B.40 par ex.) avec $\|\boldsymbol{u}\| = 1$. On peut donc appliquer le résultat de la question précédente : il existe une matrice unitaire $\mathbb{P}_2 \in \mathcal{M}_2(\mathbb{C})$ telle que la matrice $\mathbb{B}_2 = \mathbb{P}_2\mathbb{A}_2\mathbb{P}_2^*$ ait comme premier vecteur colonne $(\lambda, 0)^t$. La matrice $\mathbb{B}_2$ est donc triangulaire supérieure et comme $\mathbb{P}_2$ est unitaire on en déduit

$$\mathbb{A}_2 = \mathbb{P}_2^*\mathbb{B}_2\mathbb{P}_2.$$

On pose $\mathbb{U}_2 = \mathbb{P}_2^*$ matrice unitaire et $\mathbb{T}_2 = \mathbb{B}_2$ matrice triangulaire supérieure pour conclure que la propostion $(\mathcal{P}_2)$ est vraie.

Hérédité : Supposons que $(\mathcal{P}_{n-1})$ soit vérifiée. Montrons que $(\mathcal{P}_n)$ est vraie.
Soit $\mathbb{A}_n \in \mathcal{M}_n(\mathbb{C})$. Elle admet au moins un élément propre $(\lambda, \boldsymbol{u})$ (voir Proposition B.40 par ex.) avec $\|\boldsymbol{u}\| = 1$. On peut donc appliquer le résultat de la question précédente : il existe une matrice unitaire $\mathbb{P}_n \in \mathcal{M}_n(\mathbb{C})$ telle que la matrice $\mathbb{B}_n = \mathbb{P}_n\mathbb{A}_n\mathbb{P}_n^*$ s'écrivent

$$\mathbb{B}_n = \left(\begin{array}{c:c} \lambda & \boldsymbol{c}_{n-1}^* \\ \hdashline 0 & \\ \vdots & \mathbb{A}_{n-1} \\ 0 & \end{array}\right)$$

où $\boldsymbol{c}_{n-1} \in \mathcal{M}_{n-1,1}(\mathbb{C})$ et $\mathbb{A}_{n-1} \in \mathcal{M}_{n-1}(\mathbb{C})$. Par hypothèse de récurrence, $\exists \mathbb{U}_{n-1} \in \mathcal{M}_{n-1}(\mathbb{C})$ unitaire et $\mathbb{T}_{n-1} \in \mathcal{M}_{n-1}(\mathbb{C})$ triangulaire supérieure telles que

$$\mathbb{A}_{n-1} = \mathbb{U}_{n-1}\mathbb{T}_{n-1}\mathbb{U}_{n-1}^*$$

ou encore

$$\mathbb{T}_{n-1} = \mathbb{U}_{n-1}^*\mathbb{A}_{n-1}\mathbb{U}_{n-1}.$$

Soit $\mathbb{Q}_n \in \mathcal{M}_n(\mathbb{C})$ la matrice définie par

$$\mathbb{Q}_n = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1} & \\ 0 & & & \end{array}\right).$$

La matrice $\mathbb{Q}_n$ est unitaire. En effet on a

$$\mathbb{Q}_n\mathbb{Q}_n^* = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1} & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1}^* & \\ 0 & & & \end{array}\right) = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \underbrace{\mathbb{U}_{n-1}\mathbb{U}_{n-1}^*}_{=\mathbb{I}_{n-1}} & \\ 0 & & & \end{array}\right) = \mathbb{I}_n.$$

On note $\mathbb{T}_n$ la matrice définie par $\mathbb{T}_n = \mathbb{Q}_n^*\mathbb{B}_n\mathbb{Q}_n$. On a alors

$$\begin{aligned}\mathbb{T}_n &= \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1}^* & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} \lambda & \boldsymbol{c}_{n-1}^* & & \\ \hline 0 & & & \\ \vdots & & \mathbb{A}_{n-1} & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1} & \\ 0 & & & \end{array}\right) \\ &= \left(\begin{array}{c|ccc} \lambda & \boldsymbol{c}_{n-1}^* & & \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1}^*\mathbb{A}_{n-1} & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1} & \\ 0 & & & \end{array}\right) = \left(\begin{array}{c|ccc} \lambda & \boldsymbol{c}_{n-1}^*\mathbb{U}_{n-1}^* & & \\ \hline 0 & & & \\ \vdots & & \underbrace{\mathbb{U}_{n-1}^*\mathbb{A}_{n-1}\mathbb{U}_{n-1}}_{=\mathbb{T}_{n-1}} = & \\ 0 & & & \end{array}\right)\end{aligned}$$

La matrice $\mathbb{T}_n$ est donc triangulaire supérieure et on a par définition de $\mathbb{B}_n$

$$\mathbb{T}_n = \mathbb{Q}_n^*\mathbb{P}_n\mathbb{A}_n\mathbb{P}_n^*\mathbb{Q}_n.$$

On note $\mathbb{U}_n = \mathbb{P}_n^*\mathbb{Q}_n$. Cette matrice est unitaire car les matrices $\mathbb{Q}_n$ et $\mathbb{P}_n$ le sont. En effet, on a

$$\mathbb{U}_n\mathbb{U}_n^* = \mathbb{P}_n^*\mathbb{Q}_n(\mathbb{P}_n^*\mathbb{Q}_n)^* = \mathbb{P}_n^*\underbrace{\mathbb{Q}_n\mathbb{Q}_n^*}_{=\mathbb{I}_n}\mathbb{P}_n = \mathbb{P}_n^*\mathbb{P}_n = \mathbb{I}_n.$$

On a $\mathbb{T}_n = \mathbb{U}_n^*\mathbb{A}_n\mathbb{U}_n$ et en multipliant cette équation à gauche par $\mathbb{U}_n$ et à droite par $\mathbb{U}_n^*$ on obtient l'équation équivalente $\mathbb{A}_n = \mathbb{U}_n\mathbb{T}_n\mathbb{U}_n^*$. La propriété $(\mathcal{P}_n)$ est donc vérifiée. Ce qui achève la démonstration.

Q. 4 Résoudre $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ est équivalent à résoudre

$$\mathbb{U}\mathbb{T}\mathbb{U}^*\boldsymbol{x} = \boldsymbol{b}. \tag{3.4}$$

Comme $\mathbb{U}$ est unitaire, on a $\mathbb{U}\mathbb{U}^* = \mathbb{I}$ et $\mathbb{U}^*$ inversible. Donc en multipliant (B.55) par $\mathbb{U}^*$ on obtient le système équivalent

$$\underbrace{\mathbb{U}^*\mathbb{U}}_{=\mathbb{I}}\mathbb{T}\mathbb{U}^*\boldsymbol{x} = \mathbb{U}^*\boldsymbol{b} \iff \mathbb{T}\mathbb{U}^*\boldsymbol{x} = \mathbb{U}^*\boldsymbol{b}.$$

On pose $\boldsymbol{y} = \mathbb{U}^*\boldsymbol{x}$. Le système précédant se résoud en deux étapes

1. on cherche $\boldsymbol{y}$ solution de $\mathbb{T}\boldsymbol{y} = \mathbb{U}^*\boldsymbol{b}$. Comme $\mathbb{U}$ est unitaire on a $\det(\mathbb{U})\det(\mathbb{U}^*) = \det(\mathbb{I}) = 1$ et donc

$$\begin{aligned}\det(\mathbb{A}) &= \det(\mathbb{U}\mathbb{T}\mathbb{U}^*) = \det(\mathbb{U})\det(\mathbb{T})\det(\mathbb{U}^*)\\ &= \det(\mathbb{T})\end{aligned}$$

   Or $\mathbb{A}$ inversible équivalent à $\det(\mathbb{A}) \neq 0$ et donc la matrice $\mathbb{T}$ est inversible. La matrice $\mathbb{T}$ étant triangulaire supérieure on peut résoudre facilement le système par la *méthode de remontée*.

2. une fois $\boldsymbol{y}$ déterminé, on résoud $\mathbb{U}^*\boldsymbol{x} = \boldsymbol{y}$. Comme $\mathbb{U}$ est unitaire, on obtient directement $\boldsymbol{x} = \mathbb{U}\boldsymbol{y}$.

◇

On tire de cet exercice le théorème suivant

Théorème 3.1: Décomposition de Schur ★★★★☆

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$. Il existe une matrice unitaire $\mathbb{U}$ et une matrice triangulaire supérieure $\mathbb{T}$ telles que

$$\mathbb{A} = \mathbb{U}\mathbb{T}\mathbb{U}^* \tag{3.5}$$

Preuve. voir Exercice 3.1.2 □

Théorème 3.2: Réduction de matrices ★★★★★

1. Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$. Il existe une matrice **unitaire** $\mathbb{U}$ telle que $\mathbb{U}^{-1}\mathbb{A}\mathbb{U}$ soit **triangulaire**.
2. Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice **normale**. Il existe une matrice **unitaire** $\mathbb{U}$ telle que $\mathbb{U}^{-1}\mathbb{A}\mathbb{U}$ soit **diagonale**.
3. Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{R})$ une matrice **symétrique**. Il existe une matrice **orthogonale** $\mathbb{P}$ telle que $\mathbb{P}^{-1}\mathbb{A}\mathbb{P}$ soit **diagonale**.

Preuve. voir [1] Théorème 1.2-1 page 9 □

Exercice 3.1.3: Matrice d'élimination

Soit $\boldsymbol{v} \in \mathbb{C}^n$ avec $v_1 \neq 0$. On note $\mathbb{E}^{[\boldsymbol{v}]} \in \mathcal{M}_n(\mathbb{C})$ la matrice triangulaire inférieure à diagonale unité définie par

$$\mathbb{E}^{[\boldsymbol{v}]} = \left(\begin{array}{c|cccc} 1 & 0 & \dots & \dots & 0 \\ \hline -v_2/v_1 & 1 & 0 & \dots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -v_n/v_1 & 0 & \dots & 0 & 1 \end{array}\right) \tag{3.6}$$

Q. 1 *1. Calculer le déterminant de* $\mathbb{E}^{[\boldsymbol{v}]}$.

2. Déterminer l'inverse de $\mathbb{E}^{[\boldsymbol{v}]}$.

$\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ avec $A_{1,1} \neq 0$. On note $\boldsymbol{A}_{:,j}$ le j-ème vecteur colonne de $\mathbb{A}$ et $\boldsymbol{A}_{i,:}$ son i-ème vecteur ligne. On pose $\boldsymbol{A}_1 = \boldsymbol{A}_{:,1}$.

Q. 2 *1. Calculer* $\tilde{\mathbb{A}} = \mathbb{E}^{[\boldsymbol{A}_1]}\mathbb{A}$ *en fonction des vecteurs lignes de* $\mathbb{A}$.

2. Montrer que la première colonne de $\tilde{\mathbb{A}}$ *est le vecteur* $(A_{1,1}, 0, \dots, 0)^t$ *i.e.*

$$\mathbb{E}^{[\boldsymbol{A}_1]}\mathbb{A}\boldsymbol{e}_1 = A_{1,1}\boldsymbol{e}_1 \tag{3.7}$$

où $\boldsymbol{e}_1$ *est le premier vecteur de la base canonique de* $\mathbb{C}^n$.

Soit $m \in \mathbb{N}^*$. On note $\mathbb{E}^{[m,\boldsymbol{v}]} \in \mathcal{M}_{m+n}(\mathbb{C})$ la matrice triangulaire inférieure à diagonale unité définie

par

$$\mathbb{E}^{[m,\boldsymbol{v}]} = \left(\begin{array}{c|c} \mathbb{I}_m & \mathbb{O} \\ \hline \mathbb{O} & \mathbb{E}^{[\boldsymbol{v}]} \end{array}\right) \tag{3.8}$$

Q. 3 *1. Calculer le déterminant de* $\mathbb{E}^{[m,\boldsymbol{v}]}$.

2. Déterminer l'inverse de $\mathbb{E}^{[m,\boldsymbol{v}]}$ *en fonction de l'inverse de* $\mathbb{E}^{[\boldsymbol{v}]}$.

Soit $\mathbb{C}$ la matrice bloc définie par

$$\mathbb{C} = \left(\begin{array}{c|c} \mathbb{C}_{1,1} & \mathbb{C}_{1,2} \\ \hline \mathbb{O} & \mathbb{A} \end{array}\right)$$

où $\mathbb{C}_{1,1} \in \mathcal{M}_m(\mathbb{C})$ et $\mathbb{C}_{1,2} \in \mathcal{M}_{m,n}(\mathbb{C})$.

Q. 4 *Déterminer la matrice produit* $\mathbb{E}^{[m,\boldsymbol{A}_1]}\mathbb{C}$ *en fonction des matrices* $\mathbb{C}_{1,1}$, $\mathbb{C}_{1,2}$ *et* $\tilde{\mathbb{A}}$.

Correction Exercice 3.1.3

Q. 1 1. La matrice $\mathbb{E}^{[\boldsymbol{v}]}$ est triangulaire : son déterminant est donc le produit de ses éléments diagonaux (voir Proposition B.53, page 209). On a alors $\det(\mathbb{E}^{[\boldsymbol{v}]}) = 1$.

2. Pour calculer son inverse qui existe puisque $\det(\mathbb{E}^{[\boldsymbol{v}]}) \neq 0$, on écrit $\mathbb{E}^{[\boldsymbol{v}]}$ sous forme bloc :

$$\mathbb{E}^{[\boldsymbol{v}]} = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline \boldsymbol{e} & & \mathbb{I}_{n-1} & \end{array}\right)$$

avec $\boldsymbol{e} = (-v_2/v_1, \dots, -v_n/v_1)^{\mathsf{t}} \in \mathbb{C}^{n-1}$ On note $\mathbb{X} \in \mathcal{M}_n(\mathbb{C})$ son inverse qui s'écrit avec la même structure bloc

$$\mathbb{X} = \left(\begin{array}{c|c} a & \boldsymbol{b}^* \\ \hline \boldsymbol{c} & \mathbb{D} \end{array}\right)$$

avec $a \in \mathbb{K}$, $\boldsymbol{b} \in \mathbb{K}^{n-1}$, $\boldsymbol{c} \in \mathbb{K}^{n-1}$ et $\mathbb{D} \in \mathcal{M}_{n-1}(\mathbb{C})$.
La matrice $\mathbb{X}$ est donc solution de $\mathbb{E}^{[\boldsymbol{v}]}\mathbb{X} = \mathbb{I}$. Grace à l'écriture bloc des matrices on en déduit rapidement la matrice $\mathbb{X}$. En effet, en utilisant les produits blocs des matrices, on obtient

$$\mathbb{E}^{[\boldsymbol{v}]}\mathbb{X} = \left(\begin{array}{c|c} 1 & \mathbf{0}_{n-1}^{\mathsf{t}} \\ \hline \boldsymbol{e} & \mathbb{I}_{n-1} \end{array}\right)\left(\begin{array}{c|c} a & \boldsymbol{b}^* \\ \hline \boldsymbol{c} & \mathbb{D} \end{array}\right) = \left(\begin{array}{c|c} 1 \times a & 1 \times \boldsymbol{b}^* + \mathbf{0}_{n-1}^{\mathsf{t}} \times \mathbb{D} \\ \hline \boldsymbol{e} \times a + \mathbb{I}_{n-1} \times \boldsymbol{c} & \boldsymbol{e} \times \boldsymbol{b}^* + \mathbb{I}_{n-1} \times \mathbb{D} \end{array}\right)$$

$$= \left(\begin{array}{c|c} a & \boldsymbol{b}^* \\ \hline a\boldsymbol{e} + \boldsymbol{c} & \boldsymbol{e}\boldsymbol{b}^* + \mathbb{D} \end{array}\right)$$

Comme $\mathbb{X}$ est l'inverse de $\mathbb{E}^{[\boldsymbol{v}]}$, on a $\mathbb{E}^{[\boldsymbol{v}]}\mathbb{X} = \mathbb{I}$ et donc en écriture bloc

$$\left(\begin{array}{c|c} a & \boldsymbol{b}^* \\ \hline a\boldsymbol{e} + \boldsymbol{c} & \boldsymbol{e}\boldsymbol{b}^* + \mathbb{D} \end{array}\right) = \left(\begin{array}{c|c} 1 & \mathbf{0}_{n-1}^{\mathsf{t}} \\ \hline \mathbf{0}_{n-1} & \mathbb{I}_{n-1} \end{array}\right).$$

Ceci revient à résoudre les 4 équations

$$a = 1, \quad \boldsymbol{b}^* = \mathbf{0}_{n-1}^{\mathsf{t}}, \quad a\boldsymbol{e} + \boldsymbol{c} = \mathbf{0}_{n-1} \quad \text{et} \quad \boldsymbol{e}\boldsymbol{b}^* + \mathbb{D} = \mathbb{I}_{n-1}$$

qui donnent immédiatement $a = 1$, $\boldsymbol{b} = \mathbf{0}_{n-1}$, $\boldsymbol{c} = -\boldsymbol{e}$ et $\mathbb{D} = \mathbb{I}_{n-1}$. On obtient le résultat suivant

$$\left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline -\boldsymbol{e} & & \mathbb{I}_{n-1} & \end{array}\right)\left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline \boldsymbol{e} & & \mathbb{I}_{n-1} & \end{array}\right) = \mathbb{I}_n.$$

Il aurait été plus rapide d'utiliser la Proposition B.54, page 209.

Q. 2 1. Pour simplifier les notations, on note $\mathbb{E} = \mathbb{E}^{[\boldsymbol{A}_1]}$. Par définition du produit de deux matrices on a

$$\tilde{A}_{i,j} = \sum_{k=1}^{n} E_{i,k} A_{k,j}, \quad \forall (i,j) \in [\![1,n]\!]^2.$$

Quand $i = 1$, on a par construction $E_{1,k} = \delta_{1,k}$ et donc

$$\tilde{A}_{1,j} = A_{1,j}, \ \forall j \in [\![1,n]\!] \iff \tilde{\boldsymbol{A}}_{1,:} = \boldsymbol{A}_{1,:}. \tag{3.9}$$

Pour $i \geqslant 2$, on a $E_{i,1} = -\frac{v_i}{v_1}$ et $E_{i,k} = \delta_{i,k}$, $\forall k \in [\![2,n]\!]$. On obtient alors pour tout $j \in [\![1,n]\!]$

$$\tilde{A}_{i,j} = E_{i,1}A_{1,j} + \sum_{k=2}^{n} E_{i,k}A_{k,j} = -\frac{v_i}{v_1}A_{1,j} + \sum_{k=2}^{n} \delta_{i,k}A_{k,j} = -\frac{v_i}{v_1}A_{1,j} + A_{i,j}$$

ce qui donne pour tout $i \in [\![2,n]\!]$

$$\tilde{A}_{i,j} = A_{i,j} - \frac{v_i}{v_1}A_{1,j}, \ \forall j \in [\![1,n]\!] \iff \tilde{\boldsymbol{A}}_{i,:} = -\frac{v_i}{v_1}\boldsymbol{A}_{1,:} + \boldsymbol{A}_{i,:} \tag{3.10}$$

En conclusion, la matrice $\tilde{\mathbb{A}}$ s'écrit

$$\tilde{\mathbb{A}} = \begin{pmatrix} \boldsymbol{A}_{1,:} \\ \hdashline \boldsymbol{A}_{2,:} - (v_2/v_1)\boldsymbol{A}_{1,:} \\ \hdashline \vdots \\ \hdashline \boldsymbol{A}_{n,:} - (v_n/v_1)\boldsymbol{A}_{1,:} \end{pmatrix}$$

2. De (3.9), on tire $\tilde{A}_{1,1} = A_{1,1}$. A partir de (3.10) on obtient pour tout $i \in [\![2,n]\!]$, $\tilde{A}_{i,1} = A_{i,1} - \frac{v_i}{v_1}A_{1,1}$. Par construction $v_j = A_{j,1}$ pour tout $j \in [\![1,n]\!]$, ce qui donne $\tilde{A}_{i,1} = 0$.
La première colonne de $\tilde{\mathbb{A}}$ est $(1, 0, \ldots, 0)^{\mathsf{t}}$.

Q. 3 1. La matrice $\mathbb{E}^{[m,\boldsymbol{v}]}$ est triangulaire inférieure. Son déterminant est donc le produit de ses éléments diagonaux. Comme cette matrice est à diagonale unité (i.e. tous ses éléments diagonaux valent 1), on obtient $\det \mathbb{E}^{[m,\boldsymbol{v}]} = 1$.

Une autre manière de le démontrer. On peut voir que la matrice $\mathbb{E}^{[m,\boldsymbol{v}]}$ est bloc-diagonale. D'après la Proposition B.54, page 209, son déterminant est le produit des déterminant des blocs diagonaux : $\det \mathbb{E}^{[m,\boldsymbol{v}]} = \det \mathbb{I}_m \times \det \mathbb{E}^{[\boldsymbol{v}]} = 1$.

2. On note $\mathbb{X}$ l'inverse de la matrice $\mathbb{E}^{[m,\boldsymbol{v}]}$. Cette matrice s'écrit avec la même structure bloc

$$\mathbb{X} = \left(\begin{array}{c:c} \mathbb{X}_{1,1} & \mathbb{X}_{1,2} \\ \hdashline \mathbb{X}_{2,1} & \mathbb{X}_{2,2} \end{array}\right) \text{ avec } \mathbb{X}_{1,1} \in \mathcal{M}_m(\mathbb{C}) \text{ et } \mathbb{X}_{2,2} \in \mathcal{M}_n(\mathbb{C})$$

On a donc $\mathbb{X}\mathbb{E}^{[m,\boldsymbol{v}]} = \mathbb{I}_{m+n}$ c'est à dire en écriture bloc

$$\left(\begin{array}{c:c} \mathbb{X}_{1,1} & \mathbb{X}_{1,2} \\ \hdashline \mathbb{X}_{2,1} & \mathbb{X}_{2,2} \end{array}\right)\left(\begin{array}{c:c} \mathbb{I}_m & \mathbb{O} \\ \hdashline \mathbb{O} & \mathbb{E}^{[\boldsymbol{v}]} \end{array}\right) = \left(\begin{array}{c:c} \mathbb{I}_m & \mathbb{O} \\ \hdashline \mathbb{O} & \mathbb{I}_n \end{array}\right) =$$

On doit donc résoudre les 4 équations suivantes :

$$\mathbb{X}_{1,1}\mathbb{I}_m = \mathbb{I}_m, \quad \mathbb{X}_{1,2}\mathbb{I}_n = \mathbb{O}, \quad \mathbb{X}_{2,1}\mathbb{I}_m = \mathbb{O} \quad \text{et} \quad \mathbb{X}_{2,2}\mathbb{E}^{[\boldsymbol{v}]} = \mathbb{I}_n.$$

Comme la matrice $\mathbb{E}^{[\boldsymbol{v}]}$ est inversible, on obtient

$$\mathbb{X} = \left(\begin{array}{c:c} \mathbb{I}_m & \mathbb{O} \\ \hdashline \mathbb{O} & \left(\mathbb{E}^{[\boldsymbol{v}]}\right)^{-1} \end{array}\right)$$

Plus rapidement, comme la matrice $\mathbb{E}^{[m,\boldsymbol{v}]}$ est bloc-diagonale, on en déduit (voir Proposition B.54, page 209) directement le résultat.

Q. 4 Le produit $\mathbb{E}^{[m,\boldsymbol{v}]}\mathbb{C}$ peut s'effectuer par bloc car les blocs sont de dimensions compatibles et on a

$$\mathbb{E}^{[m,\boldsymbol{v}]}\mathbb{C} = \left(\begin{array}{c|c} \mathbb{I}_m & \mathbb{O}_{m,n} \\ \hline \mathbb{O}_{n,m} & \mathbb{E} \end{array}\right)\left(\begin{array}{c|c} \mathbb{C}_{1,1} & \mathbb{C}_{1,2} \\ \hline \mathbb{O}_{n,m} & \mathbb{A} \end{array}\right) = \left(\begin{array}{c|c} \mathbb{I}_m\mathbb{C}_{1,1} + \mathbb{O}_{m,n}\mathbb{O}_{n,m} & \mathbb{I}_m\mathbb{C}_{1,2} + \mathbb{O}_{m,n}\mathbb{A} \\ \hline \mathbb{O}_{n,m}\mathbb{C}_{1,1} + \mathbb{E}\mathbb{O}_{n,m} & \mathbb{O}_{n,m}\mathbb{C}_{1,2} + \mathbb{E}\mathbb{A} \end{array}\right)$$
$$= \left(\begin{array}{c|c} \mathbb{C}_{1,1} & \mathbb{C}_{1,2} \\ \hline \mathbb{O}_{n,m} & \mathbb{E}\mathbb{A} \end{array}\right) = \left(\begin{array}{c|c} \mathbb{C}_{1,1} & \mathbb{C}_{1,2} \\ \hline \mathbb{O}_{n,m} & \tilde{\mathbb{A}} \end{array}\right)$$

◇

On tire de cet exercice le lemme suivant

Lemme 3.3

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ avec $A_{1,1} \neq 0$. Il existe une matrice $\mathbb{E} \in \mathcal{M}_n(\mathbb{C})$ triangulaire inférieure à diagonale unité telle que

$$\mathbb{E}\mathbb{A}\boldsymbol{e}_1 = A_{1,1}\boldsymbol{e}_1 \tag{3.11}$$

où $\boldsymbol{e}_1$ est le premier vecteur de la base canonique de $\mathbb{C}^n$.

Exercice 3.1.4: Matrice de permutation

Soit $(i,j) \in [\![1,n]\!]^2$, on note $\mathbb{P}_n^{[i,j]} \in \mathcal{M}_n(\mathbb{R})$ la matrice identitée dont on a permuté les lignes i et j.

Q. 1 *Représenter cette matrice et la définir proprement.*

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$.On note $\boldsymbol{A}_{r,:}$ le r-ème vecteur ligne de $\mathbb{A}$ et $\boldsymbol{A}_{:,s}$ le s-ème vecteur colonne de $\mathbb{A}$.

Q. 2
1. *Déterminer $\mathbb{P}_n^{[i,j]}\mathbb{A}$ en fonction des vecteurs lignes de $\mathbb{A}$.*
2. *Déterminer $\mathbb{A}\mathbb{P}_n^{[i,j]}$ en fonction des vecteurs colonnes de $\mathbb{A}$.*

Q. 3
1. *Calculer le déterminant de $\mathbb{P}_n^{[i,j]}$.*
2. *Déterminer l'inverse de $\mathbb{P}_n^{[i,j]}$.*

Correction Exercice 3.1.4 On note $\mathbb{P} = \mathbb{P}_n^{[i,j]}$.

Q. 1 On peut définir cette matrice par ligne,

$$\begin{cases} \forall r \in [\![1,n]\!]\backslash\{i,j\}, & P_{r,s} = \delta_{r,s}, \quad \forall s \in [\![1,n]\!], \\ & P_{i,s} = \delta_{j,s}, \quad \forall s \in [\![1,n]\!], \\ & P_{j,s} = \delta_{i,s}, \quad \forall s \in [\![1,n]\!]. \end{cases}$$

ou par colonne

$$\begin{cases} \forall s \in [\![1,n]\!]\backslash\{i,j\}, & P_{r,s} = \delta_{r,s}, \quad \forall r \in [\![1,n]\!], \\ & P_{r,i} = \delta_{r,j}, \quad \forall r \in [\![1,n]\!], \\ & P_{r,j} = \delta_{r,i}, \quad \forall r \in [\![1,n]\!]. \end{cases}$$

Ne pas utiliser les indices i et j qui sont déjà fixés dans la définition de la matrice $\mathbb{P} = \mathbb{P}_n^{[i,j]}$.

On peut noter que la matrice $\mathbb{P}$ est symétrique.

Q. 2 1. On note $\mathbb{D} = \mathbb{P}\mathbb{A}$. Par définition du produit matriciel on a

$$D_{r,s} = \sum_{k=1}^{n} P_{r,k}A_{k,s}.$$

On obtient, $\forall s \in [\![1,n]\!]$,

$$\begin{cases} D_{r,s} &= \sum_{k=1}^{n} \delta_{r,k} A_{k,s} = A_{r,s}, \quad \forall r \in [\![1,n]\!] \backslash \{i,j\}, \\ D_{i,s} &= \sum_{k=1}^{n} \delta_{j,k} A_{k,s} = A_{j,s}, \\ D_{j,s} &= \sum_{k=1}^{n} \delta_{i,k} A_{k,s} = A_{i,s}. \end{cases}$$

ce qui donne

$$\begin{cases} \boldsymbol{D}_{r,:} &= \boldsymbol{A}_{r,:}, \quad \forall r \in [\![1,n]\!] \backslash \{i,j\}, \\ \boldsymbol{D}_{i,:} &= \boldsymbol{A}_{j,:}, \\ \boldsymbol{D}_{j,:} &= \boldsymbol{A}_{i,:}. \end{cases}$$

La notation $\boldsymbol{D}_{i,:}$ correspond au vecteur ligne $(D_{i,1}, \dots, D_{i,n})$ et $\boldsymbol{D}_{:,j}$ correspond au vecteur colonne $\begin{pmatrix} D_{1,j} \\ \vdots \\ D_{n,j} \end{pmatrix}$

2. On note $\mathbb{E} = \mathbb{A}\mathbb{P}$. Par définition du produit matriciel et par symétrie de $\mathbb{P}$ on a

$$E_{r,s} = \sum_{k=1}^{n} A_{r,k} P_{k,s} = \sum_{k=1}^{n} A_{r,k} P_{s,k}.$$


Ne pas utiliser les indices i et j qui sont déjà fixés dans la définition de la matrice $\mathbb{P} = \mathbb{P}_n^{[i,j]}$.

On obtient en raisonnant par colonne, $\forall r \in [\![1,n]\!]$,

$$\begin{cases} E_{r,s} &= \sum_{k=1}^{n} A_{r,k} \delta_{s,k} = A_{r,s}, \quad \forall s \in [\![1,n]\!] \backslash \{i,j\}, \\ E_{r,i} &= \sum_{k=1}^{n} A_{r,k} \delta_{j,k} = A_{r,j}, \\ E_{r,j} &= \sum_{k=1}^{n} A_{r,k} \delta_{i,k} = A_{r,i}. \end{cases}$$

ce qui donne

$$\begin{cases} \boldsymbol{E}_{:,s} &= \boldsymbol{A}_{:,s}, \quad \forall s \in [\![1,n]\!] \backslash \{i,j\}, \\ \boldsymbol{E}_{:,i} &= \boldsymbol{A}_{:,j}, \\ \boldsymbol{E}_{:,j} &= \boldsymbol{A}_{:,i}. \end{cases}$$

Q. 3 1. $\det(\mathbb{P}) = -1$, si $i \neq j$ et $\det(\mathbb{P}) = 1$ sinon.

2. Immédiat par calcul direct on a $\mathbb{P}\mathbb{P} = \mathbb{I}$ et donc la matrice $\mathbb{P}$ est inversible et $\mathbb{P}^{-1} = \mathbb{P}$.

◇

On tire de cet exercice le lemme suivant


Lemme 3.4

Soit $(i,j) \in [\![1,n]\!]^2$. On note $\mathbb{P}_n^{[i,j]} \in \mathcal{M}_n(\mathbb{R})$ la matrice identitée dont on a permuté les lignes i et j. Alors la matrice $\mathbb{P}_n^{[i,j]}$ est **symétrique et orthogonale**. Pour toute matrice $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$,

1. la matrice $\mathbb{P}_n^{[i,j]}\mathbb{A}$ est la matrice $\mathbb{A}$ dont on a permuté les **lignes** i et j,
2. la matrice $\mathbb{A}\mathbb{P}_n^{[i,j]}$ est la matrice $\mathbb{A}$ dont on a permuté les **colonnes** i et j,

3.1.3 Méthode de Gauss-Jordan, écriture matricielle

Soient $\mathbb{A} \in \mathcal{M}_{\mathbb{K}}()$ une matrice inversible et $\boldsymbol{b} \in \mathbb{K}^n$.
On va tout d'abord rappeler (très) brièvement l'**algorithme d'élimination** ou **algorithme de Gauss-Jordan** permettant de transformer le système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ en un système linéaire équivalent dont la matrice est triangulaire supérieure. Ce dernier système se résoud par la méthode de remontée.
Ensuite, on va réécrire cet algorithme sous forme algébrique pour obtenir le théorème ...

Cette méthode doit son nom aux mathématiciens Carl Friedrich Gauss (1777-1855, mathématicien, astronome et physicien allemand) et Wilhelm Jordan (1842-1899, mathématicien et géodésien allemand) mais elle est connue des Chinois depuis au moins le Ier siècle de notre ère. Elle est référencée dans l'important livre chinois *Jiuzhang suanshu* ou *Les Neuf Chapitres sur l'art mathématique*, dont elle constitue le huitième chapitre, sous le titre « Fang cheng » (la disposition rectangulaire). La méthode est présentée au moyen de dix-huit exercices. Dans son commentaire daté de 263, Liu Hui en attribue la paternité à Chang Ts'ang, chancelier de l'empereur de Chine au IIe siècle avant notre ère.

Algorithme de Gauss-Jordan usuel

Pour la résolution du système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ l'algorithme de Gauss-Jordan produit la forme échelonnée (réduite) d'une matrice à l'aide d'opérations élémentaires sur les lignes du système. Trois types d'opérations élémentaires sont utilisées:

- Permutation de deux lignes ;
- Multiplication d'une ligne par un scalaire non nul ;
- Ajout du multiple d'une ligne à une autre ligne.

A l'aide de ces opérations élémentaires cet algorithme permet donc de transformer le système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ en le système équivalent $\mathbb{U}\boldsymbol{x} = \boldsymbol{f}$ où $\mathbb{U}$ est triangulaire supérieure. En fait, l'algorithme va transformer la matrice $\mathbb{A}$ et le second membre $\boldsymbol{b}$ pour aboutir à un système dont la matrice est triangulaire supérieure.

Algorithme 3.4 Algorithme de Gauss-Jordan formel pour la résolution de $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$

1: **Pour** $j \leftarrow 1$ **à** $n-1$ **faire**
2: Rechercher l'indice k de la ligne du pivot (sur la colonne j, $k \in [\![j, n]\!]$)
3: Permuter les lignes j ($\mathcal{L}_j$) et k ($\mathcal{L}_k$) du système si besoin.
4: **Pour** $i \leftarrow j+1$ **à** n **faire**
5: Eliminer en effectuant $\mathcal{L}_i \leftarrow \mathcal{L}_i - \frac{A_{i,j}}{A_{j,j}}\mathcal{L}_j$
6: **Fin Pour**
7: **Fin Pour**
8: Résoudre le système triangulaire supérieur par la méthode de la remontée.

On va maintenant voir comment écrire cet algorithme de manière plus détaillée. Pour conserver sa lisibilité, on choisi pour chaque ligne un peu délicate de créer et d'utiliser une fonction dédiée à cette tâche.

Algorithme 3.5 Algorithme de Gauss-Jordan avec fonctions pour la résolution de $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$ inversible.
$\boldsymbol{b}$: vecteur de $\mathbb{K}^n$.
Résultat : $\boldsymbol{x}$: vecteur de $\mathbb{K}^n$.

```
1: Fonction x ← RSLGauss ( A, b )
2:   Pour j ← 1 à n − 1 faire
3:     k ← ChercheIndPivot (A, j)                    ▷ ChercheIndPivot à écrire
4:     [A, b] ← PermLignesSys (A, b, j, k)           ▷ PermLignesSys à écrire
5:     Pour i ← j + 1 à n faire
6:       [A, b] ← CombLignesSys (A, b, j, i, −A(i, j)/A(j, j))   ▷ CombLignesSys à écrire
7:     Fin Pour
8:   Fin Pour
9:   x ← RSLTriSup(A, b)                            ▷ RSLTriSup déjà écrite
10: Fin Fonction
```

Bien évidemment, il reste à décrire et écrire les différentes fonctions utilisées dans cette fonction :

Fonction $k \leftarrow$ **ChercheIndPivot**$(\mathbb{A}, j)$ **:** recherche $k \in [\![j, n]\!]$ tel que $\forall l \in [\![j, n]\!]$, $|A_{l,j}| \leqslant |A_{k,j}|$.

Fonction $[\mathbb{A}, \boldsymbol{b}] \leftarrow$ **PermLignesSys**$(\mathbb{A}, \boldsymbol{b}, i, k)$ **:** permute les lignes i et k de la matrice $\mathbb{A}$ ainsi que celles du vecteur $\boldsymbol{b}$.

Fonction $[\mathbb{A}, \boldsymbol{b}] \leftarrow$ **CombLignesSys**$(\mathbb{A}, \boldsymbol{b}, j, i, \alpha)$ **:** remplace la ligne i de la matrice A par la combinaison linéaire $\mathcal{L}_i \leftarrow \mathcal{L}_i + \alpha\mathcal{L}_j$. De même on remplace la ligne i de $\boldsymbol{b}$ par $b_i + \alpha b_j$.

Ces trois fonctions sont simples à écrire et sont données en Algorithmes 3.6, 3.7 et 3.8.

Algorithme 3.6 Recherche d'un pivot pour l'algorithme de Gauss-Jordan.

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$.
j : entier, $1 \leqslant j \leqslant n$.
Résultat : k : entier, indice ligne pivot

```
1: Fonction k ← ChercheIndPivot ( A, j )
2:   k ← j, pivot ← |A(j, j)|
3:   Pour i ← j + 1 à n faire
4:     Si |A(i, j)| > pivot alors
5:       k ← i, pivot ← |A(i, j)|
6:     Fin Si
7:   Fin Pour
8: Fin Fonction
```

Algorithme 3.7 Permutte deux lignes d'une matrice et d'un vecteur.

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$.
$\boldsymbol{b}$: vecteur de $\mathbb{K}^n$.
j, k : entiers, $1 \leqslant j, k \leqslant n$.
Résultat : $\mathbb{A}$ et $\boldsymbol{b}$ modifiés.

```
1: Fonction [A, b] ← PermLignesSys ( A, b, j, k )
2:   Pour l ← 1 à n faire
3:     t ← A(j, l), A(j, l) ← A(k, l), A(k, l) ← t
4:   Fin Pour
5:   t ← b(j), b(j) ← b(k), b(k) ← t
6: Fin Fonction
```

Algorithme 3.8 Combinaison linéaire $\mathcal{L}_i \leftarrow \mathcal{L}_i + \alpha\mathcal{L}_j$ appliqué à une matrice et à un vecteur.

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$.
$\boldsymbol{b}$: vecteur de $\mathbb{K}^n$.
j, i : entiers, $1 \leqslant j, i \leqslant n$.
alpha : scalaire de $\mathbb{K}$
Résultat : $\mathbb{A}$ et $\boldsymbol{b}$ modifiés.

```
1: Fonction [A, b] ← CombLignesSys ( A, b, j, i, α )
2:   Pour k ← 1 à n faire
3:     A(i, k) ← A(i, k) + α ∗ A(j, k)
4:   Fin Pour
5:   b(i) ← b(i) + αb(j)
6: Fin Fonction
```

Ecriture algébrique

Sous forme d'exercice :

Exercice 3.1.5

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ inversible.

Q. 1 *Montrer qu'il existe une matrice* $\mathbb{G} \in \mathcal{M}_n(\mathbb{C})$ *telle que* $|\det(\mathbb{G})| = 1$ *et* $\mathbb{G}\mathbb{A}\boldsymbol{e}_1 = \alpha\boldsymbol{e}_1$ *avec* $\alpha \neq 0$ *et* $\boldsymbol{e}_1$ *premier vecteur de la base canonique de* C^n.

Q. 2 1. *Montrer par récurrence sur l'ordre des matrices que pour toute matrice* $\mathbb{A}_n \in \mathcal{M}_n(\mathbb{C})$ *inversible, il existe une matrice* $\mathbb{S}_n \in \mathcal{M}_n(\mathbb{C})$ *telle que* $|\det \mathbb{S}_n| = 1$ *et* $\mathbb{S}_n\mathbb{A}_n = \mathbb{U}_n$ *avec* $\mathbb{U}_n$ *matrice triangulaire supérieure inversible.*

2. *Soit* $\boldsymbol{b} \in \mathbb{C}^n$. *En supposant connue la décompostion précédente* $\mathbb{S}_n\mathbb{A}_n = \mathbb{U}_n$, *expliquer comment résoudre le système* $\mathbb{A}_n\boldsymbol{x} = \boldsymbol{b}$.

Q. 3 *Que peut-on dire si* $\mathbb{A}$ *est non inversible?*

Indication : utiliser les résultats des exercices 3.1.3 et 3.1.4.

Correction Exercice 3.1.5

Q. 1 D'après le Lemme 3.3, si $A_{1,1} \neq 0$ le résultat est immédiat. Dans l'énoncé rien ne vient corroborer cette hypothèse. Toutefois, comme la matrice $\mathbb{A}$ est inversible, il existe au moins un $p \in [\![1,n]\!]$ tel que $A_{p,1} \neq 0$. On peut même choisir le premier indice p tel que $|A_{p,1}| = \max_{i\in[\![1,n]\!]} |A_{i,1}| > 0$ (pivot de l'algorithme de Gauss-Jordan). On note $\mathbb{P} = \mathbb{P}_n^{[1,p]}$ la matrice de permutation des lignes 1 et p (voir exercice 3.1.4, page 66). De plus on a

$$|\det \mathbb{P}| = 1 \quad \text{et} \quad \mathbb{P}^{-1} = \mathbb{P}.$$

Par construction $(\mathbb{P}\mathbb{A})_{1,1} = A_{p,1} \neq 0$, et on peut alors appliquer le Lemme 3.3 à la matrice $(\mathbb{P}\mathbb{A})$ pour obtenir l'existence d'une matrice $\mathbb{E} \in \mathcal{M}_n(\mathbb{C})$ vérifiant $\det \mathbb{E} = 1$ et telle que

$$\mathbb{E}(\mathbb{P}\mathbb{A})\boldsymbol{e}_1 = A_{p,1}\boldsymbol{e}_1.$$

En posant $\mathbb{G} = \mathbb{E}\mathbb{P}$ et $\alpha = A_{p,1}$, on obtient bien $\mathbb{G}\mathbb{A}\boldsymbol{e}_1 = \alpha\boldsymbol{e}_1$. De plus, on a

$$|\det \mathbb{G}| = |\det(\mathbb{E}\mathbb{P})| = |\det \mathbb{E} \times \det \mathbb{P}| = 1.$$

Remarque 3.5 La matrice $\mathbb{G}$ étant inversible, on a

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b} \iff \mathbb{G}\mathbb{A}\boldsymbol{x} = \mathbb{G}\boldsymbol{b}$$

ce qui correspond à la première *permutation/élimination* de l'algorithme de Gauss-Jordan.

Q. 2 1. On veut démontrer par récurrence la propriété $(\mathcal{P}_n)$,

$$(\mathcal{P}_n) \quad \left\{ \begin{array}{l} \forall \mathbb{A}_n \in \mathcal{M}_n(\mathbb{C}),\ \text{inversible } \exists \mathbb{S}_n \in \mathcal{M}_n(\mathbb{C}),\ |\det \mathbb{S}_n| = 1,\ \text{tel que} \\ \text{la matrice } \mathbb{U}_n = \mathbb{S}_n\mathbb{A} \text{ soit une triangulaire supérieure inversible} \end{array} \right.$$

Initialisation : Pour $n = 2$. Soit $\mathbb{A}_2 \in \mathcal{M}_2(\mathbb{C})$ inversible. En utilisant la question précédente il existe $\mathbb{G}_2 \in \mathcal{M}_2(\mathbb{C})$ telle que $|\det \mathbb{G}_2| = 1$ et $\mathbb{G}_2\mathbb{A}_2\boldsymbol{e}_1 = \alpha\boldsymbol{e}_1$ avec $\alpha \neq 0$ et $\boldsymbol{e}_1$ premier vecteur de la base canonique de $\mathbb{C}^2$. On note $\mathbb{U}_2 = \mathbb{G}_2\mathbb{A}_2$. Cette matrice s'écrit donc sous la forme

$$\mathbb{U}_2 = \begin{pmatrix} \alpha & \bullet \\ 0 & \bullet \end{pmatrix}$$

et elle est triangulaire supérieure. Les matrices $\mathbb{G}_2$ et $\mathbb{A}_2$ étant inversible, leur produit $\mathbb{U}_2$ l'est aussi. La proposition $(\mathcal{P}_2)$ est donc vérifiée avec $\mathbb{S}_2 = \mathbb{G}_2$.

Hérédité : Soit $n \geqslant 3$. On suppose que $(\mathcal{P}_{n-1})$ est vraie. Montrons que $(\mathcal{P}_n)$ est vérifiée.
Soit $\mathbb{A}_n \in \mathcal{M}_n(\mathbb{C})$ inversible. En utilisant la question précédente il existe $\mathbb{G}_n \in \mathcal{M}_n(\mathbb{C})$ telle que $|\det \mathbb{G}_n| = 1$ et $\mathbb{G}_n\mathbb{A}_n\boldsymbol{e}_1 = \alpha_n\boldsymbol{e}_1$ avec $\alpha_n \neq 0$ et $\boldsymbol{e}_1$ premier vecteur de la base canonique de $\mathbb{C}^n$. On note $\mathbb{V}_n = \mathbb{G}_n\mathbb{A}_n$. Cette matrice s'écrit donc sous la forme

$$\mathbb{V}_n = \left(\begin{array}{c|ccc} \alpha_n & \bullet & \dots & \bullet \\ \hline 0 & \bullet & \dots & \bullet \\ \vdots & \vdots & & \vdots \\ 0 & \bullet & \dots & \bullet \end{array}\right) \overset{\text{def}}{=} \left(\begin{array}{c|c} \alpha_n & \boldsymbol{c}_{n-1}^* \\ \hline 0 & \\ \vdots & \mathbb{B}_{n-1} \\ 0 & \end{array}\right)$$

où $\boldsymbol{c}_{n-1} \in \mathbb{C}^{n-1}$ et $\mathbb{B}_{n-1} \in \mathcal{M}_{n-1}(\mathbb{C})$. Comme $\mathbb{G}_n$ et $\mathbb{A}_n$ sont inversibles, $\mathbb{V}_n$ l'est aussi. On en déduit donc que $\mathbb{B}_{n-1}$ est inversible car $0 \neq \det \mathbb{V}_n = \alpha_n \times \det \mathbb{B}_{n-1}$ et $\alpha_n \neq 0$.

On peut donc utiliser la propriété $(\mathcal{P}_{n-1})$ (hyp. de récurrence) sur la matrice $\mathbb{B}_{n-1}$: il existe donc $\mathbb{S}_{n-1} \in \mathcal{M}_{n-1}(\mathbb{C})$, avec $|\det \mathbb{S}_{n-1}| = 1$, tel que la matrice $\mathbb{U}_{n-1} = \mathbb{S}_{n-1}\mathbb{B}_{n-1}$ soit une triangulaire supérieure inversible.

Soit $\mathbb{Q}_n \in \mathcal{M}_n(\mathbb{C})$ la matrice définie par

$$\mathbb{Q}_n = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{S}_{n-1} & \\ 0 & & & \end{array}\right)$$

On a alors

$$\mathbb{Q}_n\mathbb{G}_n\mathbb{A}_n = \mathbb{Q}_n\mathbb{V}_n = \left(\begin{array}{c|c} 1 & 0 \quad \dots \quad 0 \\ \hline 0 & \\ \vdots & \mathbb{S}_{n-1} \\ 0 & \end{array}\right)\left(\begin{array}{c|c} \alpha_n & \boldsymbol{c}^*_{n-1} \\ \hline 0 & \\ \vdots & \mathbb{B}_{n-1} \\ 0 & \end{array}\right)$$

$$= \left(\begin{array}{c|c} \alpha_n & \boldsymbol{c}^*_{n-1} \\ \hline 0 & \\ \vdots & \mathbb{S}_{n-1}\mathbb{B}_{n-1} \\ 0 & \end{array}\right) = \left(\begin{array}{c|c} \alpha_n & \boldsymbol{c}^*_{n-1} \\ \hline 0 & \\ \vdots & \mathbb{U}_{n-1} \\ 0 & \end{array}\right) \stackrel{\text{def}}{=} \mathbb{U}_n$$

La matrice $\mathbb{U}_n$ est triangulaire supérieure inversible car $\mathbb{U}_{n-1}$ l'est aussi et $\alpha_n \neq 0$.

On pose $\mathbb{S}_n = \mathbb{Q}_n\mathbb{G}_n$. On a donc $\mathbb{S}_n\mathbb{A}_n = \mathbb{U}_n$ et comme $|\det \mathbb{S}_n| = 1$ car $|\det \mathbb{G}_n| = 1$ et $\det \mathbb{G}_n = 1$, ceci prouve la véracité de la proposition $(\mathcal{P}_n)$.

2. Comme $\mathbb{S}_n$ est inversible, on a en multipliant à gauche le système par $\mathbb{S}_n$

$$\mathbb{A}_n\boldsymbol{x} = \boldsymbol{b} \iff \mathbb{S}_n\mathbb{A}_n\boldsymbol{x} = \mathbb{S}_n\boldsymbol{b} \iff \mathbb{U}_n\boldsymbol{x} = \mathbb{S}_n\boldsymbol{b}$$

Pour déterminer le vecteur $\boldsymbol{x}$, on peut alors résoudre le dernier système par l'algorithme de remontée.

Q. 3 (rapide) Si $\mathbb{A}$ est non inversible, alors dans la première question nous ne sommes pas assurés d'avoir $\alpha \neq 0$. Cependant l'existence de la matrice $\mathbb{G}$ reste avérée.

Pour la deuxième question, le seule changement vient du fait que la matrice $\mathbb{U}_n$ n'est plus inversible.

◇

On a donc démontré le théorème suivant

Théorème 3.6

Soit $\mathbb{A}$ une matrice carrée, inversible ou non. Il existe (au moins) une matrice inversible $\mathbb{G}$ telle que $\mathbb{G}\mathbb{A}$ soit triangulaire supérieure.

Preuve. voir Exercice 3.1.5, page 69. □

3.1.4 Factorisation $\mathbb{L}\mathbb{U}$

Avant de citer le théorème principal, on va faire un "petit" exercice...

Exercice 3.1.6: ★★★★☆

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice dont les sous-matrices principales d'ordre i, notées Δ_i, $i \in [\![1, n]\!]$ (voir Definition B.48, page 207) sont inversibles.
Montrer qu'il existe des matrices $\mathbb{E}^{[k]} \in \mathcal{M}_n(\mathbb{C})$, $k \in [\![1, n-1]\!]$, triangulaires inférieures à diagonale unité telles que la matrice $\mathbb{U}$ définie par

$$\mathbb{U} = \mathbb{E}^{[n-1]} \dots \mathbb{E}^{[1]}\mathbb{A}$$

soit triangulaire supérieure avec $U_{i,i} = \det \Delta_i / (U_{1,1} \times \dots \times U_{i-1,i-1})$, $\forall i \in [\![1, n]\!]$.

Correction Exercice 3.1.6

On note $\mathbb{A}^{[0]} = \mathbb{A}$. On va démontrer par récurrence finie sur $k \in [\![1, n-1]\!]$, qu'il existe une matrice $\mathbb{E}^{[k]} \in \mathcal{M}_n(\mathbb{C})$, tel que la matrice $\mathbb{A}^{[k]}$ définie itérativement par

$$\mathbb{A}^{[k]} = \mathbb{E}^{[k]}\mathbb{A}^{[k-1]}$$

s'écrit sous la forme bloc

$$\mathbb{A}^{[k]} = \left(\begin{array}{cccc:ccc} \alpha_1 & \bullet & \cdots & \bullet & \bullet & \cdots & \bullet \\ 0 & \ddots & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & \bullet & \vdots & & \vdots \\ 0 & \dots & 0 & \alpha_k & \bullet & \cdots & \bullet \\ \hdashline 0 & \dots & \dots & 0 & \bullet & \cdots & \bullet \\ \vdots & \ddots & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \dots & \dots & 0 & \bullet & \cdots & \bullet \end{array}\right) \tag{3.12}$$

avec $\alpha_1 = A_{1,1}$ et $\forall i \in [\![2,k]\!]$, $\alpha_i = \det \Delta_i / (\alpha_1 \times \cdots \times \alpha_{i-1})$.

Initialisation ($k = 1$)**:** On a $A_{1,1} \neq 0$ car $\Delta_1 = A_{1,1}$ et $\det \Delta_1 \neq 0$. D'après le Lemme 3.3, il existe une matrice $\mathbb{E}^{[1]} \in \mathcal{M}_n(\mathbb{C})$, triangulaire inférieure à diagonale unité telle que $\mathbb{E}^{[1]}\mathbb{A}\boldsymbol{e}_1 = A_{1,1}\boldsymbol{e}_1$ où $\boldsymbol{e}_1$ est le premier vecteur de la base canonique de $\mathbb{C}^n$. On a alors

$$\mathbb{A}^{[1]} = \mathbb{E}^{[1]}\mathbb{A} = \left(\begin{array}{c:ccc} \alpha_1 & \bullet & \cdots & \bullet \\ \hdashline 0 & \bullet & \cdots & \bullet \\ \vdots & \vdots & & \vdots \\ 0 & \bullet & \cdots & \bullet \end{array}\right)$$

avec $\alpha_1 = A_{1,1} = \det \Delta_1$.

Hérédité ($k < n-1$)**:** Supposons construite la matrice $\mathbb{A}^{[k]}$. Il existe donc k matrices, $\mathbb{E}^{[1]}, \dots, \mathbb{E}^{[k]}$, triangulaires inférieures à diagonale unité telles que

$$\mathbb{A}^{[k]} = \mathbb{E}^{[k]} \dots \mathbb{E}^{[1]}\mathbb{A}.$$

- On va montrer que $\alpha_{k+1} \overset{\text{def}}{=} A^{[k]}_{k+1,k+1} \neq 0$. Pour celà, on réécrit la matrice $\mathbb{A}^{[k]}$ sous forme bloc, avec comme premier bloc diagonale le bloc de dimension $k+1$:

$$\mathbb{A}^{[k]} = \left(\begin{array}{ccccc:ccc} \alpha_1 & \bullet & \cdots & \bullet & \bullet & \bullet & \cdots & \bullet \\ 0 & \ddots & \ddots & \vdots & \bullet & \vdots & & \vdots \\ \vdots & \ddots & \ddots & \bullet & \bullet & \vdots & & \vdots \\ 0 & \dots & 0 & \alpha_k & \bullet & \vdots & \cdots & \vdots \\ 0 & \dots & \dots & 0 & \alpha_{k+1} & \bullet & \cdots & \bullet \\ \hdashline 0 & \cdots & \cdots & 0 & \bullet & \bullet & \cdots & \bullet \\ \vdots & & & \vdots & \bullet & \vdots & & \vdots \\ 0 & \dots & \dots & 0 & \bullet & \bullet & \cdots & \bullet \end{array}\right) \quad \overset{k+1}{\longleftrightarrow}$$

La matrice $\mathbb{G}^{[k]} \overset{\text{def}}{=} \mathbb{E}^{[k]} \dots \mathbb{E}^{[1]}$ est triangulaire inférieure à diagonale unité car produit de matrices triangulaires inférieures à diagonale unité (voir Exercice B.3.2, page 214). Le produit de $\mathbb{G}^{[k]}\mathbb{A}$ s'écrit alors sous forme bloc

$$\mathbb{G}^{[k]}\mathbb{A} = \overset{k+1}{\longleftrightarrow} \left(\begin{array}{cccc:cccc} 1 & 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ \bullet & \ddots & \ddots & \vdots & \vdots & & & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots & & & \vdots \\ \bullet & \dots & \bullet & 1 & 0 & \cdots & \cdots & 0 \\ \hdashline \bullet & \dots & \dots & \bullet & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots & \bullet & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \ddots & 0 \\ \bullet & \dots & \dots & \bullet & \bullet & \cdots & \bullet & 1 \end{array}\right) \overset{k+1}{\longleftrightarrow} \left(\begin{array}{ccc:ccc} & & & \bullet & \cdots & \bullet \\ & \Delta_{k+1} & & \vdots & & \vdots \\ & & & \bullet & \cdots & \bullet \\ \hdashline \bullet & \cdots & \bullet & \bullet & \cdots & \bullet \\ \vdots & & \vdots & \vdots & & \vdots \\ \bullet & \cdots & \bullet & \bullet & \cdots & \bullet \end{array}\right)$$

Comme $\mathbb{A}^{[k]} = \mathbb{G}^{[k]}\mathbb{A}$, en utilisant les règles de multiplication par blocs des matrices on obtient

$$\begin{pmatrix} \alpha_1 & \bullet & \cdots & \bullet & \bullet \\ 0 & \ddots & \ddots & \vdots & \bullet \\ \vdots & \ddots & \ddots & \bullet & \bullet \\ 0 & \cdots & 0 & \alpha_k & \bullet \\ 0 & \cdots & \cdots & 0 & \alpha_{k+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \bullet & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \bullet & \cdots & \bullet & 1 \end{pmatrix} \begin{pmatrix} & & & \\ & \Delta_{k+1} & & \\ & & & \end{pmatrix}$$

En prenant le déterminant de cette dernière équation, et en utilisant le fait que le determinant d'une matrice triangulaire est le produit de ses coefficients diagonaux, on obtient

$$\prod_{i=1}^{k+1} \alpha_i = \det \Delta_{k+1}.$$

Par hypothèse Δ_{k+1} inversible , ce qui entraine $\det \Delta_{k+1} \neq 0$ et donc $\alpha_i \neq 0$, $\forall i \in [\![1, k+1]\!]$. On a donc

$$\alpha_{k+1} = \frac{\det \Delta_{k+1}}{\prod_{i=1}^{k} \alpha_i} \neq 0.$$

- Montrons l'existence d'une matrice triangulaire inférieure à diagonale unité permettant d'éliminer les termes sous diagonaux de la colonne $k+1$ de $\mathbb{A}^{[k]}$.
  Revenons à l'écriture bloc de premier bloc diagonal de dimension k. On a

$$\mathbb{A}^{[k]} = \left(\begin{array}{cccc:ccc} \alpha_1 & \bullet & \cdots & \bullet & \bullet & \cdots & \bullet \\ 0 & \ddots & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & \bullet & \vdots & & \vdots \\ 0 & \cdots & 0 & \alpha_k & \bullet & \cdots & \bullet \\ \hdashline 0 & \cdots & \cdots & 0 & \alpha_{k+1} & \bullet & \cdots \\ \vdots & \ddots & \ddots & \vdots & \bullet & \cdots & \bullet \\ \vdots & \ddots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & \bullet & \cdots & \bullet \end{array}\right) \overset{\text{def}}{=} \left(\begin{array}{c:c} \mathbb{U}^{[k]} & \mathbb{F}^{[k]} \\ \hdashline \mathbb{0} & \mathbb{V}^{[k]} \end{array}\right)$$

Nous sommes exactement dans le cas de figure étudié dans l'exercice 3.1.3, page 63. En effet, avec les notations de cet exercice et si l'on pose $\boldsymbol{v} = \mathbb{V}^{[k]}_{:,1} = (A^{[k]}_{k+1,k+1}, \ldots, A^{[k]}_{n,k+1})^{\mathsf{t}} \in \mathbb{C}^{n-(k+1)}$ (en bleu dans l'expression de $\mathbb{A}^{[k]}$ précédente) on a alors $v_1 = A^{[k]}_{k+1,k+1} = \alpha_{k+1} \neq 0$ et l'on peut définir la matrice $\mathbb{E}^{[k+1]} \in \mathcal{M}_n(\mathbb{C})$, triangulaire inférieure à diagonale unité, par

$$\mathbb{E}^{[k+1]} = \mathbb{E}^{[k,\boldsymbol{v}]} \overset{\text{def}}{=} \left(\begin{array}{c:c} \mathbb{I}_k & \mathbb{0} \\ \hdashline \mathbb{0} & \mathbb{E}^{[\boldsymbol{v}]} \end{array}\right)$$

avec $\mathbb{E}^{[\boldsymbol{v}]} \in \mathcal{M}_{n-k}(\mathbb{C})$ triangulaire inférieure à diagonale unité (définie dans l'exercice 3.1.3) telle que

$$\mathbb{E}^{[\boldsymbol{v}]}\mathbb{V}^{[k]} = \begin{pmatrix} \alpha_{k+1} & \bullet & \cdots & \bullet \\ 0 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & \bullet & \cdots & \bullet \end{pmatrix}$$

On a alors

$$\begin{aligned} \mathbb{A}^{[k+1]} \overset{\text{def}}{=} \mathbb{E}^{[k+1]}\mathbb{A}^{[k]} &= \left(\begin{array}{c:c} \mathbb{I}_k & \mathbb{0} \\ \hdashline \mathbb{0} & \mathbb{E}^{[\boldsymbol{v}]} \end{array}\right) \left(\begin{array}{c:c} \mathbb{U}^{[k]} & \mathbb{F}^{[k]} \\ \hdashline \mathbb{0} & \mathbb{V}^{[k]} \end{array}\right) \\ &= \left(\begin{array}{c:c} \mathbb{U}^{[k]} & \mathbb{F}^{[k]} \\ \hdashline \mathbb{0} & \mathbb{E}^{[\boldsymbol{v}]}\mathbb{V}^{[k]} \end{array}\right) \end{aligned}$$

et donc $\mathbb{A}^{[k+1]}$ s'écrit bien sous la forme (3.12) au rang $k+1$.

Final ($k = n-1$)**:** On a donc

$$\mathbb{U} = \mathbb{A}^{[n-1]} \overset{\text{def}}{=} \mathbb{E}^{[n-1]} \times \cdots \times \mathbb{E}^{[1]} \times \mathbb{A} = \left(\begin{array}{cccc:c} \alpha_1 & \bullet & \cdots & \cdots & \bullet \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \alpha_{n-1} & \bullet \\ \hdashline 0 & \cdots & \cdots & 0 & \bullet \end{array}\right) \tag{3.13}$$

où pour tout $k \in [\![1, n-1]\!]$ les matrices $\mathbb{E}^{[k]}$ sont triangulaires inférieures à diagonale unité.

Pour achever l'exercice, il reste à démontrer que

$$U_{n,n} = \det \Delta_n / (U_{1,1} \times \cdots \times U_{n-1,n-1}).$$

En effet, en prenant le déterminant dans (3.13) on obtient

$$\det\left(\mathbb{E}^{[n-1]} \times \cdots \times \mathbb{E}^{[1]} \times \mathbb{A}\right) = \det \left(\begin{array}{cccc:c} U_{1,1} & \bullet & \cdots & \cdots & \bullet \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & U_{n-1,n-1} & \bullet \\ \hdashline 0 & \cdots & \cdots & 0 & U_{n,n} \end{array}\right)$$

Comme le déterminant d'un produit de matrices est égale au produit des déterminants des matrices on a

$$\begin{aligned} \det\left(\mathbb{E}^{[n-1]} \times \cdots \times \mathbb{E}^{[1]} \times \mathbb{A}\right) &= \det \mathbb{E}^{[n-1]} \times \cdots \times \det \mathbb{E}^{[1]} \times \det \mathbb{A} \\ &= \det \mathbb{A} \end{aligned}$$

car les matrices $\mathbb{E}^{[k]}$ sont triangulaires inférieures à diagonale unité et donc $\det \mathbb{E}^{[k]} = 1$, $\forall k \in [\![1, n-1]\!]$. De plus, le déterminant d'une matrice traingulaire supérieure est égale au produit de ses coefficients diagonaux et donc

$$\det \left(\begin{array}{cccc:c} U_{1,1} & \bullet & \cdots & \cdots & \bullet \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & U_{n-1,n-1} & \bullet \\ \hdashline 0 & \cdots & \cdots & 0 & U_{n,n} \end{array}\right) = U_{n,n} \prod_{k=1}^{n-1} Uk, k.$$

On a alors

$$\det \mathbb{A} = \det \Delta_n = U_{n,n} \prod_{k=1}^{n-1} Uk, k \neq 0.$$

◇

Théorème 3.7: Factorisation $\mathbb{LU}$ ★★★★★

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice dont les sous-matrices principales sont inversibles alors il existe une unique matrice $\mathbb{L} \in \mathcal{M}_n(\mathbb{C})$ triangulaire inférieure (*lower triangular* en anglais) à diagonale unité et une unique matrice $\mathbb{U} \in \mathcal{M}_n(\mathbb{C})$ triangulaire supérieure (*upper triangular* en anglais) inversible telles ques

$$\mathbb{A} = \mathbb{LU}.$$

Preuve. Nous allons voir deux méthodes pour démontrer l'existence.

méthode 1 En utilisant le résultat de l'exercice 3.1.6, il existe des matrices $\mathbb{E}^{[k]} \in \mathcal{M}_n(\mathbb{C})$, $k \in [\![1, n-1]\!]$, triangulaires inférieures à diagonale unité telles que la matrice $\mathbb{U}$ définie par

$$\mathbb{U} = \mathbb{E}^{[n-1]} \dots \mathbb{E}^{[1]} \mathbb{A}$$

soit triangulaire supérieure avec $U_{i,i} \neq 0$, $\forall i \in [\![1,n]\!]$. On pose $\mathbb{G} = \mathbb{E}^{[n-1]} \cdots \mathbb{E}^{[1]}$. La matrice $\mathbb{G}$ est donc aussi triangulaire inférieure à diagonale unité. Elle est donc inversible est son inverse est triangulaire inférieure à diagonale unité (voir Proposition B.46, page 207). En notant $\mathbb{L} = \mathbb{G}^{-1}$ on a $\mathbb{A} = \mathbb{L}\mathbb{U}$.

méthode 2 Nous allons effectuer une démonstration par récurrence sur l'ordre n de la matrice $\mathbb{A}$.

Propriété: Pour tout $n \in \mathbb{N}^*$, si une matrice $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ a toutes ses sous-matrices principales inversibles alors il existe une unique matrice $\mathbb{L} \in \mathcal{M}_n(\mathbb{C})$ triangulaire inférieure à diagonale unité et une unique matrice $\mathbb{U} \in \mathcal{M}_n(\mathbb{C})$ triangulaire supérieure inversible telles ques $\mathbb{A} = \mathbb{L}\mathbb{U}$.

Initialisation: Pour $n = 1$, c'est trivial $\mathbb{L} = 1$ et $\mathbb{U} = \mathbb{A}$.

Hérédité: Supposons la propriété vraie au rang n, montrons qu'alors, la propriété est vraie au rang $n+1$.

Soit $\mathbb{A} \in \mathcal{M}_{n+1}(\mathbb{C})$ telle que toutes les sous-matrices principales soient inversibles. On peut la décomposer sous la forme bloc suivante

$$\mathbb{A} = \left(\begin{array}{c|c} \underline{\mathbb{A}} & \boldsymbol{f} \\ \hline \boldsymbol{e}^* & d \end{array}\right)$$

où

- $\underline{\mathbb{A}}$ est la matrice de $\mathcal{M}_n(\mathbb{C})$ telle que $\underline{\mathrm{A}}_{i,j} = \mathrm{A}_{i,j}$, $\forall (i,j) \in [\![1,n]\!]^2$,
- $\boldsymbol{f}$ est le vecteur de $\mathbb{C}^n$ tel que $f_i = \mathrm{A}_{i,n+1}$, $\forall i \in [\![1,n]\!]$,
- $\boldsymbol{e}$ est le vecteur de $\mathbb{C}^n$ tel que $e_i = \overline{\mathrm{A}_{n+1,i}}$, $\forall i \in [\![1,n]\!]$
- $d \in \mathbb{C}$ est le scalaire $d = \mathrm{A}_{n+1,n+1}$.

Comme les sous-matrices principales de $\underline{\mathbb{A}}$ sont les n premières sous-matrices principales de $\mathbb{A}$, elles sont donc inversibles. Par hypothèse de récurrence, il existe une unique matrice $\underline{\mathbb{L}} \in \mathcal{M}_n(\mathbb{C})$ triangulaire inférieure à diagonale unité et une unique matrice $\underline{\mathbb{U}} \in \mathcal{M}_n(\mathbb{C})$ triangulaire supérieure inversible telles ques

$$\underline{\mathbb{A}} = \underline{\mathbb{L}}\,\underline{\mathbb{U}}.$$

On va construire des matrices $\mathbb{L}$ et $\mathbb{U}$ vérifiant la propriété. Soient $\mathbb{L} \in \mathcal{M}_{n+1}(\mathbb{C})$ et $\mathbb{U} \in \mathcal{M}_{n+1}(\mathbb{C})$ décomposées sous la forme bloc

$$\mathbb{L} = \left(\begin{array}{c|c} \underline{\mathbb{L}} & \boldsymbol{0} \\ \hline \boldsymbol{g}^* & 1 \end{array}\right) \quad \text{et} \quad \mathbb{U} = \left(\begin{array}{c|c} \underline{\mathbb{U}} & \boldsymbol{h} \\ \hline \boldsymbol{0}^* & \alpha \end{array}\right)$$

où $\boldsymbol{g} \in \mathbb{C}^n$, $\boldsymbol{h} \in \mathbb{C}^n$ and $\alpha \in \mathbb{C}$. Les deux matrices sont blocs compatibles pour la multiplication et on a alors

$$\mathbb{L}\mathbb{U} = \left(\begin{array}{c|c} \underline{\mathbb{L}} & \boldsymbol{0} \\ \hline \boldsymbol{g}^* & 1 \end{array}\right)\left(\begin{array}{c|c} \underline{\mathbb{U}} & \boldsymbol{h} \\ \hline \boldsymbol{0}^* & \alpha \end{array}\right) = \left(\begin{array}{c|c} \underline{\mathbb{L}}\,\underline{\mathbb{U}} & \underline{\mathbb{L}}\boldsymbol{h} \\ \hline \boldsymbol{g}^*\underline{\mathbb{U}} & \boldsymbol{g}^*\boldsymbol{h} + \alpha \end{array}\right)$$

Comme $\underline{\mathbb{A}} = \underline{\mathbb{L}}\,\underline{\mathbb{U}}$, on va idenfier bloc par bloc les matrices $\mathbb{L}\mathbb{U}$ et $\mathbb{A}$, puis vérifier que le système est resolvable. On obtient donc par identification le système:

$$\left\{\begin{array}{lcl} \underline{\mathbb{L}}\boldsymbol{h} & = & \boldsymbol{f} \\ \boldsymbol{g}^*\underline{\mathbb{U}} & = & \boldsymbol{e}^* \\ \boldsymbol{g}^*\boldsymbol{h} + \alpha & = & d \end{array}\right.$$

Par hypothèse de récurrence $\underline{\mathbb{L}}$ est triangulaire inférieure à diagonale unité et $\underline{\mathbb{U}}$ triangulaire superieure inversible: le système précédant admet donc comme solution

$$\left\{\begin{array}{lcl} \boldsymbol{h} & = & \underline{\mathbb{L}}^{-1}\boldsymbol{f} \\ \boldsymbol{g} & = & \underline{\mathbb{U}}^{*-1}\boldsymbol{e} \\ \alpha & = & d - \boldsymbol{g}^*\boldsymbol{h} \end{array}\right.$$

On a donc construit une matrice $\mathbb{L}$ triangulaire inférieure à diagonale unité et une matrice $\mathbb{U}$ triangulaire superieure telles que $\mathbb{A} = \mathbb{L}\mathbb{U}$. Comme la matrice $\mathbb{A}$ est inversible, on en déduit que

$$\det \mathbb{A} = \det(\mathbb{L}\mathbb{U}) = \det \mathbb{L} \times \det \mathbb{U} \neq 0$$

On obtient alors $\det \mathbb{U} \neq 0$ ce qui entraine l'inversibilité de la matrice $\mathbb{U}$.

On vient de démontrer l'existence d'une factorisation $\mathbb{L}\mathbb{U}$ de la matrice $\mathbb{A}$. Pour démontrer l'unicité, on va supposer qu'il existe deux factorisations $\mathbb{L}\mathbb{U}$ de $\mathbb{A}$ i.e.

$$\mathbb{A} = \mathbb{L}_1\mathbb{U}_1 = \mathbb{L}_2\mathbb{U}_2.$$

avec $\mathbb{L}_1, \mathbb{L}_2$ matrices triangulaires inférieures à diagonale unité et $\mathbb{U}_1, \mathbb{U}_2$ matrices triangulaires supérieures (inversibles). En multipliant l'équation $\mathbb{L}_1\mathbb{U}_1 = \mathbb{L}_2\mathbb{U}_2$ à gauche par $\mathbb{L}_1^{-1}$ et à droite par $\mathbb{U}_2^{-1}$ on obtient

$$\mathbb{U}_1\mathbb{U}_2^{-1} = \mathbb{L}_1^{-1}\mathbb{L}_2. \tag{3.14}$$

La matrice $\mathbb{L}_1^{-1}\mathbb{L}_2$ est triangulaire inférieure à diagonale unité car produit de deux matrices triangulaires inférieures à diagonale unité. Elle est égale à la matrice $\mathbb{U}_1\mathbb{U}_2^{-1}$ qui elle est triangulaire supérieure (car prduit de deux matrices triangulaires supérieures). Donc $\mathbb{L}_1^{-1}\mathbb{L}_2$ est à la fois une matrice triangulaire supérieure et inférieure : elle est donc diagonale. Comme elle est à diagonale unité on en déduit que $\mathbb{L}_1^{-1}\mathbb{L}_2 = \mathbb{I}$ et donc $\mathbb{L}_1 = \mathbb{L}_2$. De l'équation (3.14), on tire alors $\mathbb{U}_1 = \mathbb{U}_2$. □

Exercice 3.1.7

Montrer que si $\mathbb{A}$ inversible admet une factorisation $\mathbb{L}\mathbb{U}$ alors toutes ses sous-matrices principales sont inversibles.

Remarque 3.8 Si la matrice $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ est inversible et si ses sous-matrices principales ne sont pas toutes inversibles, il est possible par des permutations préalables de lignes de la matrice de se ramener à une matrice telle que ses sous-matrices principales soient inversibles.

Théorème 3.9: Factorisation $\mathbb{L}\mathbb{U}$ avec permutations ★★★★★

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice inversible. Il existe une matrice $\mathbb{P}$, produit de matrices de permutation, une matrice $\mathbb{L} \in \mathcal{M}_n(\mathbb{C})$ triangulaire inférieure à diagonale unité et une matrice $\mathbb{U} \in \mathcal{M}_n(\mathbb{C})$ triangulaire supérieure telles ques

$$\mathbb{P}\mathbb{A} = \mathbb{L}\mathbb{U}. \tag{3.15}$$

Corollaire 3.10: ★★★☆☆

Si $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ est une matrice hermitienne définie positive alors elle admet une unique factorisation $\mathbb{L}\mathbb{U}$.

Preuve. (*indication*) Si la matrice $\mathbb{A}$ est hermitienne définie positive alors toutes ses sous-matrices principales sont définies positives et donc inversibles. □

Résolution d'un système linéaire par factorisation $\mathbb{L}\mathbb{U}$

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$ une matrice dont les sous-matrices principales sont inversibles et $\boldsymbol{b} \in \mathbb{K}^n$. On veut résoudre le système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$. Pour celà, grâce au théorème 3.7, on obtient :

Trouver $\boldsymbol{x} \in \mathbb{K}^n$ tel que

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b}. \tag{3.16}$$

est équivalent à

Trouver $\boldsymbol{x} \in \mathbb{K}^n$ solution de

$$\mathbb{U}\boldsymbol{x} = \boldsymbol{y} \tag{3.17}$$

avec $\boldsymbol{y} \in \mathbb{K}^n$ solution de

$$\mathbb{L}\boldsymbol{y} = \boldsymbol{b}. \tag{3.18}$$

Ceci permet donc de découper le problème initial en trois sous-problèmes plus simples. De plus, ceux-ci peuvent se traiter de manière indépendante. On obtient alors l'algorithme suivant

Algorithme 3.9 Fonction RSLFactLU permettant de résoudre, par une factorisation $\mathbb{LU}$, le système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ où $\mathbb{A}$ une matrice de $\mathcal{M}_n(\mathbb{R})$ définie positive et $\boldsymbol{b} \in \mathbb{R}^n$.

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{R})$ dont les sous-matrices principales sont inversibles définie positive,
$\boldsymbol{b}$: vecteur de $\mathbb{R}^n$.
Résultat : $\boldsymbol{x}$: vecteur de $\mathbb{R}^n$.

1: **Fonction** $\boldsymbol{x} \leftarrow$ RSLFactLU ($\mathbb{A}, \boldsymbol{b}$)
2: $[\mathbb{L}, \mathbb{U}] \leftarrow$ FactLU($\mathbb{A}$) ▷ Factorisation $\mathbb{LU}$
3: $\boldsymbol{y} \leftarrow$ RSLTriInf($\mathbb{L}, \boldsymbol{b}$) ▷ Résolution du système $\mathbb{L}\boldsymbol{y} = \boldsymbol{b}$
4: $\boldsymbol{x} \leftarrow$ RSLTriSup($\mathbb{U}, \boldsymbol{y}$) ▷ Résolution du système $\mathbb{U}\boldsymbol{x} = \boldsymbol{y}$
5: **Fin Fonction**

Il nous faut donc écrire la fonction FactLU (les deux fonctions RSLTriInf et RSLTriSup ayant déjà été écrites).

Détermination des matrices $\mathbb{L}$ et $\mathbb{U}$

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$ une matrice dont les sous-matrices principales sont inversibles. D'après le théorème 3.7, il existe une unique matrice $\mathbb{L} \in \mathcal{M}_n(\mathbb{K})$ triangulaire inférieure avec $L_{i,i} = 1$, $\forall i \in [\![1, n]\!]$, et une unique matrice $\mathbb{L} \in \mathcal{M}_n(\mathbb{K})$ triangulaire supérieure telles que

$$\mathbb{A} = \mathbb{LU} \tag{3.19}$$

c'est à dire

$$\begin{pmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & \ddots & \dots \\ A_{n,1} & \dots & A_{n,n} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ L_{2,1} & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ L_{n,1} & \dots & L_{n,n-1} & 1 \end{pmatrix} \begin{pmatrix} U_{1,1} & \dots & \dots & U_{n,1} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & U_{n,n} \end{pmatrix}. \tag{3.20}$$

Pour déterminer les matrices $\mathbb{L}$ et $\mathbb{U}$, on remarque que la 1ère ligne de $\mathbb{L}$ est déjà déterminée. On peut alors l'utiliser pour calculer la première ligne de $\mathbb{U}$: $\forall j \in [\![1, n]\!]$

$$\begin{aligned} A_{1,j} &= \sum_{k=1}^{n} L_{1,k} U_{k,j} \\ &= L_{1,1} U_{1,j} \text{ car } \mathbb{L} \text{ triangulaire inférieure} \\ &= U_{1,j} \end{aligned}$$

On a donc

$$U_{1,j} = A_{1,j}, \quad \forall j \in [\![1, n]\!]. \tag{3.21}$$

La première colonne de $\mathbb{U}$ est aussi déterminée, on peut alors l'utiliser pour calculer la première colonne de $\mathbb{L}$: $\forall i \in [\![2, n]\!]$

$$\begin{aligned} A_{i,1} &= \sum_{k=1}^{n} L_{i,k} U_{1,j} \\ &= L_{i,1} U_{1,1} \text{ car } \mathbb{U} \text{ triangulaire supérieure} \end{aligned}$$

On peut démontrer, de part les hypothèses sur la matrice $\mathbb{A}$, que $U_{1,1} \neq 0$ et alors

$$L_{1,j} = A_{i,j}/U_{1,1}, \quad \forall i \in [\![2, n]\!]. \tag{3.22}$$

La première ligne de $\mathbb{U}$ et la première colonne de $\mathbb{L}$ sont donc déterminées par les formules (3.21) et (3.22).

Par récurrence, on suppose connues les $i-1$ premières lignes de $\mathbb{U}$ et les $i-1$ premières colonnes de $\mathbb{L}$. On va montrer que l'on peux expliciter la i-ème ligne de $\mathbb{U}$ et la i-ème colonne de $\mathbb{L}$.
En effet, $\forall j \in [\![i, n]\!]$, on a

$$\begin{aligned} A_{i,j} &= \sum_{k=1}^{n} L_{i,k}U_{k,j} \\ &= \sum_{k=1}^{i-1} L_{i,k}U_{k,j} + L_{i,i}U_{i,j} + \sum_{k=i+1}^{n} L_{i,k}U_{k,j} \\ &= \sum_{k=1}^{i-1} L_{i,k}U_{k,j} + L_{i,i}U_{i,j} \text{ car } \mathbb{L} \text{ triangulaire inférieure} \end{aligned}$$

Dans l'expression $\sum_{k=1}^{i-1} L_{i,k}U_{k,j}$ tous les termes sont connus (hypothèse de récurrence) et $L_{i,i} = 1$. On en déduit,

$$\text{Pour } i \text{ allant de 1 à } n: \quad U_{i,j} = \begin{cases} A_{i,j} - \sum_{k=1}^{i-1} L_{i,k}U_{k,j}, & \forall j \in [\![i, n]\!]. \\ 0, & \forall j \in [\![1, i-1]\!]. \end{cases} \tag{3.23}$$

Maintenant, on calcule, $\forall j \in [\![i+1, n]\!]$,

$$\begin{aligned} A_{j,i} &= \sum_{k=1}^{n} L_{j,k}U_{k,i} \\ &= \sum_{k=1}^{i-1} L_{j,k}U_{k,i} + L_{j,i}U_{i,i} + \sum_{k=i+1}^{n} L_{j,k}U_{k,i} \\ &= \sum_{k=1}^{i-1} L_{j,k}U_{k,i} + L_{j,i}U_{i,i} \text{ car } \mathbb{U} \text{ triangulaire supérieure} \end{aligned}$$

Dans l'expression $\displaystyle\sum_{k=1}^{i-1} L_{j,k}U_{k,i}$ tous les termes sont connus (hypothèse de récurrence). De plus $U_{i,i}$ est donné par (3.23) et on peut démontrer, de part les hypothèses sur la matrice $\mathbb{A}$, que $U_{i,i} \neq 0$. On a alors

$$\text{Pour } i \text{ allant de 1 à } n: \quad L_{j,i} = \begin{cases} 0, & \forall j \in [\![1, i-1]\!]. \\ 1, & j = i \\ \frac{1}{U_{i,i}} \left(A_{j,i} - \displaystyle\sum_{k=1}^{i-1} L_{j,k}U_{k,i} \right), & \forall j \in [\![i+1, n]\!], \end{cases} \tag{3.24}$$

On écrit en détail les raffinements successifs permettant d'aboutir à l'algorithme final de telle sorte que le passage entre deux raffinements successifs soit le plus compréhensible possible.

Algorithme 3.10 $\mathcal{R}_0$

1: Calculer les matrices $\mathbb{L}$ et $\mathbb{U}$

Algorithme 3.10 $\mathcal{R}_1$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: Calculer la ligne i de $\mathbb{U}$.
3: Calculer la colonne i de $\mathbb{L}$.
4: **Fin Pour**

Algorithme 3.10 $\mathcal{R}_1$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: Calculer la ligne i de $\mathbb{U}$.
3: Calculer la colonne i de $\mathbb{L}$.
4: **Fin Pour**

Algorithme 3.10 $\mathcal{R}_2$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
3: $U(i,j) \leftarrow 0$
4: **Fin Pour**
5: **Pour** $j \leftarrow i$ **à** n **faire**
6: $U_{i,j} \leftarrow A_{i,j} - \sum_{k=1}^{i-1} L_{i,k}U_{k,j}$
7: **Fin Pour**
8: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
9: $L_{j,i} \leftarrow 0$
10: **Fin Pour**
11: $L_{i,i} \leftarrow 1$
12: **Pour** $j \leftarrow i+1$ **à** n **faire**
13: $L_{j,i} \leftarrow \frac{1}{U_{i,i}} \left(A_{j,i} - \sum_{k=1}^{i-1} L_{j,k}U_{k,i} \right)$
14: **Fin Pour**
15: **Fin Pour**

Algorithme 3.10 $\mathcal{R}_2$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
3: $U(i,j) \leftarrow 0$
4: **Fin Pour**
5: **Pour** $j \leftarrow i$ **à** n **faire**
6: $U_{i,j} \leftarrow A_{i,j} - \sum_{k=1}^{i-1} L_{i,k}U_{k,j}$
7: **Fin Pour**
8: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
9: $L_{j,i} \leftarrow 0$
10: **Fin Pour**
11: $L_{i,i} \leftarrow 1$
12: **Pour** $j \leftarrow i+1$ **à** n **faire**
13: $L_{j,i} \leftarrow \frac{1}{U_{i,i}} \left(A_{j,i} - \sum_{k=1}^{i-1} L_{j,k}U_{k,i} \right)$
14: **Fin Pour**
15: **Fin Pour**

Algorithme 3.10 $\mathcal{R}_3$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
3: $U(i,j) \leftarrow 0$
4: **Fin Pour**
5: **Pour** $j \leftarrow i$ **à** n **faire**
6: $S_1 \leftarrow \sum_{k=1}^{i-1} L_{i,k}U_{k,j}$
7: $U_{i,j} \leftarrow A_{i,j} - S_1$
8: **Fin Pour**
9: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
10: $L_{j,i} \leftarrow 0$
11: **Fin Pour**
12: $L_{i,i} \leftarrow 1$
13: **Pour** $j \leftarrow i+1$ **à** n **faire**
14: $S_2 \leftarrow \sum_{k=1}^{i-1} L_{j,k}U_{k,i}$
15: $L_{j,i} \leftarrow \frac{1}{U_{i,i}} \left(A_{j,i} - S_2 \right).$
16: **Fin Pour**
17: **Fin Pour**

Algorithme 3.10 $\mathcal{R}_3$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
3: $U(i,j) \leftarrow 0$
4: **Fin Pour**
5: **Pour** $j \leftarrow i$ **à** n **faire**
6: $S_1 \leftarrow \sum_{k=1}^{i-1} L_{i,k}U_{k,j}$
7: $U_{i,j} \leftarrow A_{i,j} - S_1$
8: **Fin Pour**
9: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
10: $L_{j,i} \leftarrow 0$
11: **Fin Pour**
12: $L_{i,i} \leftarrow 1$
13: **Pour** $j \leftarrow i+1$ **à** n **faire**
14: $S_2 \leftarrow \sum_{k=1}^{i-1} L_{j,k}U_{k,i}$
15: $L_{j,i} \leftarrow \frac{1}{U_{i,i}}\left(A_{j,i} - S_2\right).$
16: **Fin Pour**
17: **Fin Pour**

Algorithme 3.10 $\mathcal{R}_4$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
3: $U(i,j) \leftarrow 0$
4: **Fin Pour**
5: **Pour** $j \leftarrow i$ **à** n **faire**
6: $S_1 \leftarrow 0$
7: **Pour** $k \leftarrow 1$ **à** $i-1$ **faire**
8: $S_1 \leftarrow S_1 + L_{i,k} * U_{k,j}$
9: **Fin Pour**
10: $U_{i,j} \leftarrow A_{i,j} - S_1$
11: **Fin Pour**
12: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
13: $L_{j,i} \leftarrow 0$
14: **Fin Pour**
15: $L_{i,i} \leftarrow 1$
16: **Pour** $j \leftarrow i+1$ **à** n **faire**
17: $S_2 \leftarrow 0$
18: **Pour** $k \leftarrow 1$ **à** $i-1$ **faire**
19: $S_2 \leftarrow S_2 + L_{j,k} * U_{k,i}$
20: **Fin Pour**
21: $L_{j,i} \leftarrow \frac{1}{U_{i,i}}\left(A_{j,i} - S_2\right).$
22: **Fin Pour**
23: **Fin Pour**

L'algorithme peut être amélioré, pour gagner en lisibilité... En effet, il est possible d'initialiser la matrice $\mathbb{U}$ par la matrice nulle et la matrice $\mathbb{L}$ par la matrice identitée, ce qui permet alors de supprimer les boucles $U(i,j) \leftarrow 0$ et $L(j,i) \leftarrow 0$ ainsi que la commande $L(i,i) \leftarrow 1$. On obtient alors l'algorithme final

Algorithme 3.10 Fonction FactLU permet de calculer les matrices $\mathbb{L}$ et $\mathbb{U}$ dites matrice de factorisation $\mathbb{LU}$ associée à la matrice $\mathbb{A}$, telle que

$$\mathbb{A} = \mathbb{LU}$$

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$ dont les sous-matrices principales sont inversibles.
Résultat : $\mathbb{L}$: matrice de $\mathcal{M}_n(\mathbb{K})$ triangulaire inférieure avec $L_{i,i} = 1,\ \forall i \in [\![1, n]\!]$
$\mathbb{U}$: matrice de $\mathcal{M}_n(\mathbb{K})$ triangulaire supérieure.

Fonction $[\mathbb{L}, \mathbb{U}] \leftarrow$ FactLU ($\mathbb{A}$)
  $\mathbb{U} \leftarrow \mathbb{O}_n$ ▷ $\mathbb{O}_n$ matrice nulle $n \times n$
  $\mathbb{L} \leftarrow \mathbb{I}_n$ ▷ $\mathbb{I}_n$ matrice identitée $n \times n$
  Pour $i \leftarrow 1$ **à** n **faire**
    Pour $j \leftarrow i$ **à** n **faire** ▷ Calcul de la ligne i de $\mathbb{U}$
      $S_1 \leftarrow 0$
      Pour $k \leftarrow 1$ **à** $i-1$ **faire**
        $S_1 \leftarrow S_1 + L(i,k) * U(k,j)$
      Fin Pour
      $U(i,j) \leftarrow A(i,j) - S_1$
    Fin Pour
    Pour $j \leftarrow i+1$ **à** n **faire** ▷ Calcul de la colonne i de $\mathbb{L}$
      $S_2 \leftarrow 0$
      Pour $k \leftarrow 1$ **à** $i-1$ **faire**
        $S_2 \leftarrow S_2 + L(j,k) * U(k,i)$
      Fin Pour
      $L(j,i) \leftarrow (A_{j,i} - S_2)/U(i,i)$.
    Fin Pour
  Fin Pour
Fin Fonction

Remarque 3.11 Pour optimiser en mémoire cette fonction, il est possible de stocker les matrices $\mathbb{L}$ et $\mathbb{U}$ dans une même matrice de $\mathcal{M}_n(\mathbb{K})$ ...

Pour faciliter la lecture d'un tel algorithme, il aurait pu être judiscieux d'utiliser deux fonctions intermédiaires FactLUligU et FactLUcolL qui à l'étape i de l'algorithme calculent respectivement la ligne i de $\mathbb{U}$ et la colonne i de $\mathbb{L}$.

Algorithme 3.11 Fonction FactLUligU permet de calculer la ligne i de $\mathbb{U}$ à partir de (3.23)

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$ dont les sous-matrices principales sont inversibles.
$\mathbb{L}$: matrice de $\mathcal{M}_n(\mathbb{K})$ dont les $i-1$ premières colonnes ont été calculées
$\mathbb{U}$: matrice de $\mathcal{M}_n(\mathbb{K})$ dont les $i-1$ premières lignes ont été calculées
Résultat : $\mathbb{U}$: matrice de $\mathcal{M}_n(\mathbb{K})$ dont les i premières lignes ont été calculées

Fonction $\mathbb{U} \leftarrow$ FactLUligU ($\mathbb{A}, \mathbb{L}, \mathbb{U}, i$)
  Pour $j \leftarrow 1$ **à** $i-1$ **faire**
    $U(i,j) \leftarrow 0$
  Fin Pour
  Pour $j \leftarrow i$ **à** n **faire**
    $S_1 \leftarrow 0$
    Pour $k \leftarrow 1$ **à** $i-1$ **faire**
      $S_1 \leftarrow S_1 + L(i,k) * U(k,j)$
    Fin Pour
    $U(i,j) \leftarrow A(i,j) - S_1$
  Fin Pour
Fin Fonction

Algorithme 3.12 Fonction FactLUcolL permet de calculer la colonne i de $\mathbb{L}$ à partir de (3.24)

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$ dont les sous-matrices principales sont inversibles.
$\mathbb{L}$: matrice de $\mathcal{M}_n(\mathbb{K})$ dont les $i-1$ premières colonnes ont été calculées
$\mathbb{U}$: matrice de $\mathcal{M}_n(\mathbb{K})$ dont les i premières lignes ont été calculées
Résultat : $\mathbb{L}$: matrice de $\mathcal{M}_n(\mathbb{K})$ dont les i premières colonnes ont été calculées

Fonction $\mathbb{L} \leftarrow$ FactLUcolL ($\mathbb{A}, \mathbb{L}, \mathbb{U}, i$)
  Pour $j \leftarrow 1$ **à** $i-1$ **faire**
    $L(j,i) \leftarrow 0$
  Fin Pour
  $L(i,i) \leftarrow 1$
  Pour $j \leftarrow i+1$ **à** n **faire**
    $S_2 \leftarrow 0$
    Pour $k \leftarrow 1$ **à** $i-1$ **faire**
      $S_2 \leftarrow S_2 + L(j,k) * U(k,i)$
    Fin Pour
    $L(j,i) \leftarrow (A_{j,i} - S_2)/U(i,i)$.
  Fin Pour
Fin Fonction

On a alors

Algorithme 3.13 Fonction FactLU permet de calculer les matrices $\mathbb{L}$ et $\mathbb{U}$ dites matrice de factorisation $\mathbb{LU}$ associée à la matrice $\mathbb{A}$, telle que

$$\mathbb{A} = \mathbb{LU}$$

en utilisant des fonctions intermédiaires.

```
1: Fonction [L, U] ← FactLU ( A )
2:   Pour i ← 1 à n faire
3:     U ← FactLUligU(A, L, U, i)
4:     L ← FactLUcolL(A, L, U, i)
5:   Fin Pour
6: Fin Fonction
```

3.1.5 Factorisation $\mathbb{LDL}^*$

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice hermitienne inversible admettant une factorisation $\mathbb{LU}$. On note $\mathbb{D}$ la matrice diagonale inversible définie par $\mathbb{D} = \operatorname{diag} \mathbb{U}$ (i.e. $D_{i,i} = U_{i,i} \neq 0$, $\forall i \in [\![1, n]\!]$) et $\mathbb{R} = \mathbb{D}^{-1}\mathbb{U}$. La matrice $\mathbb{R}$ est donc matrice triangulaire supérieure à diagonale unité car

$$R_{i,i} = (\mathbb{D}^{-1}\mathbb{U})_{i,i} = \sum_{k=1}^{n} (\mathbb{D}^{-1})_{i,k} U_{k,i} = (\mathbb{D}^{-1})_{i,i} U_{i,i} = \frac{1}{U_{i,i}} U_{i,i} = 1.$$

On a alors

$$\mathbb{A} = \mathbb{LU} = \mathbb{LDD}^{-1}\mathbb{U} = \mathbb{LDR}.$$

De plus comme $\mathbb{A}$ est hermitienne on a

$$\mathbb{A}^* = \mathbb{A} \iff \mathbb{R}^*\mathbb{D}^*\mathbb{L}^* = \mathbb{LDR}$$

Ce qui donne

$$\mathbb{A} = \mathbb{R}^*(\mathbb{D}^*\mathbb{L}^*) = \mathbb{L}(\mathbb{DR})$$

et donc par unicité de la factorisation $\mathbb{LU}$ on a $\mathbb{R}^* = \mathbb{L}$ et $\mathbb{D}^*\mathbb{L}^* = \mathbb{DR}$. On obtient alors

$$\mathbb{R}^* = \mathbb{L} \text{ et } \mathbb{D}^* = \mathbb{D}$$

et on a le théorème suivant

Théorème 3.12: Factorisation $\mathbb{LDL}^*$ ★★☆☆☆

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice hermitienne inversible admettant une factorisation $\mathbb{LU}$. Alors $\mathbb{A}$ s'écrit sous la forme

$$\mathbb{A} = \mathbb{LDL}^* \tag{3.25}$$

où $\mathbb{D} = \operatorname{diag} \mathbb{U}$ est une matrice à coefficients réels.

Corollaire 3.13: ★★☆☆☆

Une matrice $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ admet une factorisation $\mathbb{LDL}^*$ avec $\mathbb{L} \in \mathcal{M}_n(\mathbb{C})$ matrice triangulaire inférieure à diagonale unité et $\mathbb{D} \in \mathcal{M}_n(\mathbb{R})$ matrice diagonale à coefficients diagonaux strictement positifs **si et seulement si** la matrice $\mathbb{A}$ est hermitienne définie positive.

Preuve. $\boxed{\Longrightarrow}$ Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ admettant une factorisation $\mathbb{LDL}^*$ avec $\mathbb{L} \in \mathcal{M}_n(\mathbb{C})$ matrice triangulaire inférieure à diagonale unité et $\mathbb{D} \in \mathcal{M}_n(\mathbb{R})$ matrice diagonale à coeffcients diagonaux strictement positifs.
La matrice $\mathbb{A}$ est alors hermitienne car

$$\mathbb{A}^* = (\mathbb{LDL}^*)^* = \mathbb{L}^{**}\mathbb{D}^*\mathbb{L}^* = \mathbb{LDL}^*.$$

De plus $\forall \boldsymbol{x} \in \mathbb{C}^n \backslash \{0\}$ on a

$$\langle \mathbb{A}\boldsymbol{x}, \boldsymbol{x} \rangle = \langle \mathbb{LDL}^*\boldsymbol{x}, \boldsymbol{x} \rangle = \langle \mathbb{DL}^*\boldsymbol{x}, \mathbb{L}^*\boldsymbol{x} \rangle$$

On pose $\boldsymbol{y} = \mathbb{L}^*\boldsymbol{x} \neq 0$ car $\boldsymbol{x} \neq 0$ et $\mathbb{L}^*$ inversible. On obtient alors

$$\langle \mathbb{A}\boldsymbol{x}, \boldsymbol{x}\rangle = \langle \mathbb{D}\boldsymbol{y}, \boldsymbol{y}\rangle = \sum_{i=1}^{n} D_{i,i}|y_i|^2 > 0$$

car $\mathbb{D}$ diagonale, $D_{i,i} > 0$, $\forall i \in [\![1, n]\!]$ et $\boldsymbol{y} \neq 0$.
La matrice hermitienne $\mathbb{A}$ est donc bien définie positive.

$\boxed{\Longleftarrow}$ Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice hermitienne définie positive.
D'après le Corollaire 3.10, page 76, la matrice $\mathbb{A}$ admet une unique factorisation $\mathbb{LU}$ et donc d'après le Théorème 3.13, page 82, la matrice hermitienne $\mathbb{A}$ peut s'écrire sous la forme $\mathbb{A} = \mathbb{LDL}^*$ où $\mathbb{D}$ est diagonale à coefficients réels et $\mathbb{L}$ triangulaire inférieure à diagonale unité. Il reste à démontrer que $D_{i,i} > 0$, $\forall i \in [\![1, n]\!]$.
Comme $\mathbb{A}$ est définie positive, on a $\forall \boldsymbol{x} \in \mathbb{C}^n\backslash\{0\}$, $\langle \mathbb{A}\boldsymbol{x}, \boldsymbol{x}\rangle > 0$. Or on a

$$\langle \mathbb{A}\boldsymbol{x}, \boldsymbol{x}\rangle = \langle \mathbb{LDL}^*\boldsymbol{x}, \boldsymbol{x}\rangle = \langle \mathbb{DL}^*\boldsymbol{x}, \mathbb{L}^*\boldsymbol{x}\rangle$$

On note $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_n\}$, la base canonique de $\mathbb{C}^n$ et on rappelle que $\forall i \in [\![1, n]\!]$, $\langle \mathbb{D}\boldsymbol{e}_i, \boldsymbol{e}_i\rangle = D_{i,i}$. Soit $i \in [\![1, n]\!]$. En choisissant $\boldsymbol{x} = (\mathbb{L}^*)^{-1}\boldsymbol{e}_i \neq 0$, on obtient alors

$$\langle \mathbb{DL}^*\boldsymbol{x}, \mathbb{L}^*\boldsymbol{x}\rangle = \langle \mathbb{D}\boldsymbol{e}_i, \boldsymbol{e}_i\rangle = D_{i,i} > 0.$$

□

3.1.6 Factorisation de Cholesky

Definition 3.14

Une **factorisation régulière de Cholesky** d'une matrice $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ est une factorisation $\mathbb{A} = \mathbb{BB}^*$ où $\mathbb{B}$ est une matrice triangulaire inférieure inversible.
Si les coefficients diagonaux de $\mathbb{B}$ sont positifs, on parle alors d'une **factorisation positive de Cholesky**.

Théorème 3.15: Factorisation de Cholesky ★★★★★

La matrice $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ admet une factorisation régulière de Cholesky **si et seulement si** la matrice $\mathbb{A}$ est hermitienne définie positive. Dans ce cas, elle admet une unique factorisation positive.

Preuve. $\boxed{\Longrightarrow}$ Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ admettant une factorisation régulière de Cholesky $\mathbb{A} = \mathbb{BB}^*$ avec $\mathbb{B}$ est une matrice triangulaire inférieure inversible.
La matrice $\mathbb{A}$ est hermitienne car

$$\mathbb{A}^* = (\mathbb{BB}^*)^* = (\mathbb{B}^*)^*\mathbb{B}^* = \mathbb{BB}^* = \mathbb{A}.$$

Soit $\boldsymbol{x} \in \mathbb{C}^n\backslash\{0\}$, on a

$$\langle \mathbb{A}\boldsymbol{x}, \boldsymbol{x}\rangle = \langle \mathbb{BB}^*\boldsymbol{x}, \boldsymbol{x}\rangle = \langle \mathbb{B}^*\boldsymbol{x}, \mathbb{B}^*\boldsymbol{x}\rangle = \|\mathbb{B}^*\boldsymbol{x}\|^2 > 0$$

car $\mathbb{B}^*\boldsymbol{x} \neq 0$ ($\mathbb{B}^*$ inversible et $\boldsymbol{x} \neq 0$). Donc la matrice $\mathbb{A}$ est bien hermitienne définie positive.

$\boxed{\Longleftarrow}$ Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice hermitienne définie positive.
D'après le Corollaire 3.13, page 82, il existe alors une matrice $\mathbb{L} \in \mathcal{M}_n(\mathbb{C})$ triangulaire inférieure à diagonale unité et une matrice $\mathbb{D} \in \mathcal{M}_n(\mathbb{R})$ diagonale à coefficient strictement positifs telles que

$$\mathbb{A} = \mathbb{LDL}^*.$$

On note $\mathbb{H} \in \mathcal{M}_n(\mathbb{R})$ une matrice diagonale inversible vérifiant $\mathbb{H}^2 = \mathbb{D}$ (i.e. $H_{i,i} = \pm\sqrt{D_{i,i}} \neq 0$, $\forall i \in [\![1, n]\!]$). On a alors

$$\mathbb{A} = \mathbb{LHHL}^* = (\mathbb{LH})(\mathbb{LH})^*$$

En posant $\mathbb{B} = \mathbb{L}\mathbb{H}$, la matrice $\mathbb{B}$ est bien triangulaire inférieure inversible car produit d'une matrice triangulaire inférieure inversible par une matrice diagonale inversible et on a $\mathbb{A} = \mathbb{B}\mathbb{B}^*$.

Montrons qu'une factorisation positive de Cholesky est unique.
Soient $\mathbb{B}_1$ et $\mathbb{B}_2$ deux factorisations positives de la matrice $\mathbb{A}$, on a donc

$$\mathbb{A} = \mathbb{B}_1\mathbb{B}_1^* = \mathbb{B}_2\mathbb{B}_2^*.$$

En multipliant à gauche par $\mathbb{B}_2^{-1}$ et à droite par $(\mathbb{B}_1^*)^{-1}$ cette équation on obtient

$$\mathbb{B}_2^{-1}\mathbb{B}_1 = \mathbb{B}_2^*(\mathbb{B}_1^*)^{-1} = \mathbb{B}_2^*(\mathbb{B}_1^{-1})^* = (\mathbb{B}_1^{-1}\mathbb{B}_2)^*$$

En notant $\mathbb{G} = \mathbb{B}_2^{-1}\mathbb{B}_1$, on tire de l'équation précédente

$$\mathbb{G} = (\mathbb{G}^{-1})^*. \tag{3.26}$$

On déduit de la (voir Proposition B.46, page 207), que l'inverse d'une matrice triangulaire inférieure à coefficients diagonaux réels strictement positifs est aussi une matrice triangulaire inférieure à coefficients diagonaux réels strictement positifs. De la (voir Proposition B.45, page 207), on obtient que le produit de matrices triangulaires inférieures à coefficients diagonaux réels strictement positifs reste triangulaire inférieure à coefficients diagonaux réels strictement positifs, on en déduit que les matrices $\mathbb{G} = \mathbb{B}_2^{-1}\mathbb{B}_1$ et $\mathbb{G}^{-1} = \mathbb{B}_1^{-1}\mathbb{B}_2$ sont triangulaires inférieures à coefficients diagonaux réels strictement positifs. Or l'équation (3.26) identifie la matrice triangulaire inférieure $\mathbb{G}$ à la matrice triangulaire supérieure $(\mathbb{G}^{-1})^*$: ce sont donc des matrices diagonales à coefficients diagonaux réels strictement positifs et on a alors $(\mathbb{G}^{-1})^* = \mathbb{G}^{-1}$. De l'équation (3.26), on obtient alors $\mathbb{G} = \mathbb{G}^{-1}$, c'est à dire $\mathbb{G} = \mathbb{I} = \mathbb{B}_2^{-1}\mathbb{B}_1$ et donc $\mathbb{B}_1 = \mathbb{B}_2$. □

Résolution d'un système linéaire par la factorisation de Cholesky

Pour commencer, avec la factorisation de Cholesky point de salut sans matrice hermitienne définie positive!

> ⚠ N'utiliser la factorisation de Cholesky pour la résolution d'un système linéaire que si la matrice du système est hermitienne définie positive.

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice hermitienne définie positive et $\boldsymbol{b} \in \mathbb{C}^n$. Grâce au théorème 3.15, on obtient :

Trouver $\boldsymbol{x} \in \mathbb{C}^n$ tel que

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b}. \tag{3.27}$$

est équivalent à

Trouver $\boldsymbol{x} \in \mathbb{C}^n$ solution de

$$\mathbb{B}^*\boldsymbol{x} = \boldsymbol{y} \tag{3.28}$$

avec $\mathbb{B}$ la matrice de factorisation positive de Cholesky de la matrice $\mathbb{A}$ avec $\boldsymbol{y} \in \mathbb{C}^n$ solution de

$$\mathbb{B}\boldsymbol{y} = \boldsymbol{b}. \tag{3.29}$$

On est donc ramené à

Algorithme 3.14 Algorithme de base permettant de résoudre, par une factorisation de Cholesky positive, le système linéaire

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$$

où $\mathbb{A}$ une matrice de $\mathcal{M}_n(\mathbb{C})$ hermitienne définie positive et $\boldsymbol{b} \in \mathbb{C}^n$.

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{C})$ hermitienne définie positive,
$\boldsymbol{b}$: vecteur de $\mathbb{C}^n$.
Résultat : $\boldsymbol{x}$: vecteur de $\mathbb{C}^n$.

1: Trouver la factorisation positive de Cholesky $\mathbb{B}$ de la matrice $\mathbb{A}$,
2: résoudre le système triangulaire inférieur $\mathbb{B}\boldsymbol{y} = \boldsymbol{b}$,
3: résoudre le système triangulaire supérieur $\mathbb{B}^*\boldsymbol{x} = \boldsymbol{y}$.

Ceci permet donc de découper le problème initial en trois sous-problèmes plus simples. De plus, ceux-ci peuvent se traiter de manière indépendante.

Algorithme 3.15 Fonction RSLCholesky permettant de résoudre, par une factorisation de Cholesky positive, le système linéaire

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$$

où $\mathbb{A}$ une matrice hermitienne de $\mathcal{M}_n(\mathbb{C})$ définie positive et $\boldsymbol{b} \in \mathbb{C}^n$.

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{C})$ hermitienne définie positive,
$\boldsymbol{b}$: vecteur de $\mathbb{C}^n$.
Résultat : $\boldsymbol{x}$: vecteur de $\mathbb{C}^n$.

1: **Fonction** $\boldsymbol{x} \leftarrow$ RSLCholesky ($\mathbb{A}, \boldsymbol{b}$)
2: $\mathbb{B} \leftarrow$ Cholesky($\mathbb{A}$) ▷ Factorisation positive de Cholesky
3: $\boldsymbol{y} \leftarrow$ RSLTriInf($\mathbb{B}, \boldsymbol{b}$) ▷ Résolution du système $\mathbb{B}\boldsymbol{y} = \boldsymbol{b}$
4: $\mathbb{U} \leftarrow$ MatAdjointe($\mathbb{B}$) ▷ Calcul de la matrice adjointe de $\mathbb{B}$
5: $\boldsymbol{x} \leftarrow$ RSLTriSup($\mathbb{U}, \boldsymbol{y}$) ▷ Résolution du système $\mathbb{B}^*\boldsymbol{x} = \boldsymbol{y}$
6: **Fin Fonction**

Il nous faut donc écrire la fonction Cholesky (les deux fonctions RSLTriInf et RSLTriSup ayant déjà été écrites et la fonction MatAdjointe étant simple à écrire).

Factorisation positive de Cholesky : écriture de l'algorithme

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice hermitienne définie positive. D'après le Théorème 3.15, il existe une unique matrice $\mathbb{B} \in \mathcal{M}_n(\mathbb{C})$ triangulaire inférieure avec $\mathrm{B}_{i,i} \in \mathbb{R}^{+*}$, $\forall i \in [\![1, n]\!]$, telle que

$$\mathbb{A} = \mathbb{B}\mathbb{B}^* \tag{3.30}$$

c'est à dire

$$\begin{pmatrix} \mathrm{A}_{1,1} & \dots & \mathrm{A}_{1,n} \\ \vdots & \ddots & \dots \\ \mathrm{A}_{n,1} & \dots & \mathrm{A}_{n,n} \end{pmatrix} = \begin{pmatrix} \mathrm{B}_{1,1} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 \\ \mathrm{B}_{n,1} & \dots & \dots & \mathrm{B}_{n,n} \end{pmatrix} \begin{pmatrix} \overline{\mathrm{B}_{1,1}} & \dots & \dots & \overline{\mathrm{B}_{n,1}} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \overline{\mathrm{B}_{n,n}} \end{pmatrix}. \tag{3.31}$$

Pour déterminer la matrice $\mathbb{B}$, on commence par calculer $\mathrm{B}_{1,1}$ (la 1ère ligne de $\mathbb{B}$ est donc déterminée) ce qui nous permet de calculer la 1ère colonne de $\mathbb{B}$.
Ensuite, on calcule $\mathrm{B}_{2,2}$ (la 2ème ligne de $\mathbb{B}$ est donc déterminée) ce qui nous permet de calculer la 2ème colonne de $\mathbb{B}$. Etc ...

On écrit en détail les raffinements successifs permettant d'aboutir à l'algorithme final de telle manière que le passage entre deux raffinements successifs soit le plus simple possible.

Algorithme 3.16 $\mathcal{R}_0$

1: Calculer la matrice $\mathbb{B}$

Algorithme 3.16 $\mathcal{R}_1$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: Calculer $\mathrm{B}_{i,i}$, connaissant les $i-1$ premières colonnes de $\mathbb{B}$.
3: Calculer la $i^{\text{ème}}$ colonne de $\mathbb{B}$.
4: **Fin Pour**

Pour calculer $\mathrm{B}_{i,i}$, connaissant les $i-1$ premières colonnes de $\mathbb{B}$. on utilise (3.30) :

$$\mathrm{A}_{i,i} = \sum_{j=1}^{n} \mathrm{B}_{i,j}(\mathbb{B}^*)_{j,i} = \sum_{j=1}^{n} |\mathrm{B}_{i,j}|^2,$$

et comme $\mathbb{B}$ est triangulaire inférieure on obtient

$$\mathrm{A}_{i,i} = \sum_{j=1}^{i} |\mathrm{B}_{i,j}|^2 = \sum_{j=1}^{i-1} |\mathrm{B}_{i,j}|^2 + |\mathrm{B}_{i,i}|^2.$$

On a donc

$$\mathrm{B}_{i,i} = \left(\mathrm{A}_{i,i} - \sum_{j=1}^{i-1} |\mathrm{B}_{i,j}|^2 \right)^{1/2}, \ \forall i \in [\![1, n]\!], \tag{3.32}$$

Comme les $i-1$ premières colonnes de $\mathbb{B}$ ont déjà été calculées, $\mathrm{B}_{i,i}$ est parfaitement déterminée par la formule précédente.

Remarque 3.16 Les hypothèses sur la matrice $\mathbb{A}$ permettent d'affirmer que, $\forall i \in [\![1, n]\!]$, $\mathrm{A}_{i,i} - \sum_{j=1}^{i-1} |\mathrm{B}_{i,j}|^2 > 0$ et donc $\mathrm{B}_{i,i} > 0$.

Pour calculer la $i^{\text{ème}}$ colonne de $\mathbb{B}$, il suffit de déterminer $\mathrm{B}_{j,i}$, $\forall j \in [\![i+1, n]\!]$. Pour celà on utilise (3.30)

$$\mathrm{A}_{j,i} = \sum_{k=1}^{n} \mathrm{B}_{j,k}(\mathbb{B}^*)_{k,i} = \sum_{k=1}^{n} \mathrm{B}_{j,k}\overline{\mathrm{B}_{i,k}}, \ \forall j \in [\![i+1, n]\!]$$

Comme $\mathbb{L}$ est triangulaire inférieure on obtient

$$\mathrm{A}_{j,i} = \sum_{k=1}^{i} \mathrm{B}_{j,k}\overline{\mathrm{B}_{i,k}} = \sum_{k=1}^{i-1} \mathrm{B}_{j,k}\overline{\mathrm{B}_{i,k}} + \mathrm{B}_{j,i}\overline{\mathrm{B}_{i,i}}, \ \forall j \in [\![i+1, n]\!]$$

Or $\mathrm{B}_{i,i} > 0$ vient d'être calculé et les $i-1$ premières colonnes de $\mathbb{B}$ ont déjà été calculées, ce qui donne

$$\mathrm{B}_{j,i} = \frac{1}{\mathrm{B}_{i,i}} \left(\mathrm{A}_{j,i} - \sum_{k=1}^{i-1} \mathrm{B}_{j,k}\overline{\mathrm{B}_{i,k}} \right), \ \forall j \in [\![i+1, n]\!] \tag{3.33}$$

$$\mathrm{B}_{j,i} = 0, \ \forall j \in [\![1, i-1]\!]. \tag{3.34}$$

Avec (3.32) et (3.33), on a

Algorithme 3.16 $\mathcal{R}_1$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: Calculer $\mathrm{B}_{i,i}$, connaissant les $i-1$ premières colonnes de $\mathbb{B}$.
3: Calculer la $i^{\text{ème}}$ colonne de $\mathbb{B}$.
4: **Fin Pour**

Algorithme 3.16 $\mathcal{R}_2$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: $\mathrm{B}_{i,i} \leftarrow \left(\mathrm{A}_{i,i} - \sum_{j=1}^{i-1} |\mathrm{B}_{i,j}|^2 \right)^{1/2}$
3: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
4: $\mathrm{B}_{j,i} \leftarrow 0$
5: **Fin Pour**
6: **Pour** $j \leftarrow i+1$ **à** n **faire**
7: $\mathrm{B}_{j,i} \leftarrow \frac{1}{\mathrm{B}_{i,i}} \left(\mathrm{A}_{j,i} - \sum_{k=1}^{i-1} \mathrm{B}_{j,k}\overline{\mathrm{B}_{i,k}} \right)$.
8: **Fin Pour**
9: **Fin Pour**

Algorithme 3.16 $\mathcal{R}_2$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: $\mathrm{B}_{i,i} \leftarrow \left(\mathrm{A}_{i,i} - \sum_{j=1}^{i-1} |\mathrm{B}_{i,j}|^2\right)^{1/2}$
3: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
4: $\mathrm{B}_{j,i} \leftarrow 0$
5: **Fin Pour**
6: **Pour** $j \leftarrow i+1$ **à** n **faire**
7: $\mathrm{B}_{j,i} \leftarrow \frac{1}{\mathrm{B}_{i,i}} \left(\mathrm{A}_{j,i} - \sum_{k=1}^{i-1} \mathrm{B}_{j,k}\overline{\mathrm{B}_{i,k}}\right).$
8: **Fin Pour**
9: **Fin Pour**

Algorithme 3.16 $\mathcal{R}_3$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: $S_1 \leftarrow \sum_{j=1}^{i-1} |\mathrm{B}_{i,j}|^2$
3: $\mathrm{B}_{i,i} \leftarrow (\mathrm{A}_{i,i} - S_1)^{1/2}$
4: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
5: $\mathrm{B}_{j,i} \leftarrow 0$
6: **Fin Pour**
7: **Pour** $j \leftarrow i+1$ **à** n **faire**
8: $S_2 \leftarrow \sum_{k=1}^{i-1} \mathrm{B}_{j,k}\overline{\mathrm{B}_{i,k}}$
9: $\mathrm{B}_{j,i} \leftarrow \frac{1}{\mathrm{B}_{i,i}} (\mathrm{A}_{j,i} - S_2).$
10: **Fin Pour**
11: **Fin Pour**

Algorithme 3.16 $\mathcal{R}_3$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: $S_1 \leftarrow \sum_{j=1}^{i-1} |\mathrm{B}_{i,j}|^2$
3: $\mathrm{B}_{i,i} \leftarrow (\mathrm{A}_{i,i} - S_1)^{1/2}$
4: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
5: $\mathrm{B}_{j,i} \leftarrow 0$
6: **Fin Pour**
7: **Pour** $j \leftarrow i+1$ **à** n **faire**
8: $S_2 \leftarrow \sum_{k=1}^{i-1} \mathrm{B}_{j,k}\overline{\mathrm{B}_{i,k}}$
9: $\mathrm{B}_{j,i} \leftarrow \frac{1}{\mathrm{B}_{i,i}} (\mathrm{A}_{j,i} - S_2).$
10: **Fin Pour**
11: **Fin Pour**

Algorithme 3.16 $\mathcal{R}_4$

1: **Pour** $i \leftarrow 1$ **à** n **faire**
2: $S_1 \leftarrow 0$
3: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
4: $S_1 \leftarrow S_1 + |\mathrm{B}_{i,j}|^2$
5: **Fin Pour**
6: $\mathrm{B}_{i,i} \leftarrow (\mathrm{A}_{i,i} - S_1)^{1/2}$
7: **Pour** $j \leftarrow 1$ **à** $i-1$ **faire**
8: $\mathrm{B}_{j,i} \leftarrow 0$
9: **Fin Pour**
10: **Pour** $j \leftarrow i+1$ **à** n **faire**
11: $S_2 \leftarrow 0$
12: **Pour** $k \leftarrow 1$ **à** $i-1$ **faire**
13: $S_2 \leftarrow S_2 + \mathrm{B}_{j,k}\overline{\mathrm{B}_{i,k}}$
14: **Fin Pour**
15: $\mathrm{B}_{j,i} \leftarrow \frac{1}{\mathrm{B}_{i,i}} (\mathrm{A}_{j,i} - S_2).$
16: **Fin Pour**
17: **Fin Pour**

On obtient alors l'algorithme final

Algorithme 3.16 Fonction CHOLESKY permettant de calculer la matrice $\mathbb{B}$, dites matrice de factorisation positive de Cholesky associée à la matrice $\mathbb{A}$, telle que $\mathbb{A} = \mathbb{B}\mathbb{B}^*$.

Données : A : matrice de $\mathcal{M}_n(\mathbb{C})$ hermitienne définie positive.
Résultat : B : matrice de $\mathcal{M}_n(\mathbb{C})$ triangulaire inférieure avec $\mathrm{B}(i,i) > 0,\ \forall i \in [\![1,n]\!]$

Fonction B ← CHOLESKY (A)
  Pour $i \leftarrow 1$ **à** n **faire**
    $S_1 \leftarrow 0$
    Pour $j \leftarrow 1$ **à** $i-1$ **faire**
      $S_1 \leftarrow S_1 + |\mathrm{B}(i,j)|^2$
    Fin Pour
    $\mathrm{B}(i,i) \leftarrow \mathrm{SQRT}(\mathrm{A}(i,i) - S_1)$
    Pour $j \leftarrow 1$ **à** $i-1$ **faire**
      $\mathrm{B}(j,i) \leftarrow 0$
    Fin Pour
    Pour $j \leftarrow i+1$ **à** n **faire**
      $S_2 \leftarrow 0$
      Pour $k \leftarrow 1$ **à** $i-1$ **faire**
        $S_2 \leftarrow S_2 + \mathrm{B}(j,k) * \overline{\mathrm{B}(i,k)}$
      Fin Pour
      $\mathrm{B}(j,i) \leftarrow (\mathrm{A}(j,i) - S_2)/\mathrm{B}(i,i).$
    Fin Pour
  Fin Pour
Fin Fonction

3.1.7 Factorisation $\mathbb{QR}$

La tranformation de Householder

Definition 3.17: Matrice élémentaire de Householder

Soit $\boldsymbol{u} \in \mathbb{C}^n$ tel que $\|\boldsymbol{u}\|_2 = 1$. On appelle **matrice élémentaire de Householder** la matrice $\mathbb{H}(\boldsymbol{u}) \in \mathcal{M}_n(\mathbb{C})$ définie par

$$\mathbb{H}(\boldsymbol{u}) = \mathbb{I} - 2\boldsymbol{u}\boldsymbol{u}^*. \tag{3.35}$$

Propriété 3.18

Toute matrice élémentaire de Householder est hermitienne et unitaire.

Preuve. Pour simplifier, on note $\mathbb{H} = \mathbb{H}(\boldsymbol{u})$. Cette matrice est hermitienne car

$$\mathbb{H}^* = (\mathbb{I} - 2\boldsymbol{u}\boldsymbol{u}^*)^* = \mathbb{I} - 2(\boldsymbol{u}\boldsymbol{u}^*)^* = \mathbb{I} - 2\boldsymbol{u}\boldsymbol{u}^* = \mathbb{H}.$$

Montrons qu'elle est unitaire (i.e. $\mathbb{H}^*\mathbb{H} = \mathbb{I}$). On a

$$\begin{aligned}\mathbb{H}^*\mathbb{H} = \mathbb{H}\mathbb{H} &= (\mathbb{I} - 2\boldsymbol{u}\boldsymbol{u}^*)(\mathbb{I} - 2\boldsymbol{u}\boldsymbol{u}^*)\\ &= \mathbb{I} - 4\boldsymbol{u}\boldsymbol{u}^* + 4\boldsymbol{u}\boldsymbol{u}^*\boldsymbol{u}\boldsymbol{u}^*.\end{aligned}$$

Or on a $\boldsymbol{u}^*\boldsymbol{u} = \|\boldsymbol{u}\|_2 = 1$ par hypothèse et donc

$$\mathbb{H}^*\mathbb{H} = \mathbb{I} - 4\boldsymbol{u}\boldsymbol{u}^* + 4\boldsymbol{u}(\boldsymbol{u}^*\boldsymbol{u})\boldsymbol{u}^* = \mathbb{I}.$$

□

Propriété 3.19

Soient $\boldsymbol{x} \in \mathbb{K}^n$ et $\boldsymbol{u} \in \mathbb{K}^n$, $\|\boldsymbol{u}\|_2 = 1$. On note $\boldsymbol{x}_{\|} = \mathrm{proj}_{\boldsymbol{u}}(\boldsymbol{x}) \overset{\text{def}}{=} \langle \boldsymbol{u}, \boldsymbol{x} \rangle \boldsymbol{u}$ et $\boldsymbol{x}_{\perp} = \boldsymbol{x} - \boldsymbol{x}_{\|}$. On a alors

$$\mathbb{H}(\boldsymbol{u})(\boldsymbol{x}_{\perp} + \boldsymbol{x}_{\|}) = \boldsymbol{x}_{\perp} - \boldsymbol{x}_{\|}. \tag{3.36}$$

et

$$\mathbb{H}(\boldsymbol{u})\boldsymbol{x} = \boldsymbol{x}, \quad \text{si } \langle \boldsymbol{x}, \boldsymbol{u} \rangle = 0. \tag{3.37}$$

Preuve. On note que par construction $\langle \boldsymbol{u}, \boldsymbol{x}_{\perp} \rangle = 0$. On a

$$\begin{aligned}
\mathbb{H}(\boldsymbol{u})(\boldsymbol{x}_{\perp} + \boldsymbol{x}_{\|}) &= (\mathbb{I} - 2\boldsymbol{u}\boldsymbol{u}^*)(\boldsymbol{x}_{\perp} + \boldsymbol{x}_{\|}) = \boldsymbol{x}_{\perp} + \boldsymbol{x}_{\|} - 2\boldsymbol{u} \underbrace{\boldsymbol{u}^* \boldsymbol{x}_{\perp}}_{=0} - 2\boldsymbol{u}\boldsymbol{u}^* \boldsymbol{x}_{\|} \\
&= \boldsymbol{x}_{\perp} + \boldsymbol{x}_{\|} - 2\boldsymbol{u}\boldsymbol{u}^* \boldsymbol{u} \langle \boldsymbol{u}, \boldsymbol{x} \rangle = \boldsymbol{x}_{\perp} + \boldsymbol{x}_{\|} - 2\boldsymbol{u} \underbrace{\boldsymbol{u}^* \boldsymbol{u}}_{=1} \boldsymbol{u}^* \boldsymbol{x} \\
&= \boldsymbol{x}_{\perp} + \boldsymbol{x}_{\|} - 2\boldsymbol{u}\boldsymbol{u}^* \boldsymbol{x} = \boldsymbol{x}_{\perp} + \boldsymbol{x}_{\|} - 2\boldsymbol{x}_{\|} \\
&= \boldsymbol{x}_{\perp} - \boldsymbol{x}_{\|}.
\end{aligned}$$

Si $\langle \boldsymbol{x}, \boldsymbol{u} \rangle = 0$ alors $\boldsymbol{x}_{\|} = 0$ et $\boldsymbol{x} = \boldsymbol{x}_{\perp}$. □

Théorème 3.20

Soient $\boldsymbol{a}$, $\boldsymbol{b}$ deux vecteurs non colinéaires de $\mathbb{C}^n$ avec $\|\boldsymbol{b}\|_2 = 1$. Soit $\alpha \in \mathbb{C}$ tel que $|\alpha| = \|\boldsymbol{a}\|_2$ et $\arg \alpha = -\arg \langle \boldsymbol{a}, \boldsymbol{b} \rangle \ [\pi]$. On a alors

$$\mathbb{H}\left(\frac{\boldsymbol{a} - \alpha \boldsymbol{b}}{\|\boldsymbol{a} - \alpha \boldsymbol{b}\|_2} \right) \boldsymbol{a} = \alpha \boldsymbol{b}. \tag{3.38}$$

Preuve. (voir Exercice 3.1.8, page 89) □

Exercice 3.1.8

Soient $\boldsymbol{a}$ et $\boldsymbol{b}$ deux vecteurs non colinéaires de $\mathbb{C}^n$ avec $\|\boldsymbol{b}\|_2 = 1$. On va chercher $\alpha \in \mathbb{C}$ et $\boldsymbol{u} \in \mathbb{C}^n$ vérifiant

$$\mathbb{H}(\boldsymbol{u})\boldsymbol{a} = \alpha \boldsymbol{b}. \tag{3.39}$$

Q. 1 1. *Montrer que si* α *vérifie* (3.39) *alors* $|\alpha| = \|\boldsymbol{a}\|_2$.

2. *Montrer que si* $\arg \alpha = -\arg(\langle \boldsymbol{a}, \boldsymbol{b} \rangle) \ [\pi]$ *alors* $\alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle \in \mathbb{R}$.

Q. 2 *Soient* α *et* $\boldsymbol{u}$ *vérifiant* (3.39).

1. *Montrer que*

$$|\langle \boldsymbol{u}, \boldsymbol{a} \rangle|^2 = \frac{\langle \boldsymbol{a}, \boldsymbol{a} \rangle - \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{2} \tag{3.40}$$

2. *Montrer que si* $\arg \alpha = -\arg(\langle \boldsymbol{a}, \boldsymbol{b} \rangle) \ [\pi]$ *alors* $\langle \boldsymbol{a}, \boldsymbol{a} \rangle - \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle \in \mathbb{R}^{*+}$.

3. *En déduire que*

$$\boldsymbol{u} = \frac{1}{2\lambda}(\boldsymbol{a} - \alpha \boldsymbol{b}), \quad \text{avec } \lambda = \pm \left(\frac{\langle \boldsymbol{a}, \boldsymbol{a} \rangle - \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle}{2} \right)^{1/2} \tag{3.41}$$

Correction Exercice 3.1.8 On pose $\mathbb{H} = \mathbb{H}(\boldsymbol{u})$ pour alléger les notations.

Q. 1 1. On a

$$\begin{aligned}
\|\boldsymbol{a}\|_2^2 = \langle \boldsymbol{a}, \boldsymbol{a} \rangle &= \langle \mathbb{H}^* \mathbb{H} \boldsymbol{a}, \boldsymbol{a} \rangle \quad \text{car } \mathbb{H} \text{ unitaire} \\
&= \langle \mathbb{H}\boldsymbol{a}, \mathbb{H}\boldsymbol{a} \rangle \quad \text{par définition du produit scalaire} \\
&= \|\mathbb{H}\boldsymbol{a}\|_2^2 = \|\alpha \boldsymbol{b}\|_2^2 = |\alpha|^2 \|\boldsymbol{b}\|_2^2 = |\alpha|^2.
\end{aligned}$$

2. On a par définition de l'argument $\alpha = |\alpha| e^{\imath \arg \alpha}$ et $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = |\langle \boldsymbol{a}, \boldsymbol{b} \rangle| e^{\imath \arg(\langle \boldsymbol{a}, \boldsymbol{b} \rangle)}$ ce qui donne
$$\alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle = |\alpha| |\langle \boldsymbol{a}, \boldsymbol{b} \rangle| e^{\imath(\arg \alpha + \arg(\langle \boldsymbol{a}, \boldsymbol{b} \rangle))} \tag{3.42}$$
et donc $\alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle$ est réel si $\arg \alpha + \arg(\langle \boldsymbol{a}, \boldsymbol{b} \rangle) = 0 \ [\pi]$.

Q. 2 1. On a
$$\begin{aligned} \mathbb{H}(\boldsymbol{u})\boldsymbol{a} = \alpha \boldsymbol{b} &\iff (\mathbb{I} - 2\boldsymbol{u}\boldsymbol{u}^*)\boldsymbol{a} = \alpha \boldsymbol{b} \\ &\iff \boldsymbol{a} - 2\boldsymbol{u}(\boldsymbol{u}^* \boldsymbol{a}) = \alpha \boldsymbol{b} \end{aligned}$$
et donc
$$\boldsymbol{a} - 2\langle \boldsymbol{u}, \boldsymbol{a} \rangle \boldsymbol{u} = \alpha \boldsymbol{b} \tag{3.43}$$
En effectuant le produit scalaire avec $\boldsymbol{a}$ de cette dernière équation, on obtient
$$\langle \boldsymbol{a}, \boldsymbol{a} \rangle - 2\langle \boldsymbol{u}, \boldsymbol{a} \rangle \langle \boldsymbol{a}, \boldsymbol{u} \rangle = \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle$$
ce qui prouve (3.40).

2. On a montré en Q.1 que $\alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle \in \mathbb{R}$ et donc $\langle \boldsymbol{a}, \boldsymbol{a} \rangle - \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle \in \mathbb{R}$. Il reste donc à montrer que $\langle \boldsymbol{a}, \boldsymbol{a} \rangle - \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle > 0$.
   - Si $\arg \alpha = -\arg(\langle \boldsymbol{a}, \boldsymbol{b} \rangle) + \pi \ [2\pi]$, alors de (3.42) on obtient $\alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle \leqslant 0$ et donc $\langle \boldsymbol{a}, \boldsymbol{a} \rangle - \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle \geqslant \|\boldsymbol{a}\|_2 > 0$ car $\boldsymbol{a} \neq 0$.
   - Si $\arg \alpha = -\arg(\langle \boldsymbol{a}, \boldsymbol{b} \rangle) \ [2\pi]$, alors de (3.42) on obtient $\alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle \geqslant 0$.
     Comme les vecteurs $\boldsymbol{a}$ et $\boldsymbol{b}$ ne sont pas colinéaires, on a inégalité stricte dans Cauchy-Schwarz :
     $$|\langle \boldsymbol{a}, \boldsymbol{b} \rangle| < \|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2 = \|\boldsymbol{a}\|_2 .$$
     On obtient donc
     $$0 \leqslant \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle \leqslant |\alpha| |\langle \boldsymbol{a}, \boldsymbol{b} \rangle| < |\alpha| \|\boldsymbol{a}\|_2 = \|\boldsymbol{a}\|_2^2$$
     Attention, dans ce cas $\langle \boldsymbol{a}, \boldsymbol{a} \rangle - \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle$ peut-être très petit.

3. De (3.43), on en déduit immédiatement (3.41).
   Vérifions que $\|\boldsymbol{u}\|_2 = 1$. On a
   $$\|\boldsymbol{u}\|_2^2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle = \frac{1}{4|\lambda|^2} \langle \boldsymbol{a} - \alpha \boldsymbol{b}, \boldsymbol{a} - \alpha \boldsymbol{b} \rangle$$
   Or
   $$\begin{aligned} \langle \boldsymbol{a} - \alpha \boldsymbol{b}, \boldsymbol{a} - \alpha \boldsymbol{b} \rangle &= \langle \boldsymbol{a}, \boldsymbol{a} \rangle - \overline{\alpha} \langle \boldsymbol{b}, \boldsymbol{a} \rangle - \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle + |\alpha|^2 \langle \boldsymbol{b}, \boldsymbol{b} \rangle = \|\boldsymbol{a}\|_2^2 - \overline{\alpha} \langle \boldsymbol{b}, \boldsymbol{a} \rangle - \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle + |\alpha|^2 \\ &= 2\|\boldsymbol{a}\|_2^2 - \overline{\alpha} \langle \boldsymbol{b}, \boldsymbol{a} \rangle - \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle \\ &= 2\|\boldsymbol{a}\|_2^2 - 2\alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle \quad \text{car } \alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle = \overline{(\alpha \langle \boldsymbol{a}, \boldsymbol{b} \rangle)} = \overline{\alpha} \langle \boldsymbol{b}, \boldsymbol{a} \rangle \in \mathbb{R} \end{aligned}$$
   De plus
   $$\begin{aligned} 4|\lambda|^2 &= 2(\langle \boldsymbol{a}, \boldsymbol{a} \rangle - \alpha \langle \boldsymbol{b}, \boldsymbol{a} \rangle) \in \mathbb{R} \\ &= 2\|\boldsymbol{a}\|_2^2 - 2\alpha \langle \boldsymbol{b}, \boldsymbol{a} \rangle \end{aligned}$$

◇

Exercice 3.1.9

Soient $\boldsymbol{a}$ et $\boldsymbol{b}$ deux vecteurs non nuls et non colinéaires de $\mathbb{C}^n$ avec $\|\boldsymbol{b}\|_2 = 1$.

Q. 1 *Ecrire la fonction algorithmique* HOUSEHOLDER *permettant de retourner une matrice de Householder* $\mathbb{H}$ *et* $\alpha \in \mathbb{C}$ *tels que* $\mathbb{H}(\boldsymbol{u})\boldsymbol{a} = \alpha \boldsymbol{b}$. *Le choix du* α *est fait par le paramètre* δ *(0 ou 1) de telle sorte que* $\arg \alpha = -\arg(\langle \boldsymbol{a}, \boldsymbol{b} \rangle) + \delta\pi$ *avec* $|\alpha| = \|\boldsymbol{a}\|_2$.
Des fonctions comme DOT$(\boldsymbol{a}, \boldsymbol{b})$ *(produit scalaire de deux vecteurs),* NORM$(\boldsymbol{a})$ *(norme 2 d'un vecteur),* ARG(z) *(argument d'un nombre complexe),* MATPROD$(\mathbb{A}, \mathbb{B})$ *(produit de deux matrices),* CTRANSPOSE$(\mathbb{A})$ *(adjoint d'une matrice), ... pourront être utilisées*

Q. 2 *Proposer un programme permettant de tester cette fonction. On pourra utiliser la fonction*

VECRAND(n) *retournant un vecteur aléatoire de* $\mathbb{C}^n$, *les parties réelles et imaginaires de chacune de ses composantes étant dans* $]0,1[$ *(loi uniforme).*

Q. 3 *Proposer un programme permettant de vérifier que* $\delta = 1$ *est le "meilleur" choix.*

Correction Exercice 3.1.9 Soient $\boldsymbol{a}$ et $\boldsymbol{b}$ deux vecteurs non nuls et non colinéaires de $\mathbb{C}^n$.

Q. 1 Les données du problème sont $\boldsymbol{a}$, $\boldsymbol{b}$ et δ. On veut calculer α et la matrice $\mathbb{H}(\boldsymbol{u})$.

Algorithme 3.17 Calcul du α et de la matrice de Householder $\mathbb{H}(\boldsymbol{u})$ telle que $\mathbb{H}(\boldsymbol{u})\boldsymbol{a} = \alpha\boldsymbol{b}$.

Données : $\boldsymbol{a}, \boldsymbol{b}$: deux vecteurs de $\mathbb{C}^n$ non nuls et non colinéaires.
δ : 0 ou 1, permet de déterminer α.
Résultat : $\mathbb{H}$: matrice de Householder dans $\mathcal{M}_n(\mathbb{C})$,
α : nombre complexe, de module $\|\boldsymbol{a}\|_2$ et d'argument $-\arg(\langle \boldsymbol{a}, \boldsymbol{b}\rangle) + \delta\pi$.

```
Fonction [H, α] ← HOUSEHOLDER ( a, b, δ )
    ab ← DOT(a, b)                          ▷ DOT produit scalaire dans ℂ.
    α ← NORM(a) * exp(i * (δ * π − ARG(ab)))
    u ← a − α * b
    u ← u/NORM(u)
    H ← EYE(n) − 2 * MATPROD(u, CTRANSPOSE(u))
Fin Fonction
```

Q. 2
```
n ← 100
a ← VECRAND(n)
b ← VECRAND(n)
b ← b/NORM(b, 2)
[H, α] ← HOUSEHOLDER(a, b, 0)
error ← NORM(H * a − α * b, 2)
```

Q. 3
```
n ← 100
a ← VECRAND(n)
b ← a + 1e − 6 * VECRAND(n)
b ← b/NORM(b, 2)
[H1, α1] ← HOUSEHOLDER(a, b, 1)
[H0, α0] ← HOUSEHOLDER(a, b, 0)
error0 ← NORM(H0 * a − α0 * b, 2)/(1 + ABS(α0))
error1 ← NORM(H1 * a − α1 * b, 2)/(1 + ABS(α1))
```

◇

Si l'on souhaite calculer le produit matrice-vecteur $\mathbb{H}(\boldsymbol{u})\boldsymbol{x}$, il n'est pas nécessaire de calculer explicitement la matrice $\mathbb{H}(\boldsymbol{u})$. En effet, on pose $\boldsymbol{v} = 2\lambda\boldsymbol{u} = \boldsymbol{a} - \alpha\boldsymbol{b}$ et $\beta = 2\lambda^2 = \|a\|_2^2 - \alpha\langle \boldsymbol{a}, \boldsymbol{b}\rangle > 0$. On choisi α de manière à maximiser β pour éviter une division par un nombre trop petit. On prend donc

$$\alpha = \|a\|_2 \, e^{\imath(\pi - \arg(\langle \boldsymbol{a}, \boldsymbol{b}\rangle))}$$

On obtient alors

$$\mathbb{H}(\boldsymbol{u})\boldsymbol{x} = \boldsymbol{x} - \frac{1}{\beta}\boldsymbol{v}\boldsymbol{v}^*\boldsymbol{x} = \boldsymbol{x} - \frac{\langle \boldsymbol{v}, \boldsymbol{x}\rangle}{\beta}\boldsymbol{v}. \tag{3.44}$$

Corollaire 3.21

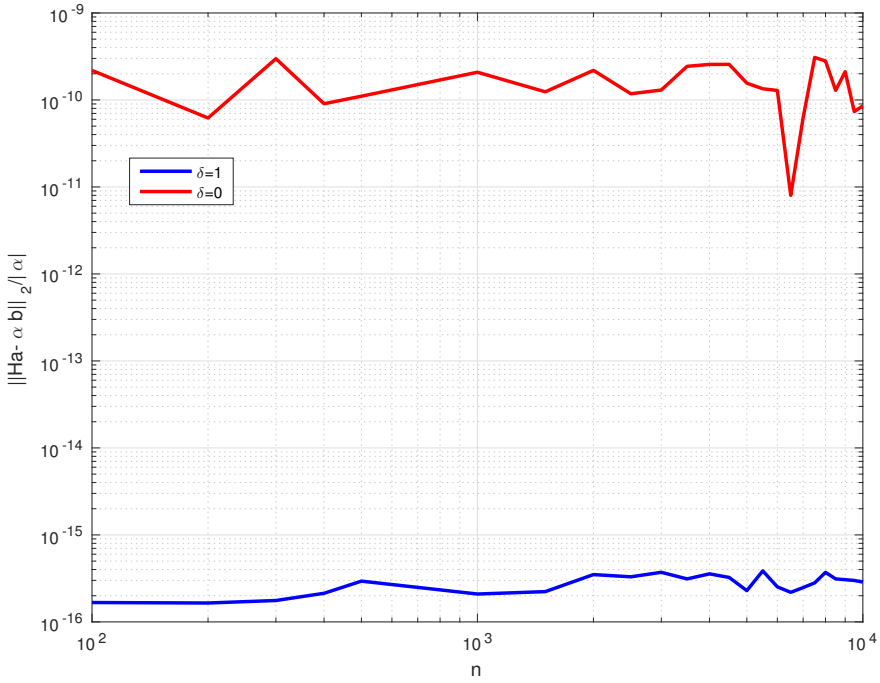


Figure 3.1: Choix de α dans HOUSEHOLDER : erreur relative en norme L_2

Soit $\boldsymbol{a} \in \mathbb{C}^n$ avec $a_1 \neq 0$ et $\exists j \in [\![2, n]\!]$ tel que $a_j \neq 0$. Soient $\theta = \arg a_1$ et

$$\boldsymbol{u}_\pm = \frac{\boldsymbol{a} \pm \|\boldsymbol{a}\|_2 \, e^{\imath\theta} \boldsymbol{e}_1}{\|\boldsymbol{a} \pm \|\boldsymbol{a}\|_2 \, e^{\imath\theta} \boldsymbol{e}_1\|}$$

Alors

$$\mathbb{H}(\boldsymbol{u}_\pm)\boldsymbol{a} = \mp \|\boldsymbol{a}\|_2 \, e^{\imath\theta} \boldsymbol{e}_1 \tag{3.45}$$

où $\boldsymbol{e}_1$ désigne le premier vecteur de la base canonique de $\mathbb{C}^n$.

Preuve. On va utiliser le Théorème 3.20.

On pose $\alpha = \pm \|\boldsymbol{a}\|_2 \, e^{\imath\theta}$ et $\boldsymbol{b} = \boldsymbol{e}_1$. Comme $\arg(\langle \boldsymbol{a}, \boldsymbol{b} \rangle) = \arg \overline{a_1} = -\arg a_1 = -\theta$, on a $\arg \alpha = \theta \ [\pi] = -\arg(\langle \boldsymbol{a}, \boldsymbol{b} \rangle) \ [\pi]$.

Les autres hypothèses du Théorème 3.20 sont vérifiées puisque le vecteur $\boldsymbol{a}$ n'est pas colinéaire à $\boldsymbol{e}_1$ et que $\|\boldsymbol{e}_1\|_2 = 1$. On a donc

$$\mathbb{H}\left(\frac{\boldsymbol{a} \pm \|\boldsymbol{a}\|_2 \, e^{\imath\theta} \boldsymbol{e}_1}{\|\boldsymbol{a} \pm \|\boldsymbol{a}\|_2 \, e^{\imath\theta} \boldsymbol{e}_1\|}\right) \boldsymbol{a} = \mp \|\boldsymbol{a}\|_2 \, e^{\imath\theta} \boldsymbol{e}_1.$$

□

Sur le même principe que l'écriture algébrique de la méthode de Gauss, nous allons transformer la matrice $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$ en une matrice triangulaire supérieure à l'aide de matrices de Householder. On a le théorème

Théorème 3.22

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice. Il existe une matrice unitaire $\mathbb{Q} \in \mathcal{M}_n(\mathbb{C})$ produit d'au plus $n-1$ matrices de Householder et une matrice triangulaire supérieure $\mathbb{R} \in \mathcal{M}_n(\mathbb{C})$ telles que

$$\mathbb{A} = \mathbb{Q}\mathbb{R}. \tag{3.46}$$

Si $\mathbb{A}$ est réelle alors $\mathbb{Q}$ et $\mathbb{R}$ sont aussi réelles et l'on peut choisir $\mathbb{Q}$ de telle sorte que les coefficients diagonaux de $\mathbb{R}$ soient positifs. De plus, si $\mathbb{A}$ est inversible alors la factorisation est unique.

Exercice 3.1.10

Soit $\mathbb{B} \in \mathcal{M}_{m+n}(\mathbb{K})$ la matrice bloc

$$\mathbb{B} = \left(\begin{array}{c|c} \mathbb{B}_{1,1} & \mathbb{B}_{1,2} \\ \hline \mathbb{0} & \mathbb{S} \end{array}\right)$$

où $\mathbb{B}_{1,1} \in \mathcal{M}_m(\mathbb{K})$ et $\mathbb{S} \in \mathcal{M}_n(\mathbb{K})$. On note $\boldsymbol{s} \in \mathbb{K}^n$ le premier vecteur colonne de $\mathbb{S}$ et on suppose que $\boldsymbol{s} \neq 0$ et $\boldsymbol{s}$ non colinéaire à $\boldsymbol{e}_1^n$ premier vecteur de la base canonique de $\mathbb{K}^n$.

Q. 1 *1. Montrer qu'il existe une matrice de Householder $\underline{\mathbb{H}} = \mathbb{H}(\underline{\boldsymbol{u}}) \in \mathcal{M}_n(\mathbb{K})$ et $\alpha \in \mathbb{K}^*$ tel que*

$$\underline{\mathbb{H}}\mathbb{S} = \left(\begin{array}{c|ccc} \pm\alpha & \bullet & \cdots & \bullet \\ \hline 0 & \bullet & \cdots & \bullet \\ \vdots & \vdots & & \vdots \\ 0 & \bullet & \cdots & \bullet \end{array}\right).$$

2. On note $\boldsymbol{u} \in \mathbb{K}^{m+n}$, le vecteur défini par $u_i = 0$, $\forall i \in [\![1, m]\!]$ et $u_{m+i} = \underline{u_i}$, $\forall i \in [\![1, n]\!]$. Montrer que

$$\mathbb{H}(\boldsymbol{u})\mathbb{B} = \left(\begin{array}{c|c} \mathbb{B}_{1,1} & \mathbb{B}_{1,2} \\ \hline \mathbb{0} & \underline{\mathbb{H}}\mathbb{S} \end{array}\right).$$

Soient $k \in [\![0, n-1]\!]$ et $\mathbb{A}^{[k]} \in \mathcal{M}_n(\mathbb{K})$ la matrice bloc définie par

$$\mathbb{A}^{[k]} = \left(\begin{array}{c|c} \mathbb{R}^{[k]} & \mathbb{F}^{[k]} \\ \hline \mathbb{0} & \underline{\mathbb{A}^{[k]}} \end{array}\right)$$

où $\mathbb{R}^{[k]}$ est une matrice triangulaire supérieure d'ordre k et $\underline{\mathbb{A}^{[k]}}$ une matrice d'ordre $n-k$.

Q. 2 *1. Sous certaines hypothèses, montrer qu'il existe une matrice de Householder $\mathbb{H}^{[k+1]}$ telle que $\mathbb{H}^{[k+1]}\mathbb{A}^{[k]} = \mathbb{A}^{[k+1]}$.*

2. Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$. Montrer qu'il existe une matrice unitaire $\mathbb{Q} \in \mathcal{M}_n(\mathbb{K})$, produit d'au plus $n-1$ matrices de Housholder, et une matrice triangulaire supérieure $\mathbb{R}$ telles que $\mathbb{A} = \mathbb{QR}$.

3. Montrer que si $\mathbb{A}$ est réelle alors les coefficient diagonaux de $\mathbb{R}$ peuvent être choisis positifs ou nuls.

4. Montrer que si $\mathbb{A}$ est réelle inversible alors la factorisation $\mathbb{QR}$, avec $\mathbb{R}$ à coefficient diagonaux strictement positifs, est unique.

Correction Exercice 3.1.10

Q. 1 1. D'après le (voir Corollaire 3.21, page 91) avec $\boldsymbol{a} = \boldsymbol{s}$, en posant $\alpha = \pm \|\boldsymbol{s}\|_2 e^{\imath \arg s_1}$ et

$$\underline{\boldsymbol{u}} = \frac{\boldsymbol{s} - \alpha \boldsymbol{e}_1^n}{\|\boldsymbol{s} - \alpha \boldsymbol{e}_1^n\|}$$

on obtient $\mathbb{H}(\underline{\boldsymbol{u}}) = \alpha \boldsymbol{e}_1^n$.
On pose $\underline{\mathbb{H}} = \mathbb{H}(\underline{\boldsymbol{u}})$. On a alors sous forme bloc

$$\underline{\mathbb{H}}\mathbb{S} = \underline{\mathbb{H}} \left(\begin{array}{c|ccc} \boldsymbol{s} & \bullet & \cdots & \bullet \\ & \vdots & & \vdots \\ & \bullet & \cdots & \bullet \end{array}\right) = \left(\begin{array}{c|ccc} \pm\alpha & \bullet & \cdots & \bullet \\ \hline 0 & \bullet & \cdots & \bullet \\ \vdots & \vdots & & \vdots \\ 0 & \bullet & \cdots & \bullet \end{array}\right)$$

2. On a $\boldsymbol{u} = \begin{pmatrix} \mathbf{0}_m \\ \hdashline \underline{\boldsymbol{u}} \end{pmatrix}$ et

$$\begin{aligned}\mathbb{H}(\boldsymbol{u}) = \mathbb{I} - 2\boldsymbol{u}\boldsymbol{u}^* &= \left(\begin{array}{c:c} \mathbb{I}_m & \mathbb{O}_{m,n} \\ \hdashline \mathbb{O}_{n,m} & \mathbb{I}_n \end{array}\right) - 2\left(\begin{array}{c} \mathbf{0}_m \\ \hdashline \underline{\boldsymbol{u}} \end{array}\right)\left(\begin{array}{c:c} \mathbf{0}_m^* & \underline{\boldsymbol{u}}^* \end{array}\right) \\ &= \left(\begin{array}{c:c} \mathbb{I}_m & \mathbb{O}_{m,n} \\ \hdashline \mathbb{O}_{n,m} & \mathbb{I}_n \end{array}\right) - 2\left(\begin{array}{c:c} \mathbb{O}_m & \mathbb{O}_{m,n} \\ \hdashline \mathbb{O}_{n,m} & \underline{\boldsymbol{u}}\underline{\boldsymbol{u}}^* \end{array}\right) \\ &= \left(\begin{array}{c:c} \mathbb{I}_m & \mathbb{O}_{m,n} \\ \hdashline \mathbb{O}_{n,m} & \mathbb{I}_n - 2\underline{\boldsymbol{u}}\underline{\boldsymbol{u}}^* \end{array}\right) = \left(\begin{array}{c:c} \mathbb{I}_m & \mathbb{O}_{m,n} \\ \hdashline \mathbb{O}_{n,m} & \underline{\mathbb{H}} \end{array}\right)\end{aligned}$$

Ce qui donne

$$\mathbb{H}(\boldsymbol{u})\mathbb{B} = \left(\begin{array}{c:c} \mathbb{I}_m & \mathbb{O}_{m,n} \\ \hdashline \mathbb{O}_{n,m} & \underline{\mathbb{H}} \end{array}\right)\left(\begin{array}{c:c} \mathbb{B}_{1,1} & \mathbb{B}_{1,2} \\ \hdashline \mathbb{O} & \mathbb{S} \end{array}\right) = \left(\begin{array}{c:c} \mathbb{B}_{1,1} & \mathbb{B}_{1,2} \\ \hdashline \mathbb{O} & \underline{\mathbb{H}}\mathbb{S} \end{array}\right).$$

Q. 2 1. On note $\underline{\boldsymbol{s}} \in \mathbb{K}^{n-k}$ le premier vecteur colonne de $\underline{\mathbb{A}^{[k]}}$. et $\boldsymbol{u} = \begin{pmatrix} \mathbf{0}_k \\ \hdashline \underline{\boldsymbol{s}} \end{pmatrix}$. D'après la question précédente si $\boldsymbol{s} \neq 0$ et $\boldsymbol{s}$ non colinéaire à $\boldsymbol{e}_1^{n-k}$ premier vecteur de la base canonique de $\mathbb{K}^{n-k}$ alors il existe une matrice de Householder $\mathbb{H}^{[k+1]} = \mathbb{H}(\boldsymbol{u})$ et $\alpha \in \mathbb{K}^*$ tels que

$$\mathbb{A}^{[k+1]} \overset{\text{def}}{=} \mathbb{H}^{[k+1]}\mathbb{A}^{[k]} = \left(\begin{array}{c:c} \mathbb{R}^{[k]} & \mathbb{F}^{[k]} \\ \hdashline \mathbb{O} & \left(\begin{array}{c:ccc} \pm\alpha & \bullet & \cdots & \bullet \\ \hdashline 0 & \bullet & \cdots & \bullet \\ \vdots & \vdots & & \vdots \\ 0 & \bullet & \cdots & \bullet \end{array}\right) \end{array}\right) = \left(\begin{array}{c:c} \mathbb{R}^{[k+1]} & \mathbb{F}^{[k+1]} \\ \hdashline \mathbb{O} & \underline{\mathbb{A}^{[k+1]}} \end{array}\right)$$

On peut remarquer que si $\boldsymbol{s} = 0$ ou $\boldsymbol{s}$ colinéaire à $\boldsymbol{e}_1^{n-k}$ alors $\mathbb{A}^{[k]}$ est déjà sous la forme $\mathbb{A}^{[k+1]}$ et donc $\mathbb{H}^{[k+1]} = \mathbb{I}$.

2. il suffit d'appliquer itérativement le résultat précédent $n-1$ fois en posant $\mathbb{A}^{[0]} = \mathbb{A}$ et $\mathbb{A}^{[k+1]} = \mathbb{H}^{[k+1]}\mathbb{A}^{[k]}$ où $\mathbb{H}^{[k+1]}$ est soit une matrice de Householder soit la matrice identité. Par construction la matrice $\mathbb{A}^{[n-1]}$ est triangulaire supérieure et l'on a

$$\mathbb{A}^{[n-1]} = \mathbb{H}^{[n-1]} \times \cdots \times \mathbb{H}^{[1]}\mathbb{A}$$

On pose $\mathbb{H} = \mathbb{H}^{[n-1]} \times \cdots \times \mathbb{H}^{[1]}$ et $\mathbb{R} = \mathbb{A}^{[n-1]}$. La matrice $\mathbb{H}$ est unitaire car produit de matrices unitaires. On note $\mathbb{Q} = \mathbb{H}^*$ On a

$$\mathbb{Q} = \mathbb{H}^{[1]} \times \cdots \times \mathbb{H}^{[n-1]}$$

car les matrices de Householder et matrice identité sont unitaires et hermitiennes.

3. Si $\mathbb{A}$ est réelle alors par construction $\mathbb{Q}$ et $\mathbb{R}$ sont réelles. Les coefficients diagonaux peuvent alors être choisi positif lors de la construction de chaque matrice de Householder.

4. Pour montrer l'unicité d'une telle factorisation, on note $\mathbb{Q}_1$, $\mathbb{Q}_2$, deux matrices orthogonales et $\mathbb{R}_1$, $\mathbb{R}_2$, deux matrices triangulaires à coefficients diagonaux positifs telles que

$$\mathbb{A} = \mathbb{Q}_1\mathbb{R}_1 = \mathbb{Q}_2\mathbb{R}_2.$$

Comme $\mathbb{A}$ est inversible les coefficients diagonaux de $\mathbb{R}_1$ et $\mathbb{R}_2$ sont strictement positifs. On a alors

$$\mathbb{I} = \mathbb{A}\mathbb{A}^{-1} = \mathbb{Q}_1\mathbb{R}_1\mathbb{R}_2^{-1}\mathbb{Q}_2^{-1}$$

et donc

$$\mathbb{Q}_1^{-1}\mathbb{Q}_2 = \mathbb{R}_1\mathbb{R}_2^{-1} \overset{\text{def}}{=} \mathbb{T}.$$

Comme $\mathbb{Q}_1$ est orthogonale on a $\mathbb{T} = \mathbb{Q}_1^{\mathsf{t}}\mathbb{Q}_2$ et

$$\mathbb{T}^{\mathsf{t}}\mathbb{T} = (\mathbb{Q}_1^{\mathsf{t}}\mathbb{Q}_2)^{\mathsf{t}}\mathbb{Q}_1^{\mathsf{t}}\mathbb{Q}_2 = \mathbb{Q}_2^{\mathsf{t}}\mathbb{Q}_1\mathbb{Q}_1^{\mathsf{t}}\mathbb{Q}_2 = \mathbb{I}.$$

La matrice $\mathbb{T}$ est donc orthogonal. De plus $\mathbb{T} = \mathbb{R}_1\mathbb{R}_2^{-1}$ est une matrice triangulaire supérieure à coefficients diagonaux strictement positifs puisque produit de triangulaire supérieure à coefficients diagonaux strictement positifs. La matrice $\mathbb{I}$ étant symétrique définie positive, d'après le Théorème 3.15 (factorisation positive de Cholesky) il existe une unique matrice $\mathbb{L}$ triangulaire inférieure à coefficients diagonaux strictement positifs telle que $\mathbb{L}\mathbb{L}^t = \mathbb{I}$. Cette matrice $\mathbb{L}$ est évidemment la matrice identité. On en déduit que $\mathbb{T} = \mathbb{L}^t = \mathbb{I}$ et donc $\mathbb{Q}_1 = \mathbb{Q}_2$ et $\mathbb{R}_1 = \mathbb{R}_2$.

◇

Exercice 3.1.11: Algorithmique

Q. 1 *Ecrire une fonction* FactQR *permettant de calculer la factorisation $\mathbb{QR}$ d'une matrice $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$.*
On pourra utiliser la fonction Householder *(voir Exercice 3.1.9, page 90).*

Q. 2 *Ecrire un programme permettant de tester cette fonction.*

Correction Exercice 3.1.11

Q. 1 L'objectif est de déterminer les matrices $\mathbb{Q}$, matrice unitaire, et $\mathbb{R}$ matrice triangulaire supérieure telle que $\mathbb{A} = \mathbb{QR}$.

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$.
Résultat : $\mathbb{Q}$: matrice unitaire de $\mathcal{M}_n(\mathbb{K})$.
$\mathbb{R}$: matrice triangulaire supérieure de $\mathcal{M}_n(\mathbb{K})$.

On rappelle la technique utilisée dans la correction de l'exercice 3.1.10 pour déterminer l'ensemble des matrices de Householder permettant de transformer la matrice $\mathbb{A}$ en une matrice triangulaire supérieure. On pose

$$\mathbb{A}^{[0]} = \mathbb{A}, \quad \mathbb{A}^{[k+1]} = \mathbb{H}^{[k+1]}\mathbb{A}^{[k]}, \ \forall k \in [\![0, n-2]\!]$$

où $\mathbb{H}^{[k+1]}$ est soit une matrice de Householder soit la matrice identité. Plus précisement, on note $\underline{\boldsymbol{s}} \in \mathbb{K}^{n-k}$ le vecteur composé des $n-k$ dernières composantes de la $k+1$-ème colonne de $\mathbb{A}^{[k]}$ et $\boldsymbol{a} = \begin{pmatrix} \mathbf{0}_k \\ \hdashline \underline{\boldsymbol{s}} \end{pmatrix}$.

- Si $\underline{s}_1 = 0$ ou $\underline{\boldsymbol{s}}$ colinéaire à $\boldsymbol{e}_1^{n-k}$ premier vecteur de la base canonique de $\mathbb{K}^{n-k}$ alors

$$\mathbb{H}^{[k+1]} = \mathbb{I}.$$

  En notant $\boldsymbol{e}_{k+1}^n$ le $k+1$-ème vecteur de la base canonique de $\mathbb{K}^n$, cette matrice peut-être calculée avec la fonction Householder par

$$[\mathbb{H}^{[k+1]}, \alpha] \leftarrow \text{Householder}(\boldsymbol{a}, \boldsymbol{e}_{k+1}^n, 1)$$

- sinon $\mathbb{H}^{[k+1]} = \mathbb{I}$.

On a vu que dans ce cas $\mathbb{A}^{[n-1]}$ est triangulaire supérieure. On pose $\mathbb{H} = \mathbb{H}^{[n-1]} \times \cdots \times \mathbb{H}^{[1]}$ qui est une matrice unitaire. On a alors $\mathbb{R} = \mathbb{A}^{[n-1]} = \mathbb{H}\mathbb{A}$ et $\mathbb{Q} = \mathbb{H}^*$.

Algorithme 3.18 $\mathcal{R}_0$

1: Calculer $\mathbb{Q}$ et $\mathbb{R}$

Algorithme 3.18 $\mathcal{R}_1$

1: $\mathbb{H} \leftarrow \mathbb{H}^{[n-1]} \times \cdots \times \mathbb{H}^{[1]}$
2: $\mathbb{R} \leftarrow \mathbb{H} * \mathbb{A}$
3: $\mathbb{Q} \leftarrow \mathbb{H}^*$

Algorithme 3.18 $\mathcal{R}_1$

1: $\mathbb{H} \leftarrow \mathbb{H}^{[n-1]} \times \cdots \times \mathbb{H}^{[1]}$
2: $\mathbb{R} \leftarrow \mathbb{H} * \mathbb{A}$
3: $\mathbb{Q} \leftarrow \mathbb{H}^*$

Algorithme 3.18 $\mathcal{R}_2$

1: $\mathbb{H} \leftarrow \mathbb{I}$
2: $\mathbb{A}^{[0]} \leftarrow \mathbb{A}$
3: **Pour** $k \leftarrow 0$ **à** $n-2$ **faire**
4: Calculer $\mathbb{H}^{[k+1]}$ à partir de $\mathbb{A}^{[k]}$
5: $\mathbb{A}^{[k+1]} \leftarrow \mathbb{H}^{[k+1]} * \mathbb{A}^{[k]}$
6: $\mathbb{H} \leftarrow \mathbb{H}^{[k+1]} * \mathbb{H}$
7: **Fin Pour**
8: $\mathbb{R} \leftarrow \mathbb{H} * \mathbb{A}$ ▷ ou $\mathbb{R} \leftarrow \mathbb{A}^{[n-1]}$
9: $\mathbb{Q} \leftarrow \mathbb{H}^*$

Algorithme 3.18 $\mathcal{R}_2$

1: $\mathbb{H} \leftarrow \mathbb{I}$
2: **Pour** $k \leftarrow 0$ **à** $n-2$ **faire**
3: Calculer $\mathbb{H}^{[k+1]}$ à partir de $\mathbb{A}^{[k]}$
4: $\mathbb{A}^{[k+1]} \leftarrow \mathbb{H}^{[k+1]} * \mathbb{A}^{[k]}$
5: $\mathbb{H} \leftarrow \mathbb{H}^{[k+1]} * \mathbb{H}$
6: **Fin Pour**
7: $\mathbb{R} \leftarrow \mathbb{A}^{[n-1]}$
8: $\mathbb{Q} \leftarrow \mathbb{H}^*$

Algorithme 3.18 $\mathcal{R}_3$

1: $\mathbb{H} \leftarrow \mathbb{I}$
2: **Pour** $k \leftarrow 0$ **à** $n-2$ **faire**
3: $\boldsymbol{a} \leftarrow [\mathbf{0}_k; \mathbb{A}^{[k]}(k+1:n, k+1)]$
4: $[\mathbb{H}^{[k+1]}, \alpha] \leftarrow$ Householder$(\boldsymbol{a}, \boldsymbol{e}^n_{k+1}, 1)$
5: $\mathbb{A}^{[k+1]} \leftarrow \mathbb{H}^{[k+1]} * \mathbb{A}^{[k]}$
6: $\mathbb{H} \leftarrow \mathbb{H}^{[k+1]}\mathbb{H}$
7: **Fin Pour**
8: $\mathbb{R} \leftarrow \mathbb{A}^{[n-1]}$
9: $\mathbb{Q} \leftarrow \mathbb{H}^*$

Ici, l'opérateur $[\bullet;\bullet]$ est l'opérateur de concaténation de deux vecteurs.

Algorithme 3.18 Fonction FactQR

Données : $\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$.
Résultat : $\mathbb{Q}$: matrice unitaire de $\mathcal{M}_n(\mathbb{K})$.
$\mathbb{R}$: matrice triangulaire supérieure de $\mathcal{M}_n(\mathbb{K})$.

1: **Fonction** $[\mathbb{Q}, \mathbb{R}] \leftarrow$ FactQR ($\mathbb{A}$)
2: $\mathbb{H} \leftarrow \mathbb{I}$
3: $\mathbb{R} \leftarrow \mathbb{A}$
4: **Pour** $k \leftarrow 0$ **à** $n-2$ **faire**
5: $\boldsymbol{a} \leftarrow [\mathbf{0}_k; \mathbb{R}(k+1:n, k+1)]$
6: $[\mathbb{S}, \alpha] \leftarrow$ Householder$(\boldsymbol{a}, \boldsymbol{e}^n_{k+1}, 1)$
7: $\mathbb{R} \leftarrow \mathbb{S} * \mathbb{R}$
8: $\mathbb{H} \leftarrow \mathbb{S} * \mathbb{H}$
9: **Fin Pour**
10: $\mathbb{Q} \leftarrow \mathbb{H}^*$
11: **Fin Fonction**

Q. 2

◇

3.2 Normes vectorielles et normes matricielles

3.2.1 Normes vectorielles

Definition 3.23

Une **norme** sur un espace vectoriel V est une application $\|\bullet\| : V \to \mathbb{R}^+$ qui vérifie les propriétés

suivantes

- $\|\boldsymbol{v}\| = 0 \Longleftrightarrow \boldsymbol{v} = 0$,
- $\|\alpha\boldsymbol{v}\| = |\alpha| \|\boldsymbol{v}\|, \ \forall \alpha \in \mathbb{K}, \ \forall \boldsymbol{v} \in V$,
- $\|\boldsymbol{u}+\boldsymbol{v}\| \leqslant \|\boldsymbol{u}\| + \|\boldsymbol{v}\|, \ \forall (\boldsymbol{u},\boldsymbol{v}) \in V^2$ (inégalité triangulaire).

Une norme sur V est également appelée **norme vectorielle** . On appelle **espace vectoriel normé** un espace vectoriel muni d'une norme.

Les trois normes suivantes sont les plus couramment utilisées :

$$\|\boldsymbol{v}\|_1 = \sum_{i=1}^{n} |v_i|$$
$$\|\boldsymbol{v}\|_2 = \left(\sum_{i=1}^{n} |v_i|^2 \right)^{1/2}$$
$$\|\boldsymbol{v}\|_\infty = \max_{i \in [\![1,n]\!]} |v_i| .$$

Proposition 3.24

Soit $\boldsymbol{v} \in \mathbb{K}^n$. Pour tout nombre réel $p \geqslant 1$, l'application $\|\bullet\|_p$ définie par

$$\|\boldsymbol{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p}$$

est une norme sur $\mathbb{K}^n$.

Lemme 3.25: Inégalité de Cauchy-Schwarz

$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{K}^n$

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leqslant \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 . \tag{3.47}$$

Cette inégalité s'appelle l'**inégalité de Cauchy-Schwarz**. On a égalité si et seulement si $\boldsymbol{x}$ et $\boldsymbol{y}$ sont colinéaires.

Preuve. Exercice B.3.17, page 224 □

Lemme 3.26: Inégalité de Hölder

Pour $p > 1$ et $\frac{1}{p} + \frac{1}{q} = 1$, on a $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{K}^n$

$$\sum_{i=1}^{n} |x_i y_i| \leqslant \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^q \right)^{1/q} = \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q . \tag{3.48}$$

Cette inégalité s'appelle l'**inégalité de Hölder**.

Preuve. Exercice B.3.19, page 227 □

Definition 3.27

Deux **normes** $\|\bullet\|$ et $\|\bullet\|'$, définies sur un même espace vectoriel V, sont **équivalentes** s'il exite

deux constantes C et C' telles que

$$\|\boldsymbol{x}\|' \leqslant C\,\|\boldsymbol{x}\| \quad \text{et} \quad \|\boldsymbol{x}\| \leqslant C'\,\|\boldsymbol{x}\|' \quad \text{pour tout } \boldsymbol{x} \in V. \tag{3.49}$$

Proposition 3.28

Sur un espace vectoriel de dimension finie toutes les normes sont équivalentes.

3.2.2 Normes matricielles

Definition 3.29

Une **norme matricielle** sur $\mathcal{M}_n(\mathbb{K})$ est une application $\|\bullet\| : \mathcal{M}_n(\mathbb{K}) \to \mathbb{R}^+$ vérifiant

1. $\|\mathbb{A}\| = 0 \iff \mathbb{A} = 0$,
2. $\|\alpha\mathbb{A}\| = |\alpha|\,\|\mathbb{A}\|$, $\forall \alpha \in \mathbb{K}$, $\forall \mathbb{A} \in \mathcal{M}_n(\mathbb{K})$,
3. $\|\mathbb{A} + \mathbb{B}\| \leqslant \|\mathbb{A}\| + \|\mathbb{B}\|\,,\ \forall\,(\mathbb{A},\mathbb{B}) \in \mathcal{M}_n(\mathbb{K})^2$ (inégalité triangulaire)
4. $\|\mathbb{A}\mathbb{B}\| \leqslant \|\mathbb{A}\|\,\|\mathbb{B}\|\,,\ \forall\,(\mathbb{A},\mathbb{B}) \in \mathcal{M}_n(\mathbb{K})^2$

Proposition 3.30

Etant donné une norme vectorielle $\|\bullet\|$ sur $\mathbb{K}^n$, l'application $\|\bullet\|_s : \mathcal{M}_n(\mathbb{K}) \to \mathbb{R}^+$ définie par

$$\|\mathbb{A}\|_s \overset{\text{def}}{=} \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n \\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} \tag{3.50}$$

est une norme matricielle, appelée **norme matricielle subordonnée** (à la norme vectorielle donnée).
Elle vérifie

$$\|\mathbb{A}\|_s = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n \\ \|\boldsymbol{v}\|\leqslant 1}} \|\mathbb{A}\boldsymbol{v}\| = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n \\ \|\boldsymbol{v}\|= 1}} \|\mathbb{A}\boldsymbol{v}\| = \inf\{\alpha \in \mathbb{R} : \|\mathbb{A}\boldsymbol{v}\| \leqslant \alpha\,\|\boldsymbol{v}\|\,,\ \forall \boldsymbol{v} \in \mathbb{K}^n\}\,. \tag{3.51}$$

De plus, pour tout $\boldsymbol{v} \in \mathbb{K}^n$ on a

$$\|\mathbb{A}\boldsymbol{v}\| \leqslant \|\mathbb{A}\|_s\,\|\boldsymbol{v}\| \tag{3.52}$$

et il existe au moins un vecteur $\boldsymbol{u} \in \mathbb{K}^n\backslash\{0\}$ tel que

$$\|\mathbb{A}\boldsymbol{u}\| = \|\mathbb{A}\|_s\,\|\boldsymbol{u}\|\,. \tag{3.53}$$

Soit $\mathbb{I}$ la matrice identité d'ordre n, on a

$$\|\mathbb{I}\|_s = 1. \tag{3.54}$$

Preuve. On note $\mathcal{B} = \{\boldsymbol{v} \in \mathbb{K}^n\ ;\ \|\boldsymbol{v}\| \leqslant 1\}$ la boule unitée de $\mathbb{K}^n$ et $\mathcal{S} = \{\boldsymbol{v} \in \mathbb{K}^n\ ;\ \|\boldsymbol{v}\| = 1\}$ la sphère unitée de $\mathbb{K}^n$. On note que les ensembles $\mathcal{B}$ et $\mathcal{S}$ sont des compacts car image réciproque de l'application continue $\boldsymbol{v} \mapsto \|\boldsymbol{v}\|$ par le fermé borné $[0,1]$ (pour la boule) et le singleton $\{1\}$ (pour la sphère).

- Vérifions que les égalités suivantes sont vraies :

$$\|\mathbb{A}\|_s \overset{\text{def}}{=} \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n \\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \sup_{\boldsymbol{v}\in\mathcal{B}} \|\mathbb{A}\boldsymbol{v}\| = \sup_{\boldsymbol{v}\in\mathcal{S}} \|\mathbb{A}\boldsymbol{v}\|$$

On a

$$\|\mathbb{A}\|_s = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n \\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n \\ \boldsymbol{v}\neq 0}} \left\|\mathbb{A}\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}\right\| = \sup_{\boldsymbol{v}\in\mathcal{S}} \|\mathbb{A}\boldsymbol{v}\|$$

Comme $\mathcal{S} \subset \mathcal{B}$ on a aussi

$$\sup_{\boldsymbol{v}\in\mathcal{B}} \|\mathbb{A}\boldsymbol{v}\| \geqslant \sup_{\boldsymbol{v}\in\mathcal{S}} \|\mathbb{A}\boldsymbol{v}\|. \tag{3.55}$$

On peut aussi remarquer que

$$\sup_{\boldsymbol{v}\in\mathcal{B}} \|\mathbb{A}\boldsymbol{v}\| = \sup_{\substack{\boldsymbol{v}\in\mathcal{B}\\ \boldsymbol{v}\neq 0}} \|\mathbb{A}\boldsymbol{v}\| \tag{3.56}$$

De plus, $\forall \boldsymbol{w} \in \mathcal{B}\backslash\{0\}$, en posant $\boldsymbol{u} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \in \mathcal{S}$. on a $w = \|\boldsymbol{w}\|\,\boldsymbol{u}$ et

$$\|\mathbb{A}\boldsymbol{w}\| = \|\boldsymbol{w}\|\,\|\mathbb{A}\boldsymbol{u}\| \leqslant \|\mathbb{A}\boldsymbol{u}\| \text{ car } \|\boldsymbol{w}\| \leqslant 1.$$

Or on a

$$\|\mathbb{A}\boldsymbol{u}\| \leqslant \sup_{\boldsymbol{v}\in\mathcal{S}} \|\mathbb{A}\boldsymbol{v}\|.$$

et on obtient alors

$$\sup_{\substack{\boldsymbol{w}\in\mathcal{B}\\ \boldsymbol{w}\neq 0}} \|\mathbb{A}\boldsymbol{w}\| \leqslant \sup_{\boldsymbol{v}\in\mathcal{S}} \|\mathbb{A}\boldsymbol{v}\|.$$

En utilisant (3.55) et (3.56), on en déduit

$$\sup_{\boldsymbol{w}\in\mathcal{B}} \|\mathbb{A}\boldsymbol{w}\| = \sup_{\boldsymbol{u}\in\mathcal{S}} \|\mathbb{A}\boldsymbol{u}\|.$$

- Vérifions que l'application $\|\bullet\|_s$ est bien définie sur $\mathcal{M}_n(\mathbb{K})$ i.e. $\forall \mathbb{A} \in \mathcal{M}_n(\mathbb{K})$, $\|\mathbb{A}\|_s < +\infty$.
  L'application $\boldsymbol{v} \mapsto \|\mathbb{A}\boldsymbol{v}\|$ est continue donc son sup sur la sphère unitée qui est compacte est atteint.
- Montrons que $\forall \boldsymbol{v} \in \mathbb{K}^n$, $\|\mathbb{A}\boldsymbol{v}\| \leqslant \|\mathbb{A}\|_s \|\boldsymbol{v}\|$.
  On a par définition du sup

$$\|\mathbb{A}\|_s = \sup_{\substack{\boldsymbol{u}\in\mathbb{K}^n\\ \boldsymbol{u}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{u}\|}{\|\boldsymbol{u}\|} \geqslant \frac{\|\mathbb{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|}, \quad \forall \boldsymbol{v} \in \mathbb{K}^n\backslash\{0\}.$$

  et donc

$$\|\mathbb{A}\boldsymbol{v}\| \leqslant \|\mathbb{A}\|_s \|\boldsymbol{v}\|, \quad \forall \boldsymbol{v} \in \mathbb{K}^n\backslash\{0\}.$$

- Montrons qu'il existe $\boldsymbol{u} \in \mathbb{K}^n$ tel que

$$\boldsymbol{u} \neq 0 \text{ et } \|\mathbb{A}\boldsymbol{u}\| = \|\mathbb{A}\|_s \|\boldsymbol{u}\|.$$

  On a

$$\|\mathbb{A}\|_s = \sup_{\boldsymbol{v}\in\mathcal{S}} \|\mathbb{A}\boldsymbol{v}\|$$

  La sphère unité étant compacte et l'application $\boldsymbol{v} \mapsto \|\mathbb{A}\boldsymbol{v}\|$ étant continue , il existe $\boldsymbol{w} \in \mathcal{S}$ tel que $\|\mathbb{A}\|_s = \|\mathbb{A}\boldsymbol{w}\|$. Soit $\lambda \in \mathbb{K}^*$ et $\boldsymbol{u} = \lambda\boldsymbol{w} \neq 0$. On a $\|\boldsymbol{u}\| = |\lambda|$ et

$$\|\mathbb{A}\|_s = \|\mathbb{A}\boldsymbol{w}\| = \left\|\mathbb{A}\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\| = \frac{1}{\|\boldsymbol{u}\|}\|\mathbb{A}\boldsymbol{u}\|.$$

- On a immédiatement

$$\|\mathbb{I}\|_s = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{I}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = 1.$$

- Montrons que $\|\bullet\|_s$ est une norme matricielle.

  1. $\|\mathbb{A}\|_s = 0 \Longleftrightarrow \mathbb{A}_s = \mathbb{O}$?
     $\boxed{\Longleftarrow}$ trivial.
     $\boxed{\Longrightarrow}$ Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$.

$$\|\mathbb{A}\|_s = 0 = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} \Longrightarrow \|\mathbb{A}\boldsymbol{v}\| = 0, \ \forall \boldsymbol{v} \in \mathbb{K}^n\backslash\{0\}$$

$$\Longrightarrow \mathbb{A}\boldsymbol{v} = \boldsymbol{0}, \ \forall \boldsymbol{v} \in \mathbb{K}^n\backslash\{0\}$$

     Soit $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ la base canonique de $\mathbb{K}^n$. On a alors $\forall j \in [\![1, n]\!]$, $\mathbb{A}\boldsymbol{e}_j = \boldsymbol{0}$ et on en déduit que

$$\mathrm{A}_{i,j} = \langle \boldsymbol{e}_i, \mathbb{A}\boldsymbol{e}_j \rangle = 0, \ \forall (i, j) \in [\![1, n]\!].$$

     et donc $\mathbb{A} = \mathbb{O}$.

2. Montrons que $\|\alpha\mathbb{A}\| = |\alpha|\,\|\mathbb{A}\|$, $\forall \alpha \in \mathbb{K}$, $\forall \mathbb{A} \in \mathcal{M}_n(\mathbb{K})$,.
Soient $\alpha \in \mathbb{K}$ et $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$. On a $\alpha\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$ (car $\mathcal{M}_n(\mathbb{K})$ est un espace vectoriel) et

$$\begin{aligned}\|\alpha\mathbb{A}\|_s &= \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\alpha\mathbb{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{|\alpha|\,\|\mathbb{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} \quad \text{car } \|\alpha\boldsymbol{u}\| = |\alpha|\,\|\boldsymbol{u}\| \\ &= |\alpha| \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = |\alpha|\,\|\mathbb{A}\|_s\,.\end{aligned}$$

3. Montrons que $\|\mathbb{A}+\mathbb{B}\|_s \leqslant \|\mathbb{A}\|_s + \|\mathbb{B}\|_s$, $\forall\,(\mathbb{A},\mathbb{B}) \in \mathcal{M}_n(\mathbb{K})^2$
Soient $\mathbb{A}$ et $\mathbb{B}$ deux matrices de $\mathcal{M}_n(\mathbb{K})$. On a $\mathbb{A}+\mathbb{B} \in \mathcal{M}_n(\mathbb{K})$ car $\mathcal{M}_n(\mathbb{K})$ est un espace vectoriel et

$$\begin{aligned}\|\mathbb{A}+\mathbb{B}\|_s &= \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|(\mathbb{A}+\mathbb{B})\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}+\mathbb{B}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} \\ &\leqslant \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|+\|\mathbb{B}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} \quad \text{par inégalité triangulaire dans } \mathbb{K}^n \\ &\leqslant \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} + \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{B}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \|\mathbb{A}\|_s + \|\mathbb{B}\|_s\,.\end{aligned}$$

4. Montrons que $\|\mathbb{A}\mathbb{B}\|_s \leqslant \|\mathbb{A}\|_s\,\|\mathbb{B}\|_s$, $\forall\,(\mathbb{A},\mathbb{B}) \in \mathcal{M}_n(\mathbb{K})^2$.
Soient $\mathbb{A}$ et $\mathbb{B}$ deux matrices de $\mathcal{M}_n(\mathbb{K})$. On a $\mathbb{A}\mathbb{B} \in \mathcal{M}_n(\mathbb{K})$ par définition du produit matriciel et

$$\begin{aligned}\|\mathbb{A}\mathbb{B}\|_s &= \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|(\mathbb{A}\mathbb{B})\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}(\mathbb{B}\boldsymbol{v})\|}{\|\boldsymbol{v}\|} \\ &\leqslant \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\|_s\,\|\mathbb{B}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} \quad \text{car } \|\mathbb{A}\boldsymbol{u}\| \leqslant \|\mathbb{A}\|_s\,\|\boldsymbol{u}\| \ \ \forall \boldsymbol{u} \in \mathbb{K}^n \\ &\leqslant \|\mathbb{A}\|_s \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{B}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \|\mathbb{A}\|_s\,\|\mathbb{B}\|_s\,.\end{aligned}$$

□

Théorème 3.31

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$. On a

$$\|\mathbb{A}\|_1 \overset{\text{def}}{=} \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|_1}{\|\boldsymbol{v}\|_1} = \max_{j\in[\![1,n]\!]} \sum_{i=1}^{n} |a_{ij}| \tag{3.57}$$

$$\|\mathbb{A}\|_2 \overset{\text{def}}{=} \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2} = \sqrt{\rho\left(\mathbb{A}^*\mathbb{A}\right)} = \sqrt{\rho\left(\mathbb{A}\mathbb{A}^*\right)} = \|\mathbb{A}^*\|_2 \tag{3.58}$$

$$\|\mathbb{A}\|_\infty \overset{\text{def}}{=} \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n\\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|_\infty}{\|\boldsymbol{v}\|_\infty} = \max_{i\in[\![1,n]\!]} \sum_{j=1}^{n} |a_{ij}| \tag{3.59}$$

La norme $\|\bullet\|_2$ est invariante par transformation unitaire :

$$\mathbb{U}\mathbb{U}^* = \mathbb{I} \Longrightarrow \|\mathbb{A}\|_2 = \|\mathbb{A}\mathbb{U}\|_2 = \|\mathbb{U}\mathbb{A}\|_2 = \|\mathbb{U}^*\mathbb{A}\mathbb{U}\|_2\,. \tag{3.60}$$

Preuve. Exercice B.3.24, page 230 □

Corollaire 3.32

1. Si une matrice $\mathbb{A}$ est hermitienne,on a $\|\mathbb{A}\|_2 = \rho(\mathbb{A})$.
2. Si une matrice $\mathbb{A}$ est unitaire, on a $\|\mathbb{A}\|_2 = 1$.

Preuve. 1. Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice hermitienne. Elle est donc normale. D'après le théorème 3.2 (réduction des matrice) page 63, il existe une matrice unitaire $\mathbb{U}$ et une matrice diagonale $\mathbb{D}$ telles que

$$\mathbb{A} = \mathbb{U}\mathbb{D}\mathbb{U}^*.$$

Les matrices $\mathbb{A}$ et $\mathbb{D}$ sont semblables: elles ont les mêmes valeurs propres et donc

$$\rho(\mathbb{A}) = \rho(\mathbb{D}).$$

De plus, comme $\mathbb{A}$ est hermitienne, ses valeurs propres sont réelles et donc $\mathbb{D} \in \mathcal{M}_n(\mathbb{R})$. Comme la norme 2 est invariante par transformation unitaire, on a

$$\|\mathbb{A}\|_2 = \|\mathbb{U}\mathbb{D}\mathbb{U}^*\|_2 = \|\mathbb{D}\|_2 .$$

De plus, $\mathbb{D}$ étant réelle, on obtient

$$\|\mathbb{D}\|_2 = \sqrt{\rho\left(\mathbb{D}^*\mathbb{D}\right)} = \sqrt{\rho\left(\mathbb{D}^2\right)}$$

La matrice $\mathbb{D}$ étant diagonale on en déduit que

$$\rho\left(\mathbb{D}^2\right) = \rho(\mathbb{D})^2.$$

On obtient donc

$$\|\mathbb{A}\|_2 = \rho(\mathbb{D}) = \rho(\mathbb{A}).$$

2. Si $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ est unitaire ou si $\mathbb{A} \in \mathcal{M}_n(\mathbb{R})$ est orthogonale alors $\mathbb{A}\mathbb{A}^* = \mathbb{I}$ et donc

$$\|\mathbb{A}\|_2 = \sqrt{\rho\left(\mathbb{A}\mathbb{A}^*\right)} = \sqrt{\rho\left(\mathbb{I}\right)} = 1.$$

□

Théorème 3.33

1. Soit $\mathbb{A}$ une matrice carrée quelconque et $\|\bullet\|$ une norme matricielle subordonnée ou non, quelconque. Alors
$$\rho(\mathbb{A}) \leqslant \|\mathbb{A}\| . \tag{3.61}$$
2. Etant donné une matrice $\mathbb{A}$ et un nombre $\varepsilon > 0$, il existe au moins une norme matricielle subordonnée telle que
$$\|\mathbb{A}\| \leqslant \rho(\mathbb{A}) + \varepsilon. \tag{3.62}$$

Preuve. 1. Soient $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$ et $(\lambda, \boldsymbol{u})$ un élément propre de $\mathbb{A}$:

$$\boldsymbol{u} \in \mathbb{C}^n \backslash \{\mathbf{0}\}, \quad \text{et} \quad \mathbb{A}\boldsymbol{u} = \lambda \boldsymbol{u}.$$

- Si la norme matricielle est subordonnée à la norme vectorielle $\|\bullet\|$ alors

$$\|\lambda \boldsymbol{u}\| = |\lambda| \, \|\boldsymbol{u}\| = \|\mathbb{A}\boldsymbol{u}\| \leqslant \|\mathbb{A}\| \, \|\boldsymbol{u}\|$$

Or $\boldsymbol{u} \in \mathbb{C}^n \backslash \{\mathbf{0}\}$ et donc on en déduit que

$$\forall \lambda \in \rho(\mathbb{A}), \quad |\lambda| \leqslant \|\mathbb{A}\| .$$

- Si la norme est quelconque (non forcément subordonnée), l'inégalité suivante n'est plus vérifiée

$$\|\mathbb{A}\boldsymbol{u}\| \leqslant \|\mathbb{A}\| \, \|\boldsymbol{u}\| .$$

Par contre on a pour toute matrice $\mathbb{B} \in \mathcal{M}_n(\mathbb{K})$

$$\|\mathbb{A}\mathbb{B}\| \leqslant \|\mathbb{A}\| \, \|\mathbb{B}\| .$$

Comme $(\lambda, \boldsymbol{u})$ un élément propre de $\mathbb{A}$, on a

$$\mathbb{A}\boldsymbol{u}\boldsymbol{u}^* = \lambda \boldsymbol{u}\boldsymbol{u}^*$$

et donc

$$\|\mathbb{A}(\boldsymbol{u}\boldsymbol{u}^*)\| = \|\lambda(\boldsymbol{u}\boldsymbol{u}^*)\| = |\lambda| \, \|\boldsymbol{u}\boldsymbol{u}^*\| .$$

De plus on a

$$\|\mathbb{A}(\boldsymbol{u}\boldsymbol{u}^*)\| \leqslant \|\mathbb{A}\| \, \|(\boldsymbol{u}\boldsymbol{u}^*)\|$$

ce qui donne

$$|\lambda| \, \|\boldsymbol{u}\boldsymbol{u}^*\| \leqslant \|\mathbb{A}\| \, \|(\boldsymbol{u}\boldsymbol{u}^*)\|$$

Comme la matrice $\mathbb{B} = \boldsymbol{u}\boldsymbol{u}^*$ est non nulle car $\boldsymbol{u}$ non nul et

$$(\boldsymbol{u}\boldsymbol{u}^*)\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{u}^*)\boldsymbol{u}) = \boldsymbol{u}\langle \boldsymbol{u}, \boldsymbol{u}\rangle \neq 0.$$

on obtient

$$|\lambda| \leqslant \|\mathbb{A}\|$$

et on en déduit alors

$$\forall \lambda \in \rho(\mathbb{A}), |\lambda| \leqslant \|\mathbb{A}\| .$$

2. voir [1], théorème 1.4-3 page 18-19.

□

Théorème 3.34

L'application $\|\bullet\|_E : \mathcal{M}_n \to \mathbb{R}^+$ définie par

$$\|\mathbb{A}\|_E = \left(\sum_{(i,j)=\in[\![1,n]\!]^2} |a_{ij}|^2 \right)^{1/2} = \sqrt{\mathrm{tr}\,(\mathbb{A}^*\mathbb{A})}, \tag{3.63}$$

pour toute matrice $\mathbb{A} = (a_{ij})$ d'ordre n, est une norme matricielle non subordonnée (pour $n \geqslant 2$), invariante par transformation unitaire et qui vérifie

$$\|\mathbb{A}\|_2 \leqslant \|\mathbb{A}\|_E \leqslant \sqrt{n} \, \|\mathbb{A}\|_2 , \ \forall \mathbb{A} \in \mathcal{M}_n. \tag{3.64}$$

De plus $\|\mathbb{I}\|_E = \sqrt{n}$.

Théorème 3.35

1. Soit $\|\bullet\|$ une norme matricielle subordonnée, et $\mathbb{B}$ une matrice vérifiant

$$\|\mathbb{B}\| < 1.$$

Alors la matrice $(\mathbb{I} + \mathbb{B})$ est inversible, et

$$\left\| (\mathbb{I} + \mathbb{B})^{-1} \right\| \leqslant \frac{1}{1 - \|\mathbb{B}\|}.$$

2. Si une matrice de la forme $(\mathbb{I} + \mathbb{B})$ est singulière, alors nécessairement

$$\|\mathbb{B}\| \geqslant 1$$

pour toute norme matricielle, subordonnée ou non.

Preuve. 1. Par l'absurde, on suppose la matrice $\mathbb{I} + \mathbb{B}$ singulière (non inversible). Alors 0 est une de ses valeurs propres. On note $(0, \boldsymbol{u})$ un élément propre de $\mathbb{I} + \mathbb{B}$. On a donc

$$\begin{aligned}(\mathbb{I} + \mathbb{B})\boldsymbol{u} = \boldsymbol{0} &\Leftrightarrow \mathbb{B}\boldsymbol{u} = -\boldsymbol{u} \\ &\Leftrightarrow (-1, \boldsymbol{u}) \quad \text{élément propre de } \mathbb{B}\end{aligned}$$

ce qui est en contradiction avec $\rho(\mathbb{B}) \leqslant \|\mathbb{B}\| < 1$.
La matrice $\mathbb{I} + \mathbb{B}$ est donc inversible.
Par définition de l'inverse

$$(\mathbb{I} + \mathbb{B})(\mathbb{I} + \mathbb{B})^{-1} = \mathbb{I}$$

et donc

$$(\mathbb{I} + \mathbb{B})^{-1} + \mathbb{B}(\mathbb{I} + \mathbb{B})^{-1} = \mathbb{I}$$

ce qui s'écrit aussi

$$(\mathbb{I} + \mathbb{B})^{-1} = \mathbb{I} - \mathbb{B}(\mathbb{I} + \mathbb{B})^{-1}.$$

En prenant sa norme on obtient

$$\begin{aligned}\left\|(\mathbb{I} + \mathbb{B})^{-1}\right\| &= \left\|\mathbb{I} - \mathbb{B}(\mathbb{I} + \mathbb{B})^{-1}\right\| \\ &\leqslant \|\mathbb{I}\| + \left\|\mathbb{B}(\mathbb{I} + \mathbb{B})^{-1}\right\| \quad \text{par l'inégalité triangulaire} \\ &\leqslant 1 + \left\|\mathbb{B}(\mathbb{I} + \mathbb{B})^{-1}\right\| \quad \text{car pour une norme subordonnée } \|\mathbb{I}\| = 1 \\ &\leqslant 1 + \|\mathbb{B}\| \left\|(\mathbb{I} + \mathbb{B})^{-1}\right\|\end{aligned}$$

Ce qui donne

$$\left\|(\mathbb{I} + \mathbb{B})^{-1}\right\| - \|\mathbb{B}\| \left\|(\mathbb{I} + \mathbb{B})^{-1}\right\| \leqslant 1$$

et donc l'inégalité est démontrée.

2. On a démontré lors de la démontration par l'absurde que si $\mathbb{I} + \mathbb{B}$ est singulière (non inversible) alors $\rho(\mathbb{B}) \geqslant 1$. Comme pour toute norme $\rho(\mathbb{B}) \leqslant \|\mathbb{B}\|$ (voir théorème 3.33, page 101), on obtient

$$\|\mathbb{B}\| \geqslant 1.$$

□

3.2.3 Suites de vecteurs et de matrices

Definition 3.36

Soit V un espace vectoriel muni d'une norme $\|\bullet\|$, on dit qu'une suite $(\boldsymbol{v}_k)$ d'éléments de V **converge vers un élément** $\boldsymbol{v} \in V$, si

$$\lim_{k\to\infty} \|\boldsymbol{v}_k - \boldsymbol{v}\| = 0$$

et on écrit

$$\boldsymbol{v} = \lim_{k\to\infty} \boldsymbol{v}_k.$$

Théorème 3.37: admis, voir [1] Théorème 1.5-1 page 21-22

Soit $\mathbb{B}$ une matrice carrée. Les conditions suivantes sont équivalentes :

1. $\lim_{k\to\infty} \mathbb{B}^k = 0$,
2. $\lim_{k\to\infty} \mathbb{B}^k \boldsymbol{v} = 0$ pour tout vecteur $\boldsymbol{v}$,
3. $\rho(\mathbb{B}) < 1$,

4. $\|\mathbb{B}\| < 1$ pour au moins une norme matricielle subordonnée $\|\bullet\|$.

Théorème 3.38: admis, voir [1] Théorème 1.5-2 page 22

Soit $\mathbb{B}$ une matrice carrée, et $\|\bullet\|$ une norme matricielle quelconque. Alors

$$\lim_{k\to\infty} \left\|\mathbb{B}^k\right\|^{1/k} = \rho(\mathbb{B}).$$

3.3 Conditionnement d'un système linéaire

Pour la résolution numérique d'un système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$, il est rare que les données $\mathbb{A}$ et $\boldsymbol{b}$ du problème ne soient pas entachées d'erreurs (aussi minimes soient-elles). La question qui se pose alors est de savoir si de petites perturbations sur les données ne peuvent pas entrainer des erreurs importantes sur le calcul de la solution.

Exemple de R.S. Wilson

Soient

$$\mathbb{A} = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, \quad \Delta\mathbb{A} = \begin{pmatrix} 0 & 0 & \frac{1}{10} & \frac{1}{5} \\ \frac{2}{25} & \frac{1}{25} & 0 & 0 \\ 0 & -\frac{1}{50} & -\frac{11}{100} & 0 \\ -\frac{1}{100} & -\frac{1}{100} & 0 & -\frac{1}{50} \end{pmatrix}$$

et $\boldsymbol{b}^{\mathsf{t}} = (32,\, 23,\, 33,\, 31)$, $(\boldsymbol{\Delta b})^{\mathsf{t}} = \left(\frac{1}{100}, -\frac{1}{100}, \frac{1}{100}, -\frac{1}{100}\right)$. Des calculs exacts donnent

$$\begin{aligned}
\mathbb{A}\boldsymbol{x} = \boldsymbol{b} \quad &\Longleftrightarrow \quad \boldsymbol{x}^{\mathsf{t}} = (1,\, 1,\, 1,\, 1) \\
\mathbb{A}\boldsymbol{u} = (\boldsymbol{b} + \boldsymbol{\Delta b}) \quad &\Longleftrightarrow \quad \boldsymbol{u}^{\mathsf{t}} = \left(\frac{91}{50}, -\frac{9}{25}, \frac{27}{20}, \frac{79}{100}\right) \\
&\qquad\qquad \approx (1.8,\, -0.36,\, 1.3,\, 0.79) \\
(\mathbb{A} + \Delta\mathbb{A})\boldsymbol{v} = \boldsymbol{b} \quad &\Longleftrightarrow \quad \boldsymbol{v}^{\mathsf{t}} = (-81,\, 137,\, -34,\, 22) \\
(\mathbb{A} + \Delta\mathbb{A})\boldsymbol{y} = (\boldsymbol{b} + \boldsymbol{\Delta b}) \quad &\Longleftrightarrow \quad \boldsymbol{y}^{\mathsf{t}} = \left(-\frac{18283543}{461600}, \frac{31504261}{461600}, -\frac{3741501}{230800}, \frac{5235241}{461600}\right) \\
&\qquad\qquad \approx (-39.61,\, 68.25,\, -16.21,\, 11.34)
\end{aligned}$$

Il est clair sur cet exemple que de petites perturbations sur les données peuvent entrainer des erreurs importantes sur la solution exacte.

On dit que le système linéaire prédédent est **mal conditionné** ou qu'il a un **mauvais conditionnement** car il est sujet à de fortes variations de la solution pour de petites perturbations des données. A contrario, on dit qu'il est **bien conditionné** ou qu'il a un **bon conditionnement** si de petites perturbations des données n'entrainent qu'une variation *raisonnable* de la solution.

Une nouvelle question : est-il possible de "mesurer" le **conditionnement** d'une matrice?

Définitions et résultats

Definition 3.39

Soit $\|.\|$ une norme matricielle subordonnée, le conditionnement d'une matrice régulière $\mathbb{A}$, associé à cette norme, est le nombre

$$\mathrm{cond}(\mathbb{A}) = \|\mathbb{A}\| \left\|\mathbb{A}^{-1}\right\|.$$

Nous noterons $\mathrm{cond}_p(\mathbb{A}) = \|\mathbb{A}\|_p \left\|\mathbb{A}^{-1}\right\|_p$.

Proposition 3.40

Soit $\mathbb{A}$ une matrice régulière. On a les propriétés suivantes

1. $\forall \alpha \in \mathbb{K}^*$, $\text{cond}(\alpha\mathbb{A}) = \text{cond}(\mathbb{A})$.
2. $\text{cond}_p(\mathbb{A}) \geqslant 1$, $\forall p \in [\![1, +\infty]\!]$.
3. $\text{cond}_2(\mathbb{A}) = 1$ si et seulement si $\mathbb{A} = \alpha\mathbb{Q}$ avec $\alpha \in \mathbb{K}^*$ et $\mathbb{Q}$ matrice unitaire

Preuve. Soit $\mathbb{A}$ une matrice régulière.

1. Soit $\alpha \in \mathbb{K}^*$, on a
$$\begin{aligned}\text{cond}(\alpha\mathbb{A}) &\overset{\text{def}}{=} \|\alpha\mathbb{A}\| \left\|(\alpha\mathbb{A})^{-1}\right\| = |\alpha| \|\mathbb{A}\| \left\|\frac{1}{\alpha}\mathbb{A}^{-1}\right\| \\ &= |\alpha| \|\mathbb{A}\| \frac{1}{|\alpha|} \left\|\mathbb{A}^{-1}\right\| = \|\mathbb{A}\| \left\|\mathbb{A}^{-1}\right\| \\ &= \text{cond}(\mathbb{A})\end{aligned}$$
2. On a $\mathbb{I} = \mathbb{A}\mathbb{A}^{-1}$. Or pour toute norme subordonnée, on a $\|\mathbb{I}\| = 1$ et donc $1 = \|\mathbb{A}\mathbb{A}^{-1}\| \leqslant \|\mathbb{A}\| \|\mathbb{A}^{-1}\| = \text{cond}(\mathbb{A})$.
3. **Admis**. Voir par exemple [7] Théorème 2 page 142-143.

□

Théorème 3.41

Soit $\mathbb{A}$ une matrice inversible. Soient $\boldsymbol{x}$ et $\boldsymbol{x} + \boldsymbol{\Delta x}$ les solutions respectives de
$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b} \quad \text{et} \quad \mathbb{A}(\boldsymbol{x} + \boldsymbol{\Delta x}) = \boldsymbol{b} + \boldsymbol{\Delta b}.$$
Supposons $\boldsymbol{b} \neq \boldsymbol{0}$, alors l'inégalité
$$\frac{\|\boldsymbol{\Delta x}\|}{\|\boldsymbol{x}\|} \leqslant \text{cond}(\mathbb{A}) \frac{\|\boldsymbol{\Delta b}\|}{\|\boldsymbol{b}\|}$$
est satisfaite, et c'est la meilleure possible : pour une matrice $\mathbb{A}$ donnée, on peut trouver des vecteurs $\boldsymbol{b} \neq \boldsymbol{0}$ et $\boldsymbol{\Delta b} \neq \boldsymbol{0}$ tels qu'elle devienne une égalité.

Preuve. On a
$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b} \quad \text{et} \quad \mathbb{A}(\boldsymbol{x} + \boldsymbol{\Delta x}) = \boldsymbol{b} + \boldsymbol{\Delta b}$$
or $\mathbb{A}(\boldsymbol{x} + \boldsymbol{\Delta x}) = \mathbb{A}\boldsymbol{x} + \mathbb{A}\boldsymbol{\Delta x}$ et donc $\mathbb{A}\boldsymbol{\Delta x} = \boldsymbol{\Delta b}$ ou encore $\boldsymbol{\Delta x} = \mathbb{A}^{-1}\boldsymbol{\Delta b}$. Ceci donne
$$\|\boldsymbol{\Delta x}\| = \left\|\mathbb{A}^{-1}\boldsymbol{\Delta b}\right\| \leqslant \left\|\mathbb{A}^{-1}\right\| \|\boldsymbol{\Delta b}\|$$
et
$$\|\boldsymbol{b}\| = \|\mathbb{A}\boldsymbol{x}\| \leqslant \|\mathbb{A}\| \|\boldsymbol{x}\|.$$
On en déduit
$$\|\boldsymbol{\Delta x}\| \|\boldsymbol{b}\| \leqslant \|\mathbb{A}\| \|\boldsymbol{x}\| \left\|\mathbb{A}^{-1}\right\| \|\boldsymbol{\Delta b}\|.$$
Comme $\boldsymbol{b} \neq \boldsymbol{0}$, on a $\boldsymbol{x} = \mathbb{A}^{-1}\boldsymbol{b} \neq \boldsymbol{0}$ et, les normes étant positives, on obtient
$$\frac{\|\boldsymbol{\Delta x}\|}{\|\boldsymbol{x}\|} \leqslant \text{cond}(\mathbb{A}) \frac{\|\boldsymbol{\Delta b}\|}{\|\boldsymbol{b}\|}$$

D'après la Proposition 3.30, pour toute norme matricielle subordonnée il existe au moins un vecteur $\boldsymbol{u} \in \mathbb{K}^n \backslash \{0\}$ et un vecteur $\boldsymbol{v} \in \mathbb{K}^n \backslash \{0\}$ tel que
$$\left\|\mathbb{A}^{-1}\boldsymbol{u}\right\| = \left\|\mathbb{A}^{-1}\right\| \|\boldsymbol{u}\| \text{ et } \|\mathbb{A}\boldsymbol{v}\| = \|\mathbb{A}\| \|\boldsymbol{v}\|.$$
En posant $\boldsymbol{b} = \mathbb{A}\boldsymbol{v}$ et $\boldsymbol{\Delta b} = \boldsymbol{u}$ on a bien égalité. □

Théorème 3.42

Soient $\mathbb{A}$ et $\mathbb{A}+\Delta\mathbb{A}$ deux matrices inversibles. Soient $\boldsymbol{x}$ et $\boldsymbol{x}+\boldsymbol{\Delta x}$ les solutions respectives de

$$\mathbb{A}\boldsymbol{x}=\boldsymbol{b} \text{ et } (\mathbb{A}+\Delta\mathbb{A})(\boldsymbol{x}+\boldsymbol{\Delta x})=\boldsymbol{b}.$$

Supposons $\boldsymbol{b}\neq\boldsymbol{0}$, alors on a

$$\frac{\|\boldsymbol{\Delta x}\|}{\|\boldsymbol{x}+\boldsymbol{\Delta x}\|}\leqslant \mathrm{cond}(\mathbb{A})\frac{\|\Delta\mathbb{A}\|}{\|\mathbb{A}\|}.$$

Preuve. On a

$$\mathbb{A}\boldsymbol{x}=\boldsymbol{b}=(\mathbb{A}+\Delta\mathbb{A})(\boldsymbol{x}+\boldsymbol{\Delta x})=\mathbb{A}\boldsymbol{x}+\mathbb{A}\boldsymbol{\Delta x}+\Delta\mathbb{A}(\boldsymbol{x}+\boldsymbol{\Delta x})$$

et donc

$$\mathbb{A}\boldsymbol{\Delta x}+\Delta\mathbb{A}(\boldsymbol{x}+\boldsymbol{\Delta x})=\boldsymbol{0}\iff \boldsymbol{\Delta x}=-\mathbb{A}^{-1}\Delta\mathbb{A}(\boldsymbol{x}+\boldsymbol{\Delta x})$$

On en déduit alors

$$\|\boldsymbol{\Delta x}\|=\left\|\mathbb{A}^{-1}\Delta\mathbb{A}(\boldsymbol{x}+\boldsymbol{\Delta x})\right\|\leqslant\left\|\mathbb{A}^{-1}\right\|\|\Delta\mathbb{A}(\boldsymbol{x}+\boldsymbol{\Delta x})\|\leqslant\left\|\mathbb{A}^{-1}\right\|\|\Delta\mathbb{A}\|\,\|\boldsymbol{x}+\boldsymbol{\Delta x}\|$$

De plus, on a $\mathrm{cond}(\mathbb{A})\overset{\text{def}}{=}\|\mathbb{A}\|\,\|\mathbb{A}^{-1}\|$ et donc $\|\mathbb{A}^{-1}\|=\frac{\mathrm{cond}(\mathbb{A})}{\|\mathbb{A}\|}$ ce qui donne

$$\|\boldsymbol{\Delta x}\|\leqslant\mathrm{cond}(\mathbb{A})\frac{\|\Delta\mathbb{A}\|}{\|\mathbb{A}\|}\|\boldsymbol{x}+\boldsymbol{\Delta x}\|.$$

Comme $\boldsymbol{b}\neq 0$, on a $\boldsymbol{x}+\boldsymbol{\Delta x}=(\mathbb{A}+\Delta\mathbb{A})^{-1}\boldsymbol{b}\neq 0$ et de l'inégalité précédente, on déduit alors

$$\frac{\|\boldsymbol{\Delta x}\|}{\|\boldsymbol{x}+\boldsymbol{\Delta x}\|}\leqslant\mathrm{cond}(\mathbb{A})\frac{\|\Delta\mathbb{A}\|}{\|\mathbb{A}\|}.$$

□

Remarque 3.43

1. Plus le conditionnement d'une matrice est proche de 1, meilleur il est.
2. Si une matrice $\mathbb{A}$ est mal conditionnée, il est possible de trouver une matrice $\mathbb{P}$ inversible tel que la matrice $\mathbb{P}^{-1}\mathbb{A}$ soit mieux conditionné... Résoudre $\mathbb{A}\boldsymbol{x}=\boldsymbol{b}$ est alors équivalent à résoudre $\mathbb{P}^{-1}\mathbb{A}\boldsymbol{x}=\mathbb{P}^{-1}\boldsymbol{b}$. La matrice $\mathbb{P}$ est appelé **préconditionneur**. Le choix $\mathbb{P}=\mathbb{A}$ est idéal mais n'est pas réellement utilisable. Si $\mathbb{A}$ est une matrice dominante alors on peut utiliser le préconditionneur de Jacobi: $\mathbb{P}=\mathrm{diag}(\mathbb{A})$.

3.4 Méthodes itératives

3.4.1 Principe

On souhaite résoudre le système linéaire $\mathbb{A}\boldsymbol{x}=\boldsymbol{b}$ par des **méthodes itératives**. Ces dernières consitent en la détermination d'une **matrice d'itération** $\mathbb{B}$ et d'un vecteur $\boldsymbol{c}$ tels que la suite de vecteurs $\boldsymbol{x}^{[k]}$ définie par

$$\boldsymbol{x}^{[k+1]}=\mathbb{B}\boldsymbol{x}^{[k]}+\boldsymbol{c},\ k\geqslant 0,\ \boldsymbol{x}^{[0]}\text{ arbitraire}$$

soit telle que $\lim_{k\to\infty}\boldsymbol{x}^{[k]}=\underline{\boldsymbol{x}}$ et $\underline{\boldsymbol{x}}$ solution de $\mathbb{A}\boldsymbol{x}=\boldsymbol{b}$. Bien évidemment les matrices $\mathbb{B}$ et les vecteurs $\boldsymbol{c}$ dépendront de $\mathbb{A}$ et $\boldsymbol{b}$.

Nous allons étudier plusieurs méthodes itératives et pour chacune d'entre elles nous expliciterons la matrice d'itération $\mathbb{B}$ et le vecteur $\boldsymbol{c}$ associé.

3.4.2 Présentation des méthodes usuelles

Soit $\mathbb{A}\in\mathcal{M}_n(\mathbb{K})$ une matrice régulière, d'éléments diagonaux non-nuls, et $\boldsymbol{b}\in\mathbb{K}^n$. On note $\mathbb{D}$ la **matrice diagonale** telle que $\mathbb{D}=\mathrm{diag}(\mathbb{A})$, $\mathbb{E}$ la **matrice triangulaire inférieure** à diagonale nulle définie par

$$\begin{cases}\mathrm{E}_{ij}=0, & i\leqslant j\\ \mathrm{E}_{ij}=-\mathrm{A}_{ij} & i>j\end{cases}\tag{3.65}$$

et $\mathbb{F}$ la **matrice triangulaire supérieure** à diagonale nulle définie par

$$\begin{cases} \mathrm{F}_{ij} = 0, & i \geqslant j \\ \mathrm{F}_{ij} = -\mathrm{A}_{ij} & i < j \end{cases} \tag{3.66}$$

On a alors

$$\begin{aligned} \mathbb{A} &= \begin{pmatrix} \mathrm{A}_{1,1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathrm{A}_{n,n} \end{pmatrix} + \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ \mathrm{A}_{2,1} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathrm{A}_{n,1} & \cdots & \mathrm{A}_{n,n-1} & 0 \end{pmatrix} + \begin{pmatrix} 0 & \mathrm{A}_{1,2} & \cdots & \mathrm{A}_{1,n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \mathrm{A}_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{pmatrix} \\ &= \mathbb{D} - \mathbb{E} - \mathbb{F} = \begin{pmatrix} \ddots & & -\mathbb{F} \\ & \mathbb{D} & \\ -\mathbb{E} & & \ddots \end{pmatrix} \end{aligned} \tag{3.67}$$

On rappelle que la i-ème équation du système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ s'écrit

$$b_i = \sum_{j=1}^{n} \mathrm{A}_{i,j} x_j = \sum_{j=1}^{i-1} \mathrm{A}_{i,j} x_j + \mathrm{A}_{i,i} x_i + \sum_{j=i+1}^{n} \mathrm{A}_{i,j} x_j. \tag{3.68}$$

Il faut aussi noter que la matrice diagonale $\mathbb{D}$ est inversible car les éléments diagonaux de $\mathbb{A}$ sont non nuls par hypothèse.

Méthode de Jacobi

Pour obtenir la méthode itérative de Jacobi, il suffit de *mettre*, dans la formule (3.4.2), l'itéré $k+1$ sur le terme diagonale et l'itéré k sur les autres termes pour avoir

$$b_i = \sum_{j=1}^{i-1} \mathrm{A}_{i,j} x_j^{[k]} + \mathrm{A}_{i,i} x_i^{[k+1]} + \sum_{j=i+1}^{n} \mathrm{A}_{i,j} x_j^{[k]}.$$

ce qui donne

$$x_i^{[k+1]} = \frac{1}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{n} \mathrm{A}_{ij} x_j^{[k]} \right) \quad \forall i \in [\![1, n]\!] \tag{3.69}$$

Méthode de Gauss-Seidel

Pour obtenir la méthode itérative de Jacobi, il suffit de *mettre*, dans la formule (3.4.2), l'itéré $k+1$ sur la partie *triangulaire inférieure* et l'itéré k sur les autres termes pour avoir

$$b_i = \sum_{j=1}^{i-1} \mathrm{A}_{i,j} x_j^{[k+1]} + \mathrm{A}_{i,i} x_i^{[k+1]} + \sum_{j=i+1}^{n} \mathrm{A}_{i,j} x_j^{[k]}.$$

ce qui donne

$$x_i^{(k+1)} = \frac{1}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1}^{i-1} \mathrm{A}_{ij} x_j^{[k+1]} - \sum_{j=i+1}^{n} \mathrm{A}_{ij} x_j^{[k]} \right) \quad \forall i \in [\![1, n]\!] \tag{3.70}$$

Méthodes de relaxation

Ces méthodes sont basées sur un paramètre de relaxation $w \in \mathbb{R}^*$ et sont données par

$$x_i^{[k+1]} = w \underline{x}_i^{[k+1]} + (1-w) x_i^{[k]}$$

où $\underline{x}_i^{[k+1]}$ est obtenu à partir de l'une des deux méthodes précédentes.

Avec la méthode de Jacobi

$$x_i^{(k+1)} = \frac{w}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{n} \mathrm{A}_{ij} x_j^{[k]} \right) + (1-w) x_i^{[k]} \ \forall i \in [\![1, n]\!].$$

Avec la méthode de Gauss-Seidel

$$x_i^{[k+1]} = \frac{w}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1}^{i-1} \mathrm{A}_{ij} x_j^{[k+1]} - \sum_{j=i+1}^{n} \mathrm{A}_{ij} x_j^{[k]} \right) + (1-w) x_i^{[k]} \quad \forall i \in [\![1, n]\!]$$

Cette dernière méthode de relaxation, utilisant la méthode de Gauss-Seidel, est appelée méthode S.O.R. (successive over relaxation)

Exercice 3.4.1

En écrivant $\mathbb{A}$ sous la forme $\mathbb{A} = \mathbb{D} - \mathbb{E} - \mathbb{F}$, montrer que les méthodes itératives de Jacobi, Gauss-Seidel et S.O.R. s'écrivent sous la forme $\boldsymbol{x}^{[k+1]} = \mathbb{B}\boldsymbol{x}^{[k]} + \boldsymbol{c}$, où l'on exprimera les matrices $\mathbb{B}$ et les vecteurs $\boldsymbol{c}$ en fonction de $\mathbb{D}$, $\mathbb{E}$, $\mathbb{F}$ et $\boldsymbol{b}$.

Correction Exercice 3.4.1 On rappelle que la i-ème équation du système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ s'écrit

$$b_i = \sum_{j=1}^{n} \mathrm{A}_{i,j} x_j = \sum_{j=1}^{i-1} \mathrm{A}_{i,j} x_j + \mathrm{A}_{i,i} x_i + \sum_{j=i+1}^{n} \mathrm{A}_{i,j} x_j.$$

et que ceci correspond à l'écriture matricielle

$$\boldsymbol{b} = \mathbb{D}\boldsymbol{x} - \mathbb{E}\boldsymbol{x} - \mathbb{F}\boldsymbol{x}$$

- Pour la **méthode de Jacobi** on a, $\forall i \in [\![1, n]\!]$,

$$b_i = \mathrm{A}_{i,i} x_i^{[k+1]} + \sum_{j=1}^{i-1} \mathrm{A}_{i,j} x_j^{[k]} + \sum_{j=i+1}^{n} \mathrm{A}_{i,j} x_j^{[k]}.$$

ce qui s'écrit matriciellement

$$\boldsymbol{b} = \mathbb{D}\boldsymbol{x}^{[k+1]} - \mathbb{E}\boldsymbol{x}^{[k]} - \mathbb{F}\boldsymbol{x}^{[k]}.$$

On a donc

$$\mathbb{D}\boldsymbol{x}^{[k+1]} = \boldsymbol{b} + (\mathbb{E} + \mathbb{F})\boldsymbol{x}^{[k]}$$

Comme la matrice $\mathbb{D}$ est inversible, on obtient

$$\boldsymbol{x}^{[k+1]} = \mathbb{D}^{-1}(\mathbb{E} + \mathbb{F})\boldsymbol{x}^{[k]} + \mathbb{D}^{-1}\boldsymbol{b}$$

La matrice d'itération de Jacobi est $\mathbb{B} = \mathbb{D}^{-1}(\mathbb{E} + \mathbb{F})$ et le vecteur $\boldsymbol{c} = \mathbb{D}^{-1}\boldsymbol{b}$.

- Pour la **méthode de Gauss-Seidel** on a, $\forall i \in [\![1, n]\!]$,

$$b_i = \mathrm{A}_{i,i} x_i^{[k+1]} + \sum_{j=1}^{i-1} \mathrm{A}_{i,j} x_j^{[k+1]} + \sum_{j=i+1}^{n} \mathrm{A}_{i,j} x_j^{[k]}.$$

ce qui s'écrit matriciellement

$$\boldsymbol{b} = \mathbb{D}\boldsymbol{x}^{[k+1]} - \mathbb{E}\boldsymbol{x}^{[k+1]} - \mathbb{F}\boldsymbol{x}^{[k]}.$$

On a donc

$$(\mathbb{D} - \mathbb{E})\boldsymbol{x}^{[k+1]} = \boldsymbol{b} + \mathbb{F}\boldsymbol{x}^{[k]}$$

Comme la matrice $\mathbb{D} - \mathbb{E}$ est inversible (matrice triangulaire inférieure d'éléments diagonaux non nuls), on obtient

$$\boldsymbol{x}^{[k+1]} = (\mathbb{D} - \mathbb{E})^{-1}\mathbb{F}\boldsymbol{x}^{[k]} + (\mathbb{D} - \mathbb{E})^{-1}\boldsymbol{b}$$

La matrice d'itération de Gauss-Seidel est $\mathbb{B} = (\mathbb{D} - \mathbb{E})^{-1}\mathbb{F}$ et le vecteur $\boldsymbol{c} = (\mathbb{D} - \mathbb{E})^{-1}\boldsymbol{b}$.

- Pour la **méthode S.O.R.** on a , $\forall i \in [\![1, n]\!]$,

$$x_i^{[k+1]} = \frac{w}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1}^{i-1} \mathrm{A}_{ij} x_j^{[k+1]} - \sum_{j=i+1}^{n} \mathrm{A}_{ij} x_j^{[k]} \right) + (1-w) x_i^{[k]}$$

ce qui s'écrit aussi

$$\frac{\mathrm{A}_{ii}}{w}x_i^{[k+1]} + \sum_{j=1}^{i-1}\mathrm{A}_{ij}x_j^{[k+1]} = b_i - \sum_{j=i+1}^{n}\mathrm{A}_{ij}x_j^{[k]} + \frac{1-w}{w}\mathrm{A}_{ii}x_i^{[k]}$$

et matriciellement on obtient

$$\left(\frac{\mathbb{D}}{w} - \mathbb{E}\right)\boldsymbol{x}^{[k+1]} = \left(\frac{1-w}{w}\mathbb{D} + \mathbb{F}\right)\boldsymbol{x}^{[k]} + \boldsymbol{b}.$$

Comme la matrice $\left(\frac{\mathbb{D}}{w} - \mathbb{E}\right)$ est inversible (car triangulaire inférieure à éléments diagonaux non nuls), on a

$$\boldsymbol{x}^{[k+1]} = \left(\frac{\mathbb{D}}{w} - \mathbb{E}\right)^{-1}\left(\frac{1-w}{w}\mathbb{D} + \mathbb{F}\right)\boldsymbol{x}^{[k]} + \left(\frac{\mathbb{D}}{w} - \mathbb{E}\right)^{-1}\boldsymbol{b}$$

La matrice d'itération de S.O.R. est $\mathbb{B} = \left(\frac{\mathbb{D}}{w} - \mathbb{E}\right)^{-1}\left(\frac{1-w}{w}\mathbb{D} + \mathbb{F}\right)$ et le vecteur $\boldsymbol{c} = \left(\frac{\mathbb{D}}{w} - \mathbb{E}\right)^{-1}\boldsymbol{b}$.

◇

Proposition 3.44

Soit $\mathbb{A}$ une matrice régulière telle que tous ses éléments diagonaux soient non nuls. On note $\mathbb{D} = \mathrm{diag}(\mathbb{A})$ et $\mathbb{E}$, $\mathbb{F}$, les matrices à diagonales nulles respectivement triangulaire inférieure et supérieure telles que $\mathbb{A} = \mathbb{D} - \mathbb{E} - \mathbb{F}$. On pose $\mathbb{L} = \mathbb{D}^{-1}\mathbb{E}$ et $\mathbb{U} = \mathbb{D}^{-1}\mathbb{F}$.
La matrice d'itération de la méthode de Jacobi, notée $\mathbb{J}$, est donnée par

$$\mathbb{J} = \mathbb{D}^{-1}(\mathbb{E} + \mathbb{F}) = \mathbb{L} + \mathbb{U}, \tag{3.71}$$

La matrice d'itération de la méthode S.O.R., notée $\mathcal{L}_w$, est donnée par

$$\mathcal{L}_w = \left(\frac{\mathbb{D}}{w} - \mathbb{E}\right)^{-1}\left(\frac{1-w}{w}\mathbb{D} + \mathbb{F}\right) = (\mathbb{I} - w\mathbb{L})^{-1}\left((1-w)\mathbb{I} + w\mathbb{U}\right). \tag{3.72}$$

et elle vérifie

$$\rho(\mathcal{L}_w) \geqslant |w-1|. \tag{3.73}$$

La matrice d'itération de Gauss-Seidel est $\mathcal{L}_1$ et elle correspond à

$$\mathcal{L}_1 = (\mathbb{D} - \mathbb{E})^{-1}\mathbb{F} = (\mathbb{I} - \mathbb{L})^{-1}\mathbb{U}. \tag{3.74}$$

Preuve. Les résultats découlent de l'Exercice 3.4.1 pour l'écriture en fonction des matrices $\mathbb{D}$, $\mathbb{E}$ et $\mathbb{F}$. Pour l'écriture en fonction des matrices $\mathbb{L}$ et $\mathbb{U}$ seule l'équation (3.72) n'est pas forcément immédiate. On a vu que

$$\mathcal{L}_w = \left(\frac{\mathbb{D}}{w} - \mathbb{E}\right)^{-1}\left(\frac{1-w}{w}\mathbb{D} + \mathbb{F}\right)$$

et comme $\mathbb{E} = \mathbb{D}\mathbb{L}$ et $\mathbb{F} = \mathbb{D}\mathbb{U}$ on obtient

$$\begin{aligned}
\mathcal{L}_w &= \left(\frac{\mathbb{D}}{w} - \mathbb{D}\mathbb{L}\right)^{-1}\left(\frac{1-w}{w}\mathbb{D} + \mathbb{D}\mathbb{U}\right) \\
&= \left(\frac{1}{w}\mathbb{D}[\mathbb{I} - w\mathbb{L}]\right)^{-1}\left(\frac{1}{w}\mathbb{D}[(1-w)\mathbb{I} + w\mathbb{U}]\right) \\
&= (\mathbb{I} - w\mathbb{L})^{-1}\left(\frac{1}{w}\mathbb{D}\right)^{-1}\left(\frac{1}{w}\mathbb{D}\right)((1-w)\mathbb{I} + w\mathbb{U}) \\
&= (\mathbb{I} - w\mathbb{L})^{-1}\left((1-w)\mathbb{I} + w\mathbb{U}\right)
\end{aligned}$$

Il reste à démontrer l'inégalité (3.73). La matrice $\mathbb{L}$ est triangulaire inférieure à diagonale nulle car elle est le produit d'une matrice diagonale (et donc triangulaire inférieure) $\mathbb{D}^{-1}$ et d'une matrice triangulaire

inférieure $\mathbb{E}$ à diagonale nulle. De même la matrice $\mathbb{U}$ est triangulaire supérieure à diagonale nulle.
On sait que le déterminant d'une matrice est égale aux produits de ses valeurs propres comptées avec leurs multiplicités. En notant n la dimension de la matrice $\mathcal{L}_w$, et en notant $\lambda_i(\mathcal{L}_w)$ ses n valeurs propres, on a donc

$$\det(\mathcal{L}_w) = \prod_{i=1}^{n} \lambda_i(\mathcal{L}_w).$$

Le rayon spectrale de $\mathcal{L}_w$, noté $\rho(\mathcal{L}_w)$, correspond au plus grand des modules des valeurs propres. On a alors

$$\rho(\mathcal{L}_w) = \max_{i\in[\![1,n]\!]} |\lambda_i(\mathcal{L}_w)| \geqslant |\det(\mathcal{L}_w)|^{1/n}$$

De plus on a

$$\det(\mathcal{L}_w) = \det\left((\mathbb{I} - w\mathbb{L})^{-1}\left((1-w)\mathbb{I} + w\mathbb{U}\right)\right) = \det\left((\mathbb{I} - w\mathbb{L})^{-1}\right)\det\left(((1-w)\mathbb{I} + w\mathbb{U})\right)$$

La matrice $\mathbb{I} - w\mathbb{L}$ est triangulaire inférieure à diagonale unité donc son inverse aussi. On en déduit $\det\left((\mathbb{I} - w\mathbb{L})^{-1}\right) = 1$. La matrice $(1-w)\mathbb{I}+w\mathbb{U}$ est triangulaire supérieure avec tous ses éléments diagonaux valant $1-w$ et donc $\det\left(((1-w)\mathbb{I} + w\mathbb{U})\right) = (1-w)^n$. On a alors $|\det(\mathcal{L}_w)| = |1-w|^n$ et

$$\rho(\mathcal{L}_w) \geqslant |\det(\mathcal{L}_w)|^{1/n} = |1-w|.$$

□

3.4.3 Etude de la convergence

Théorème 3.45

Soit $\mathbb{A}$ une matrice régulière décomposée sous la forme $\mathbb{A} = \mathbb{M} - \mathbb{N}$ avec $\mathbb{M}$ régulière. On pose

$$\mathbb{B} = \mathbb{M}^{-1}\mathbb{N} \quad \text{et} \quad \boldsymbol{c} = \mathbb{M}^{-1}\boldsymbol{b}.$$

Alors la suite définie par

$$\boldsymbol{x}^{[0]} \in \mathbb{K}^n \quad \text{et} \quad \boldsymbol{x}^{[k+1]} = \mathbb{B}\boldsymbol{x}^{[k]} + \boldsymbol{c}$$

converge vers $\bar{\boldsymbol{x}} = \mathbb{A}^{-1}\boldsymbol{b}$ quelque soit $\boldsymbol{x}^{[0]}$ si et seulement si $\rho(\mathbb{B}) < 1$.

Preuve. Comme $\bar{\boldsymbol{x}} = \mathbb{A}^{-1}\boldsymbol{b}$ (sans présupposer la convergence) on a $\mathbb{M}\bar{\boldsymbol{x}} = \mathbb{N}\bar{\boldsymbol{x}} + \boldsymbol{b}$ et alors

$$\bar{\boldsymbol{x}} = \mathbb{M}^{-1}\mathbb{N}\bar{\boldsymbol{x}} + \mathbb{M}^{-1}\boldsymbol{b} = \mathbb{B}\bar{\boldsymbol{x}} + \boldsymbol{c}$$

On obtient donc

$$\bar{\boldsymbol{x}} - \boldsymbol{x}^{[k+1]} = \mathbb{B}(\bar{\boldsymbol{x}} - \boldsymbol{x}^{[k]})$$

Or la suite $\boldsymbol{x}^{[k]}$ converge vers $\bar{\boldsymbol{x}}$ si et seulement si la suite $\boldsymbol{e}^{[k]} \overset{\text{def}}{=} \bar{\boldsymbol{x}} - \boldsymbol{x}^{[k]}$ converge vers $\mathbf{0}$. On a

$$\boldsymbol{e}^{[k]} = \mathbb{B}^k\boldsymbol{e}^{[0]}, \quad \forall k \in \mathbb{N}.$$

D'après le Théoréme 3.37, page 103, on a $\lim_{k\to+\infty} \mathbb{B}^k\boldsymbol{e}^{[0]} = 0$, $\forall \boldsymbol{e}^{[0]} \in \mathbb{K}^n$ si et seulement si $\rho(\mathbb{B}) < 1$. □

Corollaire 3.46

Soit $\mathbb{A}$ une matrice vérifiant $\mathrm{A}_{i,i} \neq 0$ $\forall i$. Une condition nécessaire de convergence pour la méthode S.O.R. est que $0 < w < 2$.

Preuve. On a vu en (voir Proposition 3.44, page 109) que $\rho(\mathcal{L}_w) \geqslant |w-1|$. Donc si $\rho(\mathcal{L}_w) \geqslant 1$, la non-convergence est certaine d'après le Théoréme 3.37, page 103. Une condition nécessaire (mais non suffisante) de convergence est que $|w-1| < 1$ i.e. $w \in]0,2[$. □

Théorème 3.47

Soit $\mathbb{A}$ une matrice à diagonale strictement dominante ou une matrice inversible à diagonale fortement dominante alors

- la méthode de Jacobi est convergente,
- si $w \in]0,1]$ la méthode de Relaxation est convergente.

Preuve. voir [7], vol.2, Théorème 19 et 20, pages 346 à 349. □

Théorème 3.48

Soit $\mathbb{A}$ une matrice hermitienne inversible en décomposée en $\mathbb{A} = \mathbb{M} - \mathbb{N}$ où $\mathbb{M}$ est inversible. Soit $\mathbb{B} = \mathbb{I} - \mathbb{M}^{-1}\mathbb{A}$, la matrice de l'itération. Supposons que $\mathbb{M}^* + \mathbb{N}$ (qui est hermitienne) soit définie positive. Alors $\rho(\mathbb{B}) < 1$ si et seulement si $\mathbb{A}$ est définie positive.

Preuve. (voir Exercice 3.4.2, page 111) □

Exercice 3.4.2

Soit $\mathbb{A} \in \mathcal{M}_{n,n}(\mathbb{C})$ une matrice hermitienne inversible décomposée en $\mathbb{A} = \mathbb{M} - \mathbb{N}$ où $\mathbb{M}$ est inversible. On note $\mathbb{B} = \mathbb{I} - \mathbb{M}^{-1}\mathbb{A}$.

Q. 1 *Montrer que la matrice* $\mathbb{M}^* + \mathbb{N}$ *est hermitienne.*

On suppose maintenant que $\mathbb{M}^* + \mathbb{N}$ est définie positive.

Q. 2 *Soit* $\boldsymbol{x}$ *un vecteur quelconque de* $\mathbb{C}^n$ *et* $\boldsymbol{y} = \mathbb{B}\boldsymbol{x}$.

1. *Montrer que*

$$\langle \boldsymbol{x}, \mathbb{A}\boldsymbol{x}\rangle - \langle \boldsymbol{y}, \mathbb{A}\boldsymbol{y}\rangle = \langle \boldsymbol{x}, \mathbb{A}\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\rangle + \langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{A}\boldsymbol{x}\rangle - \langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{A}\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\rangle \tag{3.75}$$

et

$$\boldsymbol{x} - \boldsymbol{y} = \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}. \tag{3.76}$$

2. *En déduire que*

$$\langle \boldsymbol{x}, \mathbb{A}\boldsymbol{x}\rangle - \langle \boldsymbol{y}, \mathbb{A}\boldsymbol{y}\rangle = \langle (\boldsymbol{x} - \boldsymbol{y}), (\mathbb{M}^* + \mathbb{N})(\boldsymbol{x} - \boldsymbol{y})\rangle. \tag{3.77}$$

Q. 3 *Montrer que si* $\mathbb{A}$ *est définie positive alors* $\rho(\mathbb{B}) < 1$.

Q. 4 *Démontrer par l'absurde que si* $\rho(\mathbb{B}) < 1$ *alors* $\mathbb{A}$ *est définie positive.*

Correction Exercice 3.4.2

Q. 1 On a

$$\begin{aligned}\left(\mathbb{M}^* + \mathbb{N}\right)^* &= \mathbb{M} + \mathbb{N}^* \\ &= \mathbb{A} + \mathbb{N} + \mathbb{N}^* = \mathbb{A}^* + \mathbb{N} + \mathbb{N}^* \quad \text{car } \mathbb{A} \text{ est hermitienne} \\ &= \mathbb{M}^* + \mathbb{N}.\end{aligned}$$

La matrice $\mathbb{M}^* + \mathbb{N}$ est donc hermitienne.

Q. 2 1. On a $\boldsymbol{y} = \mathbb{B}\boldsymbol{x}$ avec $\mathbb{B} = \mathbb{I} - \mathbb{M}^{-1}\mathbb{A}$ ce qui donne

$$\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{x} - \mathbb{B}\boldsymbol{x} = \left(\mathbb{I} - \mathbb{B}\right)\boldsymbol{x} = \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}.$$

L'équation (3.76) est donc démontrée. Pour prouver (3.75), on note que

$$\boldsymbol{y} = \boldsymbol{x} - \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}$$

et donc

$$\begin{aligned}\langle \boldsymbol{y}, \mathbb{A}\boldsymbol{y}\rangle &= \big\langle \boldsymbol{x} - \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{A}\big(\boldsymbol{x} - \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big)\big\rangle \\ &= \langle \boldsymbol{x}, \mathbb{A}\boldsymbol{x}\rangle - \big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{A}\boldsymbol{x}\big\rangle - \big\langle \boldsymbol{x}, \mathbb{A}\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle + \big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{A}\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle.\end{aligned}$$

On en déduit immédiatement (3.75).

2. En utilisant (3.76), on obtient

$$\begin{aligned}\big\langle \boldsymbol{x} - \boldsymbol{y}, \big(\mathbb{M}^* + \mathbb{N}\big)(\boldsymbol{x} - \boldsymbol{y})\big\rangle &= \big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \big(\mathbb{M}^* + \mathbb{N}\big)\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle \\ &= \big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \big(\mathbb{M} + \mathbb{N}^*\big)\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle \quad \text{car } \mathbb{M}^* + \mathbb{N} \text{ hermitienne} \\ &= \big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \big(\mathbb{M} + \mathbb{M}^* - \mathbb{A}^*\big)\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle \quad \text{car } \mathbb{N} = \mathbb{M} - \mathbb{A} \\ &= \big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{A}\boldsymbol{x}\big\rangle + \big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{M}^*\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle - \big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{A}\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle \text{ car } \mathbb{A} \text{ hermitienne}\end{aligned}$$

Or, par propriété du produit scalaire, on a

$$\big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{M}^*\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle = \big\langle \mathbb{M}\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle = \big\langle \mathbb{A}\boldsymbol{x}, \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle = \big\langle \boldsymbol{x}, \mathbb{A}^*\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle.$$

Comme $\mathbb{A}$ est hermitienne, on obtient

$$\big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{M}^*\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle = \big\langle \boldsymbol{x}, \mathbb{A}\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle.$$

On abouti alors à

$$\big\langle \boldsymbol{x} - \boldsymbol{y}, \big(\mathbb{M}^* + \mathbb{N}\big)(\boldsymbol{x} - \boldsymbol{y})\big\rangle = \big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{A}\boldsymbol{x}\big\rangle + \big\langle \boldsymbol{x}, \mathbb{A}\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle - \big\langle \mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}, \mathbb{A}\mathbb{M}^{-1}\mathbb{A}\boldsymbol{x}\big\rangle.$$

L'équation (3.77) est obtenue en utilisant (3.75).

Q. 3 On veut démontrer que sous les hypothèses $\mathbb{A}$ hermitienne définie positive et $\mathbb{M}^* + \mathbb{N}$ (hermitienne) définie positive on a $\rho(\mathbb{B}) < 1$, c'est à dire que pour tout élément propre $(\lambda, \boldsymbol{u})$ de $\mathbb{B}$ alors $|\lambda| < 1$.

Soit $(\lambda, \boldsymbol{u}) \in \mathbb{C} \times \mathbb{C}^n \backslash \{0\}$ un élément propre de $\mathbb{B}$. On a $\mathbb{B}\boldsymbol{u} = \lambda\boldsymbol{u}$. En prenant $\boldsymbol{x} = \boldsymbol{u}$ dans **Q**.2, on a $\boldsymbol{y} = \mathbb{B}\boldsymbol{u} = \lambda\boldsymbol{u}$ et donc $\boldsymbol{x} - \boldsymbol{y} = (1 - \lambda)\boldsymbol{u}$. De (3.77) on obtient

$$\langle \boldsymbol{u}, \mathbb{A}\boldsymbol{u}\rangle - \langle \lambda\boldsymbol{u}, \lambda\mathbb{A}\boldsymbol{u}\rangle = \langle (1 - \lambda)\boldsymbol{u}, (\mathbb{M}^* + \mathbb{N})((1 - \lambda)\boldsymbol{u})\rangle$$

c'est à dire

$$(1 - |\lambda|^2)\langle \boldsymbol{u}, \mathbb{A}\boldsymbol{u}\rangle = |1 - \lambda|^2 \langle \boldsymbol{u}, (\mathbb{M}^* + \mathbb{N})\boldsymbol{u}\rangle. \tag{3.78}$$

Comme par hypothèse, la matrice $\mathbb{M}^* + \mathbb{N}$ (hermitienne) est définie positive et $\boldsymbol{u} \neq 0$, on obtient

$$\langle \boldsymbol{u}, (\mathbb{M}^* + \mathbb{N})\boldsymbol{u}\rangle > 0$$

et donc on déduit de (3.78) que

$$(1 - |\lambda|^2)\langle \boldsymbol{u}, \mathbb{A}\boldsymbol{u}\rangle \geqslant 0.$$

Comme $\mathbb{A}$ hermitienne définie positive et $\boldsymbol{u} \neq 0$, on déduit de (3.78) que $1 - |\lambda|^2 \geqslant 0$ et donc $|\lambda| \leqslant 1$.

On va démontrer par l'absurde que $|\lambda| \neq 1$. On suppose $|\lambda| = 1$, de (3.78) on déduit

$$|1 - \lambda|^2 \langle \boldsymbol{u}, (\mathbb{M}^* + \mathbb{N})\boldsymbol{u}\rangle = 0.$$

Comme par hypothèse, la matrice $\mathbb{M}^*+\mathbb{N}$ (hermitienne) est définie positive et $\boldsymbol{u} \neq 0$, on obtient $|1-\lambda|^2 = 0$, c'est à dire $\lambda = 1$. Or dans ce cas on a

$$\boldsymbol{y} = \mathbb{B}\boldsymbol{x} = \mathbb{B}\boldsymbol{u} = \lambda\boldsymbol{u} = \boldsymbol{u}.$$

L'équation (3.76) donne alors

$$0 = \mathbb{M}^{-1}\mathbb{A}\boldsymbol{u}.$$

Comme $\boldsymbol{u} \neq 0$, et que $\mathbb{M}^{-1}\mathbb{A}$ est inversible, l'équation ci-dessus est impossible, donc $|\lambda| \neq 1$.

Q. 4 On veut démontrer par l'absurde que, sous les hypothèses $\mathbb{M}^* + \mathbb{N}$ (hermitienne) définie positive, $\mathbb{A}$ hermitienne inversible et $\rho(\mathbb{B}) < 1$, on a $\mathbb{A}$ définie positive.

On suppose que $\mathbb{A}$ n'est pas définie positive. Alors il existe $\boldsymbol{x} \in \mathbb{C}^n \backslash \{0\}$ tel que $\langle \boldsymbol{x}, \mathbb{A}\boldsymbol{x} \rangle \notin \mathbb{R}^{+*}$.

Pour une matrice quelconque que $\langle \boldsymbol{x}, \mathbb{A}\boldsymbol{x} \rangle \in \mathbb{C}$, or comme $\mathbb{A}$ est hermitienne on a $\langle \boldsymbol{x}, \mathbb{A}\boldsymbol{x} \rangle \in \mathbb{R}$. En effet, par propriété du produit scalaire on a

$$\langle \boldsymbol{x}, \mathbb{A}\boldsymbol{x} \rangle = \langle \mathbb{A}^* \boldsymbol{x}, \boldsymbol{x} \rangle = \langle \mathbb{A}\boldsymbol{x}, \boldsymbol{x} \rangle = \overline{\langle \boldsymbol{x}, \mathbb{A}\boldsymbol{x} \rangle}.$$

On note $\boldsymbol{x}^{[0]} = \boldsymbol{x}$ et $\alpha_0 = \left\langle \boldsymbol{x}^{[0]}, \mathbb{A}\boldsymbol{x}^{[0]} \right\rangle$ le nombre réel négatif ou nul ($\alpha_0 \leqslant 0$). On défini alors pour les suites

$$\boldsymbol{x}^{[k]} = \mathbb{B}\boldsymbol{x}^{[k-1]} \quad \text{et} \quad \alpha_k = \left\langle \boldsymbol{x}^{[k]}, \mathbb{A}\boldsymbol{x}^{[k]} \right\rangle.$$

On a alors

$$\boldsymbol{x}^{[k]} = \mathbb{B}^k \boldsymbol{x}^{[0]}, \ \forall k \in \mathbb{N}.$$

D'après le théorème du cours (Théorème B.72 page 195),

$$\rho(\mathbb{B}) < 1 \iff \lim_{k \to +\infty} \mathbb{B}^k \boldsymbol{v}, \ \forall \boldsymbol{v}.$$

On a donc

$$\lim_{k \to +\infty} \boldsymbol{x}^{[k]} = 0 \quad \text{et} \quad \lim_{k \to +\infty} \alpha_k = 0$$

On utilise maintenant l'égalité (3.77) avec $\boldsymbol{x} = \boldsymbol{x}^{[k-1]}$ et $\boldsymbol{y} = \mathbb{B}\boldsymbol{x} = \boldsymbol{x}^{[k]}$

$$\left\langle \boldsymbol{x}^{[k-1]}, \mathbb{A}\boldsymbol{x}^{[k-1]} \right\rangle - \left\langle \boldsymbol{x}^{[k]}, \mathbb{A}\boldsymbol{x}^{[k]} \right\rangle = \left\langle (\boldsymbol{x}^{[k-1]} - \boldsymbol{x}^{[k]}), (\mathbb{M}^* + \mathbb{N})(\boldsymbol{x}^{[k-1]} - \boldsymbol{x}^{[k]}) \right\rangle$$

On peut noter que $\boldsymbol{x}^{[k-1]} - \boldsymbol{x}^{[k]} \neq 0$, car sinon $\boldsymbol{x}^{[k-1]} = \boldsymbol{x}^{[k]} = \mathbb{B}\boldsymbol{x}^{[k-1]}$ et $\lambda = 1$ serait valeur propre de $\mathbb{B}$. il faudrait montrer que $\boldsymbol{x}^{[k-1]} \neq 0$

Dans ce cas, comme $\mathbb{M}^* + \mathbb{N}$ est définie positive, on obtient

$$\left\langle \boldsymbol{x}^{[k-1]}, \mathbb{A}\boldsymbol{x}^{[k-1]} \right\rangle - \left\langle \boldsymbol{x}^{[k]}, \mathbb{A}\boldsymbol{x}^{[k]} \right\rangle = \alpha_{k-1} - \alpha_k > 0.$$

La suite $(\alpha_k)_{k \in \mathbb{N}}$ est donc strictement décroissante de premier terme $\alpha_0 \leqslant 0$: elle ne peut converger vers 0.

Montrons par récurrence forte que $\boldsymbol{x}^{[k]} \neq 0$, $\boldsymbol{x}^{[k]} - \boldsymbol{x}^{[k-1]} \neq 0$, et $0 \geqslant \alpha_{k-1} > \alpha_k$, $k \in \mathbb{N}^*$.

- **Initialisation :** On a $\boldsymbol{x}^{[0]} \neq 0$ et $\boldsymbol{x}^{[1]} = \mathbb{B}\boldsymbol{x}^{[0]}$.
  On montre par l'absurde que $\boldsymbol{x}^{[1]} \neq 0$. Supposons $\boldsymbol{x}^{[1]} = 0$, alors $\alpha_1 = 0$ et $\boldsymbol{x}^{[0]} - \boldsymbol{x}^{[1]} = \boldsymbol{x}^{[0]} \neq 0$. Comme $\mathbb{M}^* + \mathbb{N}$ est hermitienne définie positive on obtient

  $$\alpha_0 - \alpha_1 = \left\langle (\boldsymbol{x}^{[0]} - \boldsymbol{x}^{[1]}), (\mathbb{M}^* + \mathbb{N})(\boldsymbol{x}^{[0]} - \boldsymbol{x}^{[1]}) \right\rangle > 0$$

  et contradiction avec $\alpha_0 \leqslant 0$.
  On montre ensuite par l'absurde que $\boldsymbol{x}^{[0]} \neq \boldsymbol{x}^{[1]}$. Supposons $\boldsymbol{x}^{[1]} = \boldsymbol{x}^{[0]}$. Par construction $\boldsymbol{x}^{[1]} = \mathbb{B}\boldsymbol{x}^{[0]}$, et dans ce cas, comme $\boldsymbol{x}^{[0]} \neq 0$, $(1, \boldsymbol{x}^{[0]})$ serait un élément propre de $\mathbb{B}$: contradiction avec $\rho(\mathbb{B}) < 1$.
  Comme $\boldsymbol{x}^{[0]} - \boldsymbol{x}^{[1]} \neq 0$, on a

  $$\alpha_0 - \alpha_1 = \left\langle (\boldsymbol{x}^{[0]} - \boldsymbol{x}^{[1]}), (\mathbb{M}^* + \mathbb{N})(\boldsymbol{x}^{[0]} - \boldsymbol{x}^{[1]}) \right\rangle > 0$$

  et donc $0 \geqslant \alpha_0 > \alpha_1$.

- **Hérédité :** On suppose la propriété vraie jusqu'au rang k. On a alors $\boldsymbol{x}^{[k]} \neq 0$, $\boldsymbol{x}^{[k+1]} = \mathbb{B}\boldsymbol{x}^{[k]}$ et $\alpha_k \leqslant 0$.
  On montre par l'absurde que $\boldsymbol{x}^{[k+1]} \neq 0$. Supposons $\boldsymbol{x}^{[k+1]} = 0$, alors $\alpha_{k+1} = 0$ et $\boldsymbol{x}^{[k]} - \boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]} \neq 0$. Comme $\mathbb{M}^* + \mathbb{N}$ est hermitienne définie positive on obtient

  $$\alpha_k - \alpha_{k+1} = \left\langle (\boldsymbol{x}^{[k]} - \boldsymbol{x}^{[k+1]}), (\mathbb{M}^* + \mathbb{N})(\boldsymbol{x}^{[k]} - \boldsymbol{x}^{[k+1]}) \right\rangle > 0$$

  et contradiction avec $\alpha_k \leqslant 0$.
  On montre ensuite par l'absurde que $\boldsymbol{x}^{[k]} \neq \boldsymbol{x}^{[k+1]}$. Supposons $\boldsymbol{x}^{[k+1]} = \boldsymbol{x}^{[k]}$. Par construction

$\boldsymbol{x}^{[k+1]} = \mathbb{B}\boldsymbol{x}^{[k]}$, et dans ce cas, comme $\boldsymbol{x}^{[k]} \neq 0$, $(1, \boldsymbol{x}^{[k]})$ serait un élément propre de $\mathbb{B}$: contradiction avec $\rho(\mathbb{B}) < 1$.
Comme $\boldsymbol{x}^{[k]} - \boldsymbol{x}^{[k+1]} \neq 0$, on a

$$\alpha_k - \alpha_{k+1} = \left\langle (\boldsymbol{x}^{[k]} - \boldsymbol{x}^{[k+1]}), (\mathbb{M}^* + \mathbb{N})(\boldsymbol{x}^{[k]} - \boldsymbol{x}^{[k+1]}) \right\rangle > 0$$

et donc $0 \geqslant \alpha_k > \alpha_{k+1}$.

◇

Théorème 3.49

Soit $\mathbb{A}$ une matrice hermitienne définie positive, alors la méthode de relaxation converge si et seulement si $w \in]0, 2[$.

Preuve. voir [7], vol.2, Corollaire 24, page 351. □

3.4.4 Algorithmes

Nous allons écrire des algorithmes pour chacune des méthodes itératives proposées. L'objectif n'est pas d'écrire des algorithmes "optimisés" mais de voir la méthodologie permettant la construction d'algorithmes simples et fonctionnels.

Principe de base

Les méthodes itératives pour la résolution d'un système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ s'écrivent sous la forme

$$\boldsymbol{x}^{[0]} \in \mathbb{K}^n \quad \text{et} \quad \boldsymbol{x}^{[k+1]} = \mathbb{B}\boldsymbol{x}^{[k]} + \boldsymbol{c}$$

La matrice d'itération $\mathbb{B}$ et le vecteur $\boldsymbol{c}$ sont construits de telle sorte que si la suite $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ converge vers $\bar{\boldsymbol{x}}$ alors $\bar{\boldsymbol{x}}$ est aussi solution du système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$.

Comme pour les méthodes de point fixe, La convergence de ces méthodes n'est pas assurée et si il y a convergence le nombre d'itération nécessaire n'est (à priori) pas connu. C'est pourquoi algorithmiquement on utilise une boucle **Tantque**. On défini alors un **nombre maximum d'itérations** au delà du quel les calculs itératifs sont stoppés et une valeur $\varepsilon > 0$ permettant d'arrêter les calculs lorsque $\boldsymbol{x}^{[k]}$ est suffisament proche de $\bar{\boldsymbol{x}}$. Pour être plus précis, on note $\boldsymbol{r}^{[k]} = \boldsymbol{b} - \mathbb{A}\boldsymbol{x}^{[k]}$ le résidu. On a alors

$$\boldsymbol{r}^{[k]} = \mathbb{A}\bar{\boldsymbol{x}} - \mathbb{A}\boldsymbol{x}^{[k]} = \mathbb{A}\boldsymbol{e}^{[k]}$$

Si on prend comme critère d'arrêt

$$\frac{\|\boldsymbol{r}^{[k]}\|}{\|\boldsymbol{b}\|} \leqslant \varepsilon$$

alors

$$\left\|\boldsymbol{e}^{[k]}\right\| = \left\|\mathbb{A}^{-1}\boldsymbol{r}^{[k]}\right\| \leqslant \left\|\mathbb{A}^{-1}\right\| \left\|\boldsymbol{r}^{[k]}\right\| \leqslant \varepsilon \left\|\mathbb{A}^{-1}\right\| \|\boldsymbol{b}\| = \varepsilon \left\|\mathbb{A}^{-1}\right\| \|\mathbb{A}\bar{\boldsymbol{x}}\| \leqslant \varepsilon \left\|\mathbb{A}^{-1}\right\| \|\mathbb{A}\| \|\bar{\boldsymbol{x}}\| = \varepsilon \operatorname{cond}(\mathbb{A}) \|\bar{\boldsymbol{x}}\|$$

et donc

$$\frac{\|\boldsymbol{e}^{[k]}\|}{\|\bar{\boldsymbol{x}}\|} \leqslant \varepsilon \operatorname{cond}(\mathbb{A})$$

Pour éviter des soucis lorsque $\|\boldsymbol{b}\|$ est proche de zéro, on peut utiliser comme critère d'arrêt de convergence

$$\frac{\|\boldsymbol{r}^{[k]}\|}{\|\boldsymbol{b}\| + 1} \leqslant \varepsilon.$$

Comme vu avec les méthodes itératives de type point fixe (voir section 2.2.3, page 30), il n'est pas forcément utile de stocker l'intégralité des termes de la suite $\boldsymbol{x}^{[k]}$ puisque seul le "dernier" nous intéresse. On a alors L'Algorithme 3.19 correspondant à un algorithme itératif générique sans stockage des valeurs intermédiaires.

Algorithme 3.19 Méthode itérative pour la résolution d'un système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$

Données :
$\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$,
$\boldsymbol{b}$: vecteur de $\mathbb{K}^n$,
$\boldsymbol{x}^0$: vecteur initial de $\mathbb{K}^n$,
ε : la tolérence, $\varepsilon \in \mathbb{R}^+$,
kmax : nombre maximum d'itérations, kmax $\in \mathbb{N}^*$
Résultat :
$\boldsymbol{x}^{\text{tol}}$: un vecteur de $\mathbb{K}^n$ si convergence, sinon $\varnothing$

1: $k \leftarrow 0$, $\boldsymbol{x}^{\text{tol}} \leftarrow \varnothing$
2: $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
3: tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
4: **Tantque** $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
5: $k \leftarrow k + 1$
6: $\boldsymbol{p} \leftarrow \boldsymbol{x}$ ▷ $\boldsymbol{p}$ contient le vecteur précédent
7: $\boldsymbol{x} \leftarrow$ calcul de l'itérée suivante en fonction de $\boldsymbol{p}$, $\mathbb{A}$, $\boldsymbol{b}$, ...
8: $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
9: **Fin Tantque**
10: **Si** $\|\boldsymbol{r}\| \leqslant$ tol **alors** ▷ Convergence
11: $\boldsymbol{x}^{\text{tol}} \leftarrow \boldsymbol{x}$
12: **Fin Si**

Méthode de Jacobi

Pour Jacobi, la suite des itérées est définie par

$$x_i^{[k+1]} = \frac{1}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{n} \mathrm{A}_{ij} x_j^{[k]} \right), \quad \forall i \in [\![1, n]\!].$$

Cette formule donne explicitement les composantes du vecteur $\boldsymbol{x}^{[k+1]}$ en fonction de la matrice $\mathbb{A}$, et des vecteurs $\boldsymbol{b}$ et $\boldsymbol{x}^{[k]}$. A partir de l'Algorithme 3.19, on va construire par raffinements successifs la RSLJacobi donnée dans l'Algorithme 3.20.

Algorithme 3.20 $\mathcal{R}_0$

1: $k \leftarrow 0$, $\boldsymbol{x}^{\text{tol}} \leftarrow \varnothing$
2: $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
3: tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
4: **Tantque** $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
5: $k \leftarrow k + 1$
6: $\boldsymbol{p} \leftarrow \boldsymbol{x}$
7:
8: $\boldsymbol{x} \leftarrow$ calcul par Jacobi
9: $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
10: **Fin Tantque**
11: **Si** $\|\boldsymbol{r}\| \leqslant$ tol **alors** ▷ Convergence
12: $\boldsymbol{x}^{\text{tol}} \leftarrow \boldsymbol{x}$
13: **Fin Si**

Algorithme 3.20 $\mathcal{R}_1$

1: $k \leftarrow 0$, $\boldsymbol{x}^{\text{tol}} \leftarrow \varnothing$
2: $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
3: tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
4: **Tantque** $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
5: $k \leftarrow k + 1$
6: $\boldsymbol{p} \leftarrow \boldsymbol{x}$
7: **Pour** $i \leftarrow 1$ **à** n **faire**
8: $x_i \leftarrow \frac{1}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{n} \mathrm{A}_{ij} \boldsymbol{p}_j \right)$
9: **Fin Pour**
10: $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
11: **Fin Tantque**
12: **Si** $\|\boldsymbol{r}\| \leqslant$ tol **alors** ▷ Convergence
13: $\boldsymbol{x}^{\text{tol}} \leftarrow \boldsymbol{x}$
14: **Fin Si**

Algorithme 3.20 $\mathcal{R}_1$

$k \leftarrow 0$, $\boldsymbol{x}^{\text{tol}} \leftarrow \varnothing$
$\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
Tantque $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
  $k \leftarrow k + 1$
  $\boldsymbol{p} \leftarrow \boldsymbol{x}$
  Pour $i \leftarrow 1$ **à** n **faire**

$$x_i \leftarrow \frac{1}{\mathrm{A}_{ii}}\left(b_i - \sum_{j=1, j\neq i}^{n} \mathrm{A}_{ij}\boldsymbol{p}_j\right)$$

  Fin Pour
  $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
Fin Tantque
Si $\|\boldsymbol{r}\| \leqslant$ tol **alors**
  $\boldsymbol{x}^{\text{tol}} \leftarrow \boldsymbol{x}$
Fin Si

Algorithme 3.20 $\mathcal{R}_2$

$k \leftarrow 0$, $\boldsymbol{x}^{\text{tol}} \leftarrow \varnothing$
$\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
Tantque $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
  $k \leftarrow k + 1$
  $\boldsymbol{p} \leftarrow \boldsymbol{x}$
  Pour $i \leftarrow 1$ **à** n **faire**
    S $\leftarrow 0$
    Pour $j \leftarrow 1$ **à** n $(j \neq i)$ **faire**
      S $\leftarrow$ S $+ \mathrm{A}_{i,j}p_j$
    Fin Pour
    $x_i \leftarrow \frac{1}{\mathrm{A}_{ii}}(b_i - \mathrm{S})$
  Fin Pour
  $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
Fin Tantque
Si $\|\boldsymbol{r}\| \leqslant$ tol **alors**
  $\boldsymbol{x}^{\text{tol}} \leftarrow \boldsymbol{x}$
Fin Si

On peut alors écrire la fonction RSLJacobi :

Algorithme 3.20 Méthode itérative de Jacobi pour la résolution d'un système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$

Données :
$\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$,
$\boldsymbol{b}$: vecteur de $\mathbb{K}^n$,
$\boldsymbol{x}^0$: vecteur initial de $\mathbb{K}^n$,
ε : la tolérence, $\varepsilon \in \mathbb{R}^+$,
kmax : nombre maximum d'itérations, kmax $\in \mathbb{N}^*$
Résultat :
$\boldsymbol{X}$: un vecteur de $\mathbb{K}^n$

Fonction $\boldsymbol{X} \leftarrow$ RSLJacobi ($\mathbb{A}, \boldsymbol{b}, \boldsymbol{x}^0, \varepsilon,$ kmax)
  $k \leftarrow 0$, $\boldsymbol{X} \leftarrow \varnothing$
  $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
  tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
  Tantque $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
    $k \leftarrow k + 1$
    $\boldsymbol{p} \leftarrow \boldsymbol{x}$
    Pour $i \leftarrow 1$ **à** n **faire**
      S $\leftarrow 0$
      Pour $j \leftarrow 1$ **à** n $(j \neq i)$ **faire**
        S $\leftarrow$ S $+ A(i,j) * p(j)$
      Fin Pour
      $x(i) \leftarrow (b(i) - \mathrm{S})/A(i,i)$
    Fin Pour
    $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
  Fin Tantque
  Si $\|\boldsymbol{r}\| \leqslant$ tol **alors**
    $\boldsymbol{X} \leftarrow \boldsymbol{x}$
  Fin Si
Fin Fonction

Méthode de Gauss-Seidel

Pour Gauss-Seidel, la suite des itérées est définie par

$$x_i^{(k+1)} = \frac{1}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1}^{i-1} \mathrm{A}_{ij} x_j^{[k+1]} - \sum_{j=i+1}^{n} \mathrm{A}_{ij} x_j^{[k]} \right) \quad \forall i \in [\![1, n]\!]$$

Cette formule donne explicitement la composante i du vecteur $\boldsymbol{x}^{[k+1]}$ en fonction de la matrice $\mathbb{A}$, et des vecteurs $\boldsymbol{b}$ et $\boldsymbol{x}^{[k]}$, mais aussi des $i-1$ premières composantes de $\boldsymbol{x}^{[k+1]}$. Contrairement à la méthode de Jacobi, il est impératif de calculer sussessivement $x_1^{[k+1]}$, $x_2^{[k+1]}$, ... A partir de l'Algorithme 3.19, on va construire par raffinements successifs la RSLGaussSeidel donnée dans l'Algorithme 3.21.

Algorithme 3.21 $\mathcal{R}_0$

1: $k \leftarrow 0$, $\boldsymbol{x}^{\mathrm{tol}} \leftarrow \varnothing$
2: $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
3: tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
4: **Tantque** $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
5: $k \leftarrow k + 1$
6: $\boldsymbol{p} \leftarrow \boldsymbol{x}$
7:
8: $\boldsymbol{x} \leftarrow$ calcul par Gauss-Seidel
9: $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
10: **Fin Tantque**
11: **Si** $\|\boldsymbol{r}\| \leqslant$ tol **alors**
12: $\boldsymbol{x}^{\mathrm{tol}} \leftarrow \boldsymbol{x}$
13: **Fin Si**

Algorithme 3.21 $\mathcal{R}_1$

1: $k \leftarrow 0$, $\boldsymbol{x}^{\mathrm{tol}} \leftarrow \varnothing$
2: $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
3: tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
4: **Tantque** $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
5: $k \leftarrow k + 1$
6: $\boldsymbol{p} \leftarrow \boldsymbol{x}$
7: **Pour** $i \leftarrow 1$ **à** n **faire**
8: $x_i \leftarrow \frac{1}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1}^{i-1} \mathrm{A}_{ij} x_j - \sum_{j=i+1}^{n} \mathrm{A}_{ij} p_j \right)$
9: **Fin Pour**
10: $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
11: **Fin Tantque**
12: **Si** $\|\boldsymbol{r}\| \leqslant$ tol **alors**
13: $\boldsymbol{x}^{\mathrm{tol}} \leftarrow \boldsymbol{x}$
14: **Fin Si**

Algorithme 3.21 $\mathcal{R}_0$

1: $k \leftarrow 0$, $\boldsymbol{x}^{\mathrm{tol}} \leftarrow \varnothing$
2: $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
3: tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
4: **Tantque** $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
5: $k \leftarrow k + 1$
6: $\boldsymbol{p} \leftarrow \boldsymbol{x}$
7:
8: $\boldsymbol{x} \leftarrow$ calcul par Gauss-Seidel
9: $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
10: **Fin Tantque**
11: **Si** $\|\boldsymbol{r}\| \leqslant$ tol **alors**
12: $\boldsymbol{x}^{\mathrm{tol}} \leftarrow \boldsymbol{x}$
13: **Fin Si**

Algorithme 3.21 $\mathcal{R}_1$

1: $k \leftarrow 0$, $\boldsymbol{x}^{\mathrm{tol}} \leftarrow \varnothing$
2: $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
3: tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
4: **Tantque** $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
5: $k \leftarrow k + 1$
6: $\boldsymbol{p} \leftarrow \boldsymbol{x}$
7: **Pour** $i \leftarrow 1$ **à** n **faire**
8: $x_i \leftarrow \frac{1}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1}^{i-1} \mathrm{A}_{ij} x_j - \sum_{j=i+1}^{n} \mathrm{A}_{ij} p_j \right)$
9: **Fin Pour**
10: $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
11: **Fin Tantque**
12: **Si** $\|\boldsymbol{r}\| \leqslant$ tol **alors**
13: $\boldsymbol{x}^{\mathrm{tol}} \leftarrow \boldsymbol{x}$
14: **Fin Si**

On peut alors écrire la fonction RSLGaussSeidel :

Algorithme 3.21 Méthode itérative de Gauss-Seidel pour la résolution d'un système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$

Données :
$\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$,
$\boldsymbol{b}$: vecteur de $\mathbb{K}^n$,
$\boldsymbol{x}^0$: vecteur initial de $\mathbb{K}^n$,
ε : la tolérence, $\varepsilon \in \mathbb{R}^+$,
kmax : nombre maximum d'itérations, kmax $\in \mathbb{N}^*$
Résultat :
$\boldsymbol{X}$: un vecteur de $\mathbb{K}^n$

1: **Fonction** $\boldsymbol{X} \leftarrow$ RSLGaussSeidel $(\mathbb{A}, \boldsymbol{b}, \boldsymbol{x}^0, \varepsilon, \text{kmax})$
2: $k \leftarrow 0$, $\boldsymbol{X} \leftarrow \varnothing$
3: $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
4: tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
5: **Tantque** $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
6: $k \leftarrow k + 1$
7: $\boldsymbol{p} \leftarrow \boldsymbol{x}$
8: **Pour** $i \leftarrow 1$ **à** n **faire**
9: S $\leftarrow 0$
10: **Pour** $j \leftarrow 1$ **à** $i - 1$ **faire**
11: S $\leftarrow$ S $+ A(i,j) * x(j)$
12: **Fin Pour**
13: **Pour** $j \leftarrow i + 1$ **à** n **faire**
14: S $\leftarrow$ S $+ A(i,j) * p(j)$
15: **Fin Pour**
16: $x(i) \leftarrow (b(i) -$ S$)/A(i,i)$
17: **Fin Pour**
18: $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
19: **Fin Tantque**
20: **Si** $\|\boldsymbol{r}\| \leqslant$ tol **alors**
21: $\boldsymbol{X} \leftarrow \boldsymbol{x}$
22: **Fin Si**
23: **Fin Fonction**

Jeux algorithmiques

On a bien sûr noté que les fonctions RSLJacobi et RSLGaussSeidel ont la même ossature puisque toutes deux basées sur l'Algorithme 3.19 générique. En effet seule la transcription de la ligne 7 de l'Algorithme 3.19 diffère :

Fonction $\boldsymbol{X} \leftarrow$ RSLJacobi $(\mathbb{A}, \boldsymbol{b}, \boldsymbol{x}^0, \varepsilon, \text{kmax})$
  $k \leftarrow 0$, $\boldsymbol{X} \leftarrow \varnothing$
  $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
  tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
  Tantque $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
    $k \leftarrow k + 1$
    $\boldsymbol{p} \leftarrow \boldsymbol{x}$
    Pour $i \leftarrow 1$ **à** n **faire**
      S $\leftarrow 0$
      Pour $j \leftarrow 1$ **à** n $(j \neq i)$ **faire**
        S $\leftarrow$ S $+ A(i,j) * p(j)$
      Fin Pour
      $x(i) \leftarrow (b(i) - \text{S})/A(i,i)$
    Fin Pour
    $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
  Fin Tantque
  Si $\|\boldsymbol{r}\| \leqslant$ tol **alors**
    $\boldsymbol{X} \leftarrow \boldsymbol{x}$
  Fin Si
Fin Fonction

Fonction $\boldsymbol{X} \leftarrow$ RSLGaussSeidel $(\mathbb{A}, \boldsymbol{b}, \boldsymbol{x}^0, \varepsilon, \text{kmax})$
  $k \leftarrow 0$, $\boldsymbol{X} \leftarrow \varnothing$
  $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
  tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
  Tantque $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
    $k \leftarrow k + 1$
    $\boldsymbol{p} \leftarrow \boldsymbol{x}$
    Pour $i \leftarrow 1$ **à** n **faire**
      S $\leftarrow 0$
      Pour $j \leftarrow 1$ **à** $i - 1$ **faire**
        S $\leftarrow$ S $+ A(i,j) * x(j)$
      Fin Pour
      Pour $j \leftarrow i + 1$ **à** n **faire**
        S $\leftarrow$ S $+ A(i,j) * p(j)$
      Fin Pour
      $x(i) \leftarrow (b(i) - \text{S})/A(i,i)$
    Fin Pour
    $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
  Fin Tantque
  Si $\|\boldsymbol{r}\| \leqslant$ tol **alors**
    $\boldsymbol{X} \leftarrow \boldsymbol{x}$
  Fin Si
Fin Fonction

On va écrire les fonctions algorithmiques IterJacobi et IterGaussSeidel permettant le calcul d'une itérée respectivement pour les méthodes de Jacobi et Gauss-Seidel (voir Algorithmes 3.22 et 3.23).

Algorithme 3.22 Itération de Jacobi : calcul de $\boldsymbol{x}$ tel que

$$x_i = \frac{1}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{n} \mathrm{A}_{ij} y_j \right), \quad \forall i \in [\![1, n]\!].$$

Données :
A : matrice de $\mathcal{M}_n(\mathbb{K})$,
b : vecteur de $\mathbb{K}^n$,
y : vecteur de $\mathbb{K}^n$,
Résultat :
x : un vecteur de $\mathbb{K}^n$

```
1: Fonction x ← IterJacobi ( A, b, y )
2:   Pour i ← 1 à n faire
3:     S ← 0
4:     Pour j ← 1 à n (j ≠ i) faire
5:       S ← S + A(i, j) * y(j)
6:     Fin Pour
7:     x(i) ← (b(i) − S)/A(i, i)
8:   Fin Pour
9: Fin Fonction
```

Algorithme 3.23 Itération de Gauss-Seidel : calcul de $\boldsymbol{x}$ tel que

$$x_i = \frac{1}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1}^{i-1} \mathrm{A}_{i,j} x_j - \sum_{j=i+1}^{n} \mathrm{A}_{i,j} y_j \right), \quad \forall i \in [\![1, n]\!].$$

Données :
A : matrice de $\mathcal{M}_n(\mathbb{K})$,
b : vecteur de $\mathbb{K}^n$,
y : vecteur de $\mathbb{K}^n$,
Résultat :
x : un vecteur de $\mathbb{K}^n$

```
1: Fonction x ← IterGaussSeidel ( A, b, y )
2:   Pour i ← 1 à n faire
3:     S ← 0
4:     Pour j ← 1 à i − 1 faire
5:       S ← S + A(i, j) * x(j)
6:     Fin Pour
7:     Pour j ← i + 1 à n faire
8:       S ← S + A(i, j) * y(j)
9:     Fin Pour
10:    x(i) ← (b(i) − S)/A(i, i)
11:  Fin Pour
12: Fin Fonction
```

En utilisant ces deux fonctions, les algorithmes de résolutions de systèmes linéaires par les méthodes de Jacobi et Gauss-Seidel peuvent se réécrire sous la forme suivante :

Fonction $\boldsymbol{X} \leftarrow$ RSLJacobiV2 $(\mathbb{A}, \boldsymbol{b}, \boldsymbol{x}^0, \varepsilon, \text{kmax})$
  $k \leftarrow 0$, $\boldsymbol{X} \leftarrow \varnothing$
  $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
  tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
  Tantque $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
    $k \leftarrow k + 1$
    $\boldsymbol{p} \leftarrow \boldsymbol{x}$
    $\boldsymbol{x} \leftarrow$ IterJacobi$(\mathbb{A}, \boldsymbol{b}, \boldsymbol{p})$
    $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
  Fin Tantque
  Si $\|\boldsymbol{r}\| \leqslant$ tol **alors**
    $\boldsymbol{X} \leftarrow \boldsymbol{x}$
  Fin Si
Fin Fonction

Fonction $\boldsymbol{X} \leftarrow$ RSLGaussSeidelV2 $(\mathbb{A}, \boldsymbol{b}, \boldsymbol{x}^0, \varepsilon, \text{kmax})$
  $k \leftarrow 0$, $\boldsymbol{X} \leftarrow \varnothing$
  $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
  tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
  Tantque $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
    $k \leftarrow k + 1$
    $\boldsymbol{p} \leftarrow \boldsymbol{x}$
    $\boldsymbol{x} \leftarrow$ IterGaussSeidel$(\mathbb{A}, \boldsymbol{b}, \boldsymbol{p})$
    $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A} * \boldsymbol{x}$,
  Fin Tantque
  Si $\|\boldsymbol{r}\| \leqslant$ tol **alors**
    $\boldsymbol{X} \leftarrow \boldsymbol{x}$
  Fin Si
Fin Fonction

En programmation, dès que l'on commence à faire des copier/coller[1] il faut se poser la question : est-il possible de faire sans?

La plupart du temps la réponse est oui, et celà permet souvent de simplifier, clarifier et racourcir le code ce qui simplifie grandement sa maitenance. Dans notre cas, on écrit l'Algorithme générique 3.19 sous forme d'une fonction à laquelle on ajoute aux paramètres d'entrées une fonction formelle IterFonc permettant le calcul d'une itérée :

$$\boldsymbol{x} \leftarrow \text{IterFonc}(\mathbb{A}, \boldsymbol{b}, \boldsymbol{y}).$$

Cette fonction est donc une donnée du problème. On a alors

Algorithme 3.24 Méthode itérative pour la résolution d'un système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$

Données :
$\mathbb{A}$: matrice de $\mathcal{M}_n(\mathbb{K})$,
$\boldsymbol{b}$: vecteur de $\mathbb{K}^n$,
IterFonc : fonction de paramètres une matrice d'ordre n,
: et deux vecteurs de $\mathbb{K}^n$. retourne un vecteur de $\mathbb{K}^n$.
$\boldsymbol{x}^0$: vecteur initial de $\mathbb{K}^n$,
ε : la tolérence, $\varepsilon \in \mathbb{R}^+$,
kmax : nombre maximum d'itérations, kmax $\in \mathbb{N}^*$
Résultat :
$\boldsymbol{x}^{\text{tol}}$: un vecteur de $\mathbb{K}^n$ si convergence, sinon $\varnothing$

1: **Fonction** $\boldsymbol{X} \leftarrow$ RSLMethIter $(\mathbb{A}, \boldsymbol{b}, \text{IterFonc}, \boldsymbol{x}^0, \varepsilon, \text{kmax})$
2: $k \leftarrow 0$, $\boldsymbol{x}^{\text{tol}} \leftarrow \varnothing$
3: $\boldsymbol{x} \leftarrow \boldsymbol{x}^0$, $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
4: tol $\leftarrow \varepsilon(\|\boldsymbol{b}\| + 1)$
5: **Tantque** $\|\boldsymbol{r}\| >$ tol et $k \leqslant$ kmax **faire**
6: $k \leftarrow k + 1$
7: $\boldsymbol{p} \leftarrow \boldsymbol{x}$
8: $\boldsymbol{x} \leftarrow \text{IterFonc}(\mathbb{A}, \boldsymbol{b}, \boldsymbol{p})$
9: $\boldsymbol{r} \leftarrow \boldsymbol{b} - \mathbb{A}\boldsymbol{x}$,
10: **Fin Tantque**
11: **Si** $\|\boldsymbol{r}\| \leqslant$ tol **alors**
12: $\boldsymbol{x}^{\text{tol}} \leftarrow \boldsymbol{x}$
13: **Fin Si**
14: **Fin Fonction**

En utilisant cette fonction, les algorithmes de résolutions de systèmes linéaires par les méthodes de Jacobi et Gauss-Seidel peuvent se réécrire sous la forme suivante :

Fonction $\boldsymbol{X} \leftarrow$ RSLJacobiV3 $(\mathbb{A}, \boldsymbol{b}, \boldsymbol{x}^0, \varepsilon, \text{kmax})$
$\boldsymbol{X} \leftarrow$ RSLMethIter$(\mathbb{A}, \boldsymbol{b}, \text{IterJacobi}, \boldsymbol{x}^0, \varepsilon, \text{kmax})$
Fin Fonction

Fonction $\boldsymbol{X} \leftarrow$ RSLGaussSeidelV3 $(\mathbb{A}, \boldsymbol{b}, \boldsymbol{x}^0, \varepsilon, \text{kmax})$
$\boldsymbol{X} \leftarrow$ RSLMethIter$(\mathbb{A}, \boldsymbol{b}, \text{IterGaussSeidel}, \boldsymbol{x}^0, \varepsilon, \text{kmax})$
Fin Fonction

Méthode S.O.R.

Pour la méthode de relaxation utilisant Gauss-Seidel,avec $w \in \mathbb{R}^*$, la suite des itérées est définie par

$$x_i^{[k+1]} = \frac{w}{\mathrm{A}_{ii}} \left(b_i - \sum_{j=1}^{i-1} \mathrm{A}_{ij} x_j^{[k+1]} - \sum_{j=i+1}^{n} \mathrm{A}_{ij} x_j^{[k]} \right) + (1-w) x_i^{[k]} \quad \forall i \in [\![1, n]\!]$$

[1] Opération qui pour un bon programmeur est une chose très fatiguante!

Algorithme 3.25 Itération S.O.R. : calcul de $\boldsymbol{x}$ tel que $x_i = \frac{w}{\mathrm{A}_{ii}}\left(b_i - \sum_{j=1}^{i-1} \mathrm{A}_{ij}x_j - \sum_{j=i+1}^{n} \mathrm{A}_{ij}y_j\right) + (1-w)y_i$

Données :
A : matrice de $\mathcal{M}_n(\mathbb{K})$,
b : vecteur de $\mathbb{K}^n$,
y : vecteur de $\mathbb{K}^n$,
w : réel non nul.
Résultat :
x : un vecteur de $\mathbb{K}^n$

```
1: Fonction x ← IterSOR ( A, b, y, w )
2:   Pour i ← 1 à n faire
3:     S ← 0
4:     Pour j ← 1 à i − 1 faire
5:       S ← S + A(i, j) ∗ x(j)
6:     Fin Pour
7:     Pour j ← i + 1 à n faire
8:       S ← S + A(i, j) ∗ y(j)
9:     Fin Pour
10:    x(i) ← w ∗ (b(i) − S)/A(i, i) + (1 − w) ∗ y(i)
11:  Fin Pour
12: Fin Fonction
```

```
Fonction X ← RSLSORV3 ( 𝔸, b, w, x⁰, ε, kmax )
  IterFun ← ((𝕄, r, s) ↦ IterSOR(𝕄, r, s, w))
  X ← RSLMethIter(𝔸, b, IterFun, x⁰, ε, kmax)
Fin Fonction
```

Une ch'tite remarque

Les méthodes itératives que l'on a étudiées sont basées sur l'expression

$$\boldsymbol{x}^{[k+1]} = \mathbb{B}\boldsymbol{x}^{[k]} + \boldsymbol{c}$$

Si l'on pose

$$\boldsymbol{\Phi} : \boldsymbol{x} \mapsto \mathbb{B}\boldsymbol{x} + \boldsymbol{c}$$

alors

$$\boldsymbol{x}^{[k+1]} = \boldsymbol{\Phi}(\boldsymbol{x}^{[k]}).$$

Celà vous rappelle-t'il quelque chose?

3.4.5 Exercices

Exercice 3.4.3

Q. 1 *Montrer que pour la matrice*

$$\mathbb{A}_1 = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}$$

la méthode de Jacobi converge, tandis que la méthode de Gauss-Seidel diverge.

Q. 2 *Montrer que pour la matrice*

$$\mathbb{A}_2 = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$$

la méthode de Jacobi diverge, tandis que la méthode de Gauss-Seidel converge.

Exercice 3.4.4

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice hermitienne définie positive décomposée (par points) sous la forme $\mathbb{A} = \mathbb{D} - \mathbb{E} - \mathbb{F}$ où $\mathbb{D} = \mathrm{diag}(\mathbb{A})$, $\mathbb{E}$ est triangulaire inférieure et d'éléments nuls sur la diagonale et $\mathbb{F}$ est triangulaire supérieure et d'éléments nuls sur la diagonale.

On étudie une méthode itérative de résolution su système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$. Soit $\boldsymbol{x}_0 \in \mathbb{C}^n$, on

définit la suite $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ par

$$(\mathbb{D}-\mathbb{E})\boldsymbol{x}_{k+1/2} = \mathbb{F}\boldsymbol{x}_k + \boldsymbol{b} \quad (3.79)$$
$$(\mathbb{D}-\mathbb{F})\boldsymbol{x}_{k+1} = \mathbb{E}\boldsymbol{x}_{k+1/2} + \boldsymbol{b} \quad (3.80)$$

Q. 1 *Ecrire le vecteur $\boldsymbol{x}_{k+1}$ sous la forme*

$$\boldsymbol{x}_{k+1} = \mathbb{B}\boldsymbol{x}_k + \boldsymbol{c} \quad (3.81)$$

en explicitant le matrice $\mathbb{B}$ et le vecteur $\boldsymbol{c}$.

Q. 2 *1. Montrer que*

$$\mathbb{D}^{-1} = (\mathbb{D}-\mathbb{E})^{-1} - \mathbb{D}^{-1}\mathbb{E}(\mathbb{D}-\mathbb{E})^{-1}. \quad (3.82)$$

2. Soit $(\lambda,\boldsymbol{p})$ un élément propre de la matrice $\mathbb{B}$. Montrer que

$$\lambda\mathbb{A}\boldsymbol{p} + (\lambda-1)\mathbb{E}\mathbb{D}^{-1}\mathbb{F}\boldsymbol{p} = 0. \quad (3.83)$$

Q. 3 *En déduire la convergence de cette méthode vers la solution $\underline{\boldsymbol{x}}$ de $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$.*

Q. 4 *Etendre ces résultats au cas d'une décomposition $\mathbb{A} = \mathbb{D} - \mathbb{E} - \mathbb{F}$ par blocs.*

Correction Exercice

Q. 1 La matrice $\mathbb{D}$ est inversible. En effet, pour tout $i \in [\![1,n]\!]$, $d_{i,i} = a_{i,i} = \langle \mathbb{A}\boldsymbol{e}_i, \boldsymbol{e}_i\rangle > 0$ car $\mathbb{A}$ définie positive et $\boldsymbol{e}_i$, i-ème vecteur de la base canonique est non nul.
On en déduit que les matrices $\mathbb{D}-\mathbb{E}$ (triangulaire inférieure de diagonale la diagonale de $\mathbb{D}$) et $\mathbb{D}-\mathbb{F}$ (triangulaire supérieure de diagonale la diagonale de $\mathbb{D}$) sont inversibles.

De (3.79), on obtient en multipliant à gauche par $(\mathbb{D}-\mathbb{E})^{-1}$

$$\boldsymbol{x}_{k+1/2} = (\mathbb{D}-\mathbb{E})^{-1}\mathbb{F}\boldsymbol{x}_k + (\mathbb{D}-\mathbb{E})^{-1}\boldsymbol{b}.$$

En remplaçant cette expression de $\boldsymbol{x}_{k+1/2}$ dans (3.80), on a

$$\begin{aligned}(\mathbb{D}-\mathbb{F})\boldsymbol{x}_{k+1} &= \mathbb{E}\left((\mathbb{D}-\mathbb{E})^{-1}\mathbb{F}\boldsymbol{x}_k + (\mathbb{D}-\mathbb{E})^{-1}\boldsymbol{b}\right) + \boldsymbol{b}\\ &= \mathbb{E}(\mathbb{D}-\mathbb{E})^{-1}\mathbb{F}\boldsymbol{x}_k + \left(\mathbb{E}(\mathbb{D}-\mathbb{E})^{-1} + \mathbb{I}\right)\boldsymbol{b}\end{aligned}$$

En multipliant à gauche cette équation par $(\mathbb{D}-\mathbb{F})^{-1}$, on abouti a

$$\boldsymbol{x}_{k+1} = (\mathbb{D}-\mathbb{F})^{-1}\mathbb{E}(\mathbb{D}-\mathbb{E})^{-1}\mathbb{F}\boldsymbol{x}_k(\mathbb{D}-\mathbb{F})^{-1}\left(\mathbb{E}(\mathbb{D}-\mathbb{E})^{-1} + \mathbb{I}\right)\boldsymbol{b}$$

En posant

$$\mathbb{B} = (\mathbb{D}-\mathbb{F})^{-1}\mathbb{E}(\mathbb{D}-\mathbb{E})^{-1}\mathbb{F}$$
$$\boldsymbol{c} = (\mathbb{D}-\mathbb{F})^{-1}\left(\mathbb{E}(\mathbb{D}-\mathbb{E})^{-1} + \mathbb{I}\right)\boldsymbol{b}$$

on obtient (3.81).

Q. 2 1. L'expression à démontrer est bien définie car $\mathbb{D}$ et $\mathbb{D}-\mathbb{E}$ inversibles. De plus on a

$$\mathbb{I} = (\mathbb{D}-\mathbb{E})(\mathbb{D}-\mathbb{E})^{-1}$$

En multipliant à gauche cette équation par $\mathbb{D}^{-1}$ on obtient

$$\begin{aligned}\mathbb{D}^{-1} &= \mathbb{D}^{-1}(\mathbb{D}-\mathbb{E})(\mathbb{D}-\mathbb{E})^{-1}\\ &= (\mathbb{D}-\mathbb{E})^{-1} - \mathbb{D}^{-1}\mathbb{E}(\mathbb{D}-\mathbb{E})^{-1}.\end{aligned}$$

2. Soit $(\lambda,\boldsymbol{p})$ un élément propre de la matrice $\mathbb{B}$. on a alors

$$\begin{aligned}\mathbb{B}\boldsymbol{p} = \lambda\boldsymbol{p} &\Leftrightarrow (\mathbb{D}-\mathbb{F})^{-1}\mathbb{E}(\mathbb{D}-\mathbb{E})^{-1}\mathbb{F}\boldsymbol{p} = \lambda\boldsymbol{p}\\ &\Leftrightarrow \mathbb{E}(\mathbb{D}-\mathbb{E})^{-1}\mathbb{F}\boldsymbol{p} = \lambda(\mathbb{D}-\mathbb{F})\boldsymbol{p}\\ &\Leftrightarrow \mathbb{D}^{-1}\mathbb{E}(\mathbb{D}-\mathbb{E})^{-1}\mathbb{F}\boldsymbol{p} = \lambda\mathbb{D}^{-1}(\mathbb{D}-\mathbb{F})\boldsymbol{p}\end{aligned}$$

De (3.82), on a

$$\mathbb{D}^{-1}\mathbb{E}(\mathbb{D}-\mathbb{E})^{-1} = (\mathbb{D}-\mathbb{E})^{-1} - \mathbb{D}^{-1}$$

et donc

$$\begin{aligned}
\mathbb{B}\boldsymbol{p} = \lambda\boldsymbol{p} &\Leftrightarrow \left((\mathbb{D}-\mathbb{E})^{-1} - \mathbb{D}^{-1}\right)\mathbb{F}\boldsymbol{p} = \lambda\mathbb{D}^{-1}(\mathbb{D}-\mathbb{F})\boldsymbol{p}\\
&\Leftrightarrow (\mathbb{D}-\mathbb{E})\left((\mathbb{D}-\mathbb{E})^{-1} - \mathbb{D}^{-1}\right)\mathbb{F}\boldsymbol{p} = \lambda(\mathbb{D}-\mathbb{E})\mathbb{D}^{-1}(\mathbb{D}-\mathbb{F})\boldsymbol{p}\\
&\Leftrightarrow \left(\mathbb{I} - \mathbb{I} + \mathbb{E}\mathbb{D}^{-1}\right)\mathbb{F}\boldsymbol{p} = \lambda(\mathbb{I} - \mathbb{E}\mathbb{D}^{-1})(\mathbb{D}-\mathbb{F})\boldsymbol{p}\\
&\Leftrightarrow \mathbb{E}\mathbb{D}^{-1}\mathbb{F}\boldsymbol{p} = \lambda(\mathbb{D} - \mathbb{E} - \mathbb{F} + \mathbb{E}\mathbb{D}^{-1}\mathbb{F})\boldsymbol{p}\\
&\Leftrightarrow \mathbb{E}\mathbb{D}^{-1}\mathbb{F}\boldsymbol{p} = \lambda(\mathbb{A} + \mathbb{E}\mathbb{D}^{-1}\mathbb{F})\boldsymbol{p}
\end{aligned}$$

On en déduit alors

$$\lambda\mathbb{A}\boldsymbol{p} + (\lambda-1)\mathbb{E}\mathbb{D}^{-1}\mathbb{F}\boldsymbol{p} = 0.$$

Q. 3 La matrice $\mathbb{A}$ est inversible car elle est définie positive et donc $\underline{\boldsymbol{x}}$ est bien définie. De l'équation (3.79), on déduit

$$(\mathbb{D}-\mathbb{E})\boldsymbol{x}_{k+1/2} = \mathbb{F}\boldsymbol{x}_k + \mathbb{A}\underline{\boldsymbol{x}} = \mathbb{F}\boldsymbol{x}_k + (\mathbb{D}-\mathbb{E}-\mathbb{F})\underline{\boldsymbol{x}}$$

et donc

$$(\mathbb{D}-\mathbb{E})(\boldsymbol{x}_{k+1/2} - \underline{\boldsymbol{x}}) = \mathbb{F}(\boldsymbol{x}_k - \underline{\boldsymbol{x}}) \tag{3.84}$$

De la même manière à partir de l'équation (3.80), on déduit

$$(\mathbb{D}-\mathbb{F})(\boldsymbol{x}_{k+1} - \underline{\boldsymbol{x}}) = \mathbb{E}(\boldsymbol{x}_{k+1/2} - \underline{\boldsymbol{x}}) \tag{3.85}$$

En utilisant (3.84), l'équation (3.86) devient

$$(\mathbb{D}-\mathbb{F})(\boldsymbol{x}_{k+1} - \underline{\boldsymbol{x}}) = \mathbb{E}(\mathbb{D}-\mathbb{E})^{-1}\mathbb{F}(\boldsymbol{x}_k - \underline{\boldsymbol{x}})$$

c'est à dire

$$\boldsymbol{x}_{k+1} - \underline{\boldsymbol{x}} = \mathbb{B}(\boldsymbol{x}_k - \underline{\boldsymbol{x}}).$$

En posant $\boldsymbol{e}_k = \boldsymbol{x}_k - \underline{\boldsymbol{x}}$ on a alors

$$\boldsymbol{e}_k = \mathbb{B}^k\boldsymbol{e}_0, \quad \forall k \geqslant 0.$$

Or la suite $\boldsymbol{x}_k$ converge vers $\underline{\boldsymbol{x}}$ si et seulement si la suite $\boldsymbol{e}_k$ converge vers $\mathbf{0}$. Pour celà, d'après le Théorème 3.37, page 103, il est nécessaire et suffisant d'avoir $\rho(\mathbb{B}) < 1$.
Soit $(\lambda, \boldsymbol{p})$ un élément propre de $\mathbb{B}$. Montrons que $|\lambda| < 1$.
On déduit de l'équation (3.83)

$$\begin{aligned}
0 &= \left\langle \boldsymbol{p}, \lambda\mathbb{A}\boldsymbol{p} + (\lambda-1)\mathbb{E}\mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}\right\rangle\\
&= \lambda\left\langle \boldsymbol{p}, \mathbb{A}\boldsymbol{p}\right\rangle + (\lambda-1)\left\langle \boldsymbol{p}, \mathbb{E}\mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}\right\rangle
\end{aligned} \tag{3.86}$$

Comme la matrice $\mathbb{A}$ est définie positive on a $\langle\mathbb{A}\boldsymbol{p}, \boldsymbol{p}\rangle > 0$ car $\boldsymbol{p} \neq \mathbf{0}$ (vecteur propre) et donc

$$\left\langle \boldsymbol{p}, \mathbb{A}\boldsymbol{p}\right\rangle = \overline{\left\langle \mathbb{A}\boldsymbol{p}, \boldsymbol{p}\right\rangle} = \left\langle \mathbb{A}\boldsymbol{p}, \boldsymbol{p}\right\rangle > 0.$$

De plus on a

$$\left\langle \boldsymbol{p}, \mathbb{E}\mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}\right\rangle = \left\langle \mathbb{E}^*\boldsymbol{p}, \mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}\right\rangle.$$

La matrice $\mathbb{A}$ étant hermitienne, on a $\mathbb{E}^* = \mathbb{F}$. La matrice $\mathbb{A}$ étant définie positive, la matrice diagonale $\mathbb{D}$ est définie positive car $d_{i,i} > 0$, $\forall i \in [\![1, n]\!]$, et donc $\mathbb{D}^{-1}$ aussi. Comme $\mathbb{F}\boldsymbol{p}$ n'est pas nécessairement non nul, on a

$$\left\langle \mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}, \mathbb{F}\boldsymbol{p}\right\rangle \in \mathbb{R}^+.$$

On en déduit

$$\left\langle \mathbb{F}\boldsymbol{p}, \mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}\right\rangle = \overline{\left\langle \mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}, \mathbb{F}\boldsymbol{p}\right\rangle} = \left\langle \mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}, \mathbb{F}\boldsymbol{p}\right\rangle \geqslant 0.$$

De l'équation (3.86), on obtient

$$\lambda(\left\langle \boldsymbol{p}, \mathbb{A}\boldsymbol{p}\right\rangle + \left\langle \mathbb{F}\boldsymbol{p}, \mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}\right\rangle) = \left\langle \mathbb{F}\boldsymbol{p}, \mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}\right\rangle$$

or

$$\left\langle \boldsymbol{p}, \mathbb{A}\boldsymbol{p}\right\rangle + \left\langle \mathbb{F}\boldsymbol{p}, \mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}\right\rangle \neq 0$$

ce qui donne

$$\lambda = \frac{\left\langle \mathbb{F}\boldsymbol{p}, \mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}\right\rangle}{\left\langle \boldsymbol{p}, \mathbb{A}\boldsymbol{p}\right\rangle + \left\langle \mathbb{F}\boldsymbol{p}, \mathbb{D}^{-1}\mathbb{F}\boldsymbol{p}\right\rangle}.$$

On a alors $\lambda \in [0, 1[$.

Q. 4 Comme la matrice $\mathbb{A}$ est hermitienne définie positive, chaque bloc diadonal l'est aussi. Donc la matrice diagonale bloc $\mathbb{D}$ est aussi hermitienne définie positive ainsi que son inverse . Les résultats précédents sont donc toujours valable.

◇

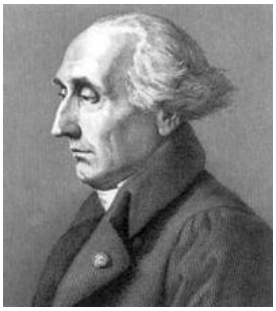
Chapitre 4

Interpolation

L'interpolation est un outil mathématique permettant de construire des fonctions à partir de la donnée d'un nombre fini de valeurs.

A developper...

Les protagonistes de cette histoire



(a) *Joseph-Louis Lagrange* 1736-1813, mathématicien italien puis français

(b) *Pafnouti Lvovitch Tchebychev* 1821-1894, mathématicien russe

(c) *Charles Hermite* 1822-1901, mathématicien français

(d) *Henri-Léon Lebesgue* 1875-1941, mathématicien français

4.1 Polynôme d'interpolation de Lagrange

Exercice 4.1.1

Soient $n \in \mathbb{N}^*$ et $n+1$ couples de $\mathbb{R}^2$, $(x_i, y_i)_{i\in[\![0,n]\!]}$, tels que les x_i sont distincts deux à deux. On note

Q. 1 *1. Soit $i \in [\![0,n]\!]$. Montrer qu'il existe un unique polynôme L_i de degré n vérifiant*

$$L_i(x_j) = \delta_{ij}, \ \forall j \in [\![0,n]\!]. \tag{4.1}$$

2. Montrer que les $(L_i)_{i\in[\![0,n]\!]}$ forment une base de $\mathbb{R}_n[X]$ (espace vectoriel des polynômes à coefficients réels de degré inférieur ou égal à n).

On défini le polynôme P_n par

$$\mathrm{P}_n(x) = \sum_{i=0}^{n} y_i L_i(x). \tag{4.2}$$

Q. 2 *Montrer que polynôme P_n est l'unique polynôme de degré au plus n vérifiant $\mathrm{P}_n(x_i) = y_i$, $\forall i \in [\![1,n]\!]$.*

Correction Exercice

Q. 1 1. De (4.1), on déduit que les n points distincts x_j pour $j \in [\![0,n]\!]\backslash\{i\}$ sont les n zéros du polynôme L_i de degré n : il s'écrit donc sous la forme

$$L_i(x) = C \prod_{\substack{j=0 \\ j\neq i}}^{n} (x - x_j) \quad \text{avec } C \in \mathbb{R}$$

Pour déterminer la constante C, on utilise (4.1) avec $j = i$

$$L_i(x_i) = 1 = C \prod_{\substack{j=0 \\ j\neq i}}^{n} (x_i - x_j)$$

Les points x_i sont distincts deux à deux, on a $\prod_{\substack{j=0 \\ j\neq i}}^{n}(x_i - x_j) \neq 0$ et donc

$$C = \frac{1}{\prod_{\substack{j=0 \\ j\neq i}}^{n}(x_i - x_j)}$$

d'où

$$L_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}, \quad \forall i \in [\![0,n]\!]. \tag{4.3}$$

Il reste à démontrer l'unicité. On suppose qu'il existe L_i et U_i deux polynômes de $\mathbb{R}_n[X]$ vérifiant (4.1). Alors $Q_i = L_i - U_i$ est polynôme de degré n (au plus) admettant $n+1$ zéros distincts, c'est donc le polynôme nul et on a nécessairement $L_i = U_i$.

2. On sait que $\dim \mathbb{R}_n[X] = n+1$. Pour que les $\{L_i\}_{i\in[\![0,n]\!]}$ forment une base de $\mathbb{R}_n[X]$ il suffit de démontrer qu'ils sont linéairement indépendants.
Soit $\lambda_0, \cdots, \lambda_n$ $n+1$ scalaires. Montrons pour celà que

$$\sum_{i=1}^{n} \lambda_i L_i = 0 \implies \lambda_i = 0, \ \forall i \in [\![0,n]\!]$$

Noter que la première égalité est dans l'espace vectoriel $\mathbb{R}_n[X]$ et donc le 0 est pris au sens polynôme nul.
On a

$$\sum_{i=1}^{n} \lambda_i L_i = 0 \iff \sum_{i=1}^{n} \lambda_i L_i(x) = 0, \ \forall x \in \mathbb{R}$$

Soit $k \in [\![0,n]\!]$. En choisissant $x = x_k$, on a par (4.1) $\sum_{i=1}^n \lambda_i L_i(x_k) = \lambda_k$ et donc

$$\sum_{i=1}^n \lambda_i L_i = 0 \implies \sum_{i=1}^n \lambda_i L_i(x_k) = 0,\ \forall k \in [\![0,n]\!] \iff \lambda_k = 0,\ \forall k \in [\![0,n]\!].$$

Les $\{L_i\}_{i\in[\![0,n]\!]}$ sont donc linéairement indépendants.

Q. 2 Par construction $\mathrm{P}_n \in \mathbb{R}_n[X]$ et on a, $\forall j \in [\![0,n]\!]$[1],

$$\begin{aligned}\mathrm{P}_n(x_j) &\overset{\text{def}}{=} \sum_{i=0}^n y_i L_i(x_j)\\ &= \sum_{i=0}^n y_i \delta_{i,j} \text{ par (4.1)}\\ &= y_j.\end{aligned}$$

Pour demontrer l'unicité, on propose ici deux méthodes

- On note P_a et P_b deux polynômes de $\mathbb{R}_n[X]$ vérifiant (4.1). Le polynôme $\mathrm{Q} = \mathrm{P}_a - \mathrm{P}_b$ appartient aussi à $\mathbb{R}_n[X]$ et il vérifie, $\forall i \in [\![0,n]\!]$,

$$\mathrm{Q}(x_i) = \mathrm{P}_a(x_i) - \mathrm{P}_b(x_i) = 0.$$

  Les $n+1$ points x_i étant distincts, ce sont donc $n+1$ racines distinctes du polynôme Q. Or tout polynôme de degré n admet au plus n racines disctinctes[2]. On en déduit que le seul polynôme de degré au plus n admettant $n+1$ racines distinctes est le polynôme nulle et donc $\mathrm{P}_a = \mathrm{P}_b$.

- c'est l'unique polynôme de degré au plus n vérifiant (4.2) car la décomposition dans la base $\{L_i\}_{i\in[\![0,n]\!]}$ est unique.

◇

Definition 4.1

Soient $n \in \mathbb{N}^*$ et $(x_i, y_i)_{i\in[\![0,n]\!]}$ avec $(x_i, y_i) \in \mathbb{R}^2$ et les x_i distincts deux à deux. Le **polynôme d'interpolation de Lagrange** associé aux $n+1$ points $(x_i, y_i)_{i\in[\![0,n]\!]}$, noté P_n, est donné par

$$\mathrm{P}_n(x) = \sum_{i=0}^n y_i \mathrm{L}_i(x),\ \forall x \in \mathbb{R} \tag{4.4}$$

avec

$$\mathrm{L}_i(x) = \prod_{\substack{j=0\\ j\neq i}}^n \frac{x - x_j}{x_i - x_j},\ \forall i \in [\![0,n]\!],\ \forall x \in \mathbb{R}. \tag{4.5}$$

Théorème 4.2

Le **polynôme d'interpolation de Lagrange**, $\mathcal{P}_n$, associé aux $n+1$ points $(x_i, y_i)_{i\in[\![0,n]\!]}$, est l'unique polynôme de degré au plus n, vérifiant

$$\mathcal{P}_n(x_i) = y_i,\ \forall i \in [\![0,n]\!]. \tag{4.6}$$

Remarque 4.3 Il est aussi possible d'obtenir l'existence et l'unicité du polynôme d'interpolation de Lagrange sans passer par sa construction. Pour celà on défini $\Phi : \mathbb{R}_n[X] \longrightarrow \mathbb{R}^{n+1}$ par

$$\forall \mathrm{P} \in \mathbb{R}_n[X], \quad \Phi(\mathrm{P}) = (\mathrm{P}(x_0), \cdots, \mathrm{P}(x_n)).$$

[1] A noter le choix de l'indice j. Que doit-on faire dans ce qui suit si l'on choisi i comme indice?

[2] Le théorème de d'Alembert-Gauss affirme que tout polynôme à coefficients complexes de degré n admet n racines complexes qui ne sont pas nécessairement distinctes

L'existence et l'unicité du polynôme P_n est équivalente à la bijectivité de l'application Φ. Or celle-ci est une application linéaire entre deux espaces de dimension $n+1$. Elle est donc bijective si et seulement si elle injective (ou surjective). Pour vérifier l'injectivité de Φ il est nécessaire et suffisant de vérifier que son noyau est réduit au polynôme nul.
Soit $P \in \ker \Phi$. On a alors $\Phi(P) = \mathbf{0}_{n+1}$ et donc $(x_0, \cdots, x_n)$ sont $n+1$ racines distinctes de P. Or le seul polynôme de $\mathbb{R}_n[X]$ ayant $n+1$ racines distinctes est le polynôme nul et donc $P = 0$.

A titre d'exemple, on représente, En figure 4.2, le polynôme d'interpolation de Lagrange associé à 7 points donnés.

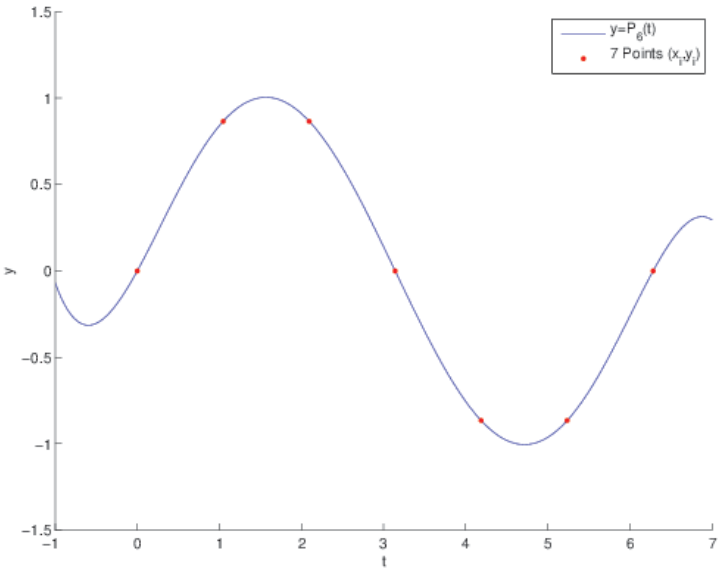


Figure 4.2: Polynôme d'interpolation de Lagrange avec 7 points donnés)

Exercice 4.1.2

Ecrire la fonction LAGRANGE permettant de calculer $\mathcal{P}_n$ (polynôme d'interpolation de Lagrange associé aux $n+1$ points $(x_i, y_i)_{i \in [\![0,n]\!]}$) au point $t \in \mathbb{R}$.

Correction Exercice

But : Calculer le polynôme $\mathcal{P}_n(t)$ définit par (4.4)
Données : $\boldsymbol{X}$: vecteur/tableau de $\mathbb{R}^{n+1}$, $X(i) = x_{i-1}$ $\forall i \in [\![1, n+1]\!]$ et $X(i) \neq X(j)$ pour $i \neq j$,
$\boldsymbol{Y}$: vecteur/tableau de $\mathbb{R}^{n+1}$, $Y(i) = y_{i-1}$ $\forall i \in [\![1, n+1]\!]$,
t : un réel.
Résultat : y : le réel $y = \mathcal{P}_n(t)$.

Algorithme 4.1 $\mathcal{R}_0$

1: Calcul de $y = \mathcal{P}_n(t) = \sum_{i=1}^{n+1} Y(i) L_{i-1}(t)$

Algorithme 4.1 $\mathcal{R}_1$

1: $y \leftarrow 0$
2: **Pour** $i \leftarrow 1$ **à** $n+1$ **faire**
3: $\quad y \leftarrow y + Y(i) * L_{i-1}(t)$
4: **Fin Pour**

Algorithme 4.1 $\mathcal{R}_1$

1: $y \leftarrow 0$
2: **Pour** $i \leftarrow 1$ **à** $n+1$ **faire**
3: $\quad y \leftarrow y + Y(i) * L_{i-1}(t)$
4: **Fin Pour**

Algorithme 4.1 $\mathcal{R}_2$

1: $y \leftarrow 0$
2: **Pour** $i \leftarrow 1$ **à** $n+1$ **faire**
3: $\quad L \leftarrow \prod_{\substack{j=1 \\ j\neq i}}^{n+1} \frac{t - X(j)}{X(i) - X(j)}$
4: $\quad y \leftarrow y + Y(i) * L$
5: **Fin Pour**

Algorithme 4.1 $\mathcal{R}_2$

1: $y \leftarrow 0$
2: **Pour** $i \leftarrow 1$ **à** $n+1$ **faire**
3: $\quad L \leftarrow \prod_{\substack{j=1 \\ j\neq i}}^{n+1} \frac{t - X(j)}{X(i) - X(j)}$
4: $\quad y \leftarrow y + Y(i) * L$
5: **Fin Pour**

Algorithme 4.1 $\mathcal{R}_3$

1: $y \leftarrow 0$
2: **Pour** $i \leftarrow 1$ **à** $n+1$ **faire**
3: $\quad L \leftarrow 1$
4: $\quad$ **Pour** $j \leftarrow 1$ **à** $n+1$, $(j \sim= i)$ **faire**
5: $\quad\quad L \leftarrow L * (t - X(j))/(X(i) - X(j))$
6: $\quad$ **Fin Pour**
7: $\quad y \leftarrow y + Y(i) * L$
8: **Fin Pour**

On obtient alors l'algorithme final

Algorithme 4.1 Fonction LAGRANGE permettant de calculer le polynôme d'interpolation de Lagrange $\mathcal{P}_n(x)$ définit par (4.4)

Données : $\boldsymbol{X}$: vecteur/tableau de $\mathbb{R}^{n+1}$, $X(i) = x_{i-1}$ $\forall i \in [\![1, n+1]\!]$ et $X(i) \neq X(j)$ pour $i \neq j$,
$\boldsymbol{Y}$: vecteur/tableau de $\mathbb{R}^{n+1}$, $Y(i) = y_{i-1}$ $\forall i \in [\![1, n+1]\!]$,
t : un réel.
Résultat : y : le réel $y = \mathcal{P}_n(t)$.

1: **Fonction** $y \leftarrow$ LAGRANGE $(\ t, X, Y\ )$
2: $\quad y \leftarrow 0$
3: $\quad$ **Pour** $i \leftarrow 1$ **à** $n+1$ **faire**
4: $\quad\quad L \leftarrow 1$
5: $\quad\quad$ **Pour** $j \leftarrow 1$ **à** $n+1$, $(j \sim= i)$ **faire**
6: $\quad\quad\quad L \leftarrow L * (t - X(j))/(X(i) - X(j))$
7: $\quad\quad$ **Fin Pour**
8: $\quad\quad y \leftarrow y + Y(i) * L$
9: $\quad$ **Fin Pour**
10: $\quad$ **return** y
11: **Fin Fonction**

◇

4.1.1 Erreur de l'interpolation

Soit une fonction $f : [a, b] \longrightarrow \mathbb{R}$. On suppose que les y_i sont donnés par

$$y_i = f(x_i), \quad \forall i \in [\![0, n]\!]. \tag{4.7}$$

On cherche à évaluer l'erreur $E_n(t) = f(t) - \mathcal{P}_n(t)$, $\forall t \in [a, b]$.

On propose en figures 4.3 à 4.5 d'etudier deux exemples. Le premier (figure de gauche) correspond à l'interpolation de la fonction $\sin(t)$ par le polynôme d'interpolation de Lagrange aux points équidistants $t_i = a + ih$ avec $a = 0$ et $h = 2\pi/n$. Le second (figure de droite) correspond à l'interpolation de la fonction $\frac{1}{1+t^2}$ par le polynôme d'interpolation de Lagrange aux points $x_i = a + ih$ avec $a = -5$ et $h = 10/n$.

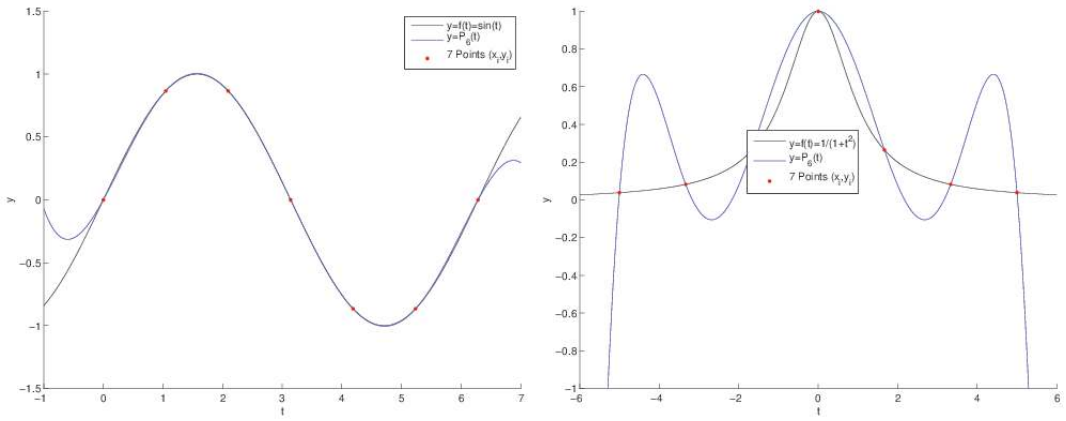


Figure 4.3: Polynômes d'interpolation de lagrange avec $n = 6$ (7 points) uniformément répartis. A gauche pour la fonction $f : t \longrightarrow sin(t)$ avec $x_0 = 0$, $x_6 = 2\pi$ et à droite pour la fonction $f : t \longrightarrow 1/(1 + t^2)$ avec $x_0 = -5$, $x_6 = 5$.

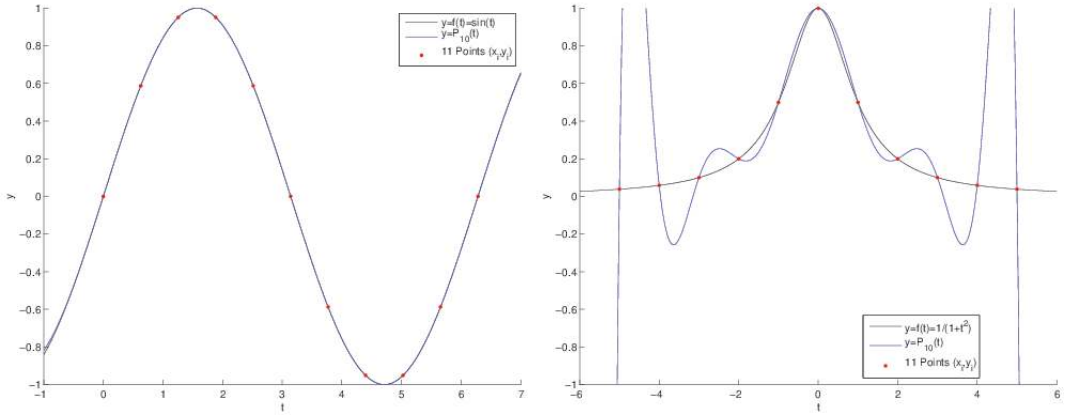


Figure 4.4: Polynômes d'interpolation de lagrange avec $n = 10$ (11 points) uniformément répartis. A gauche pour la fonction $f : t \longrightarrow sin(t)$ avec $x_0 = 0$, $x_{10} = 2\pi$ et à droite pour la fonction $f : tx \longrightarrow 1/(1 + t^2)$ avec $x_0 = -5$, $x_{10} = 5$.

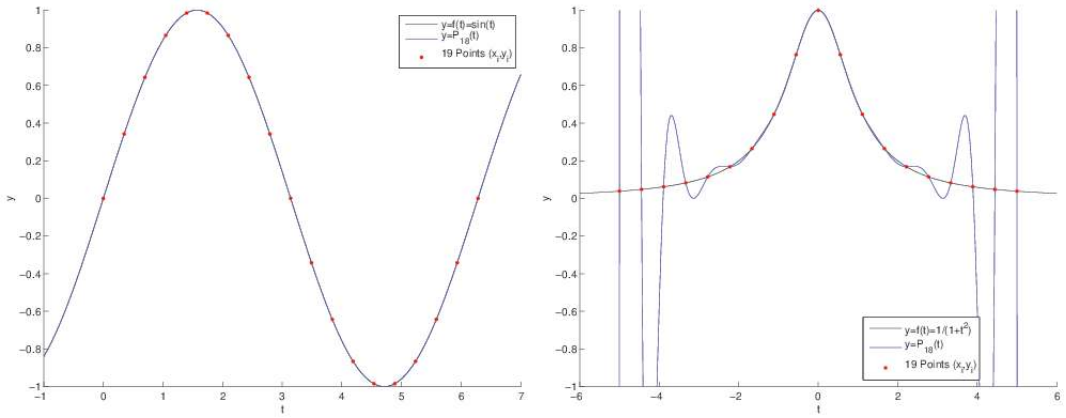


Figure 4.5: Polynômes d'interpolation de lagrange avec $n = 18$ (19 points) uniformément répartis. A gauche pour la fonction $f : t \longrightarrow sin(t)$ avec $x_0 = 0$, $x_{18} = 2\pi$ et à droite pour la fonction $f : t \longrightarrow 1/(1 + t^2)$ avec $x_0 = -5$, $x_{18} = 5$.

Exercice 4.1.3

Soit $f \in \mathcal{C}^{n+1}([a;b];\mathbb{R})$. Soient $n \in \mathbb{N}^*$ et $n+1$ couples de $\mathbb{R}^2$, $(x_i, y_i)_{i\in[\![0,n]\!]}$, tels que les x_i sont distincts deux à deux et $y_i = f(x_i)$.
On note par P_n le polynôme d'interpolation de Lagrange associé aux points $(x_i, y_i)_{i\in[\![0,n]\!]}$ et π_n le polynôme de degré $n+1$ défini par

$$\pi_n(x) = \prod_{i=0}^{n}(x - x_i). \tag{4.8}$$

Q. 1 *Montrer que, $\forall x \in [a;b]$, il existe ξ_x appartenant au plus petit intervalle fermé contenant $x, x_0, \ldots, x_n$ tel que*

$$f(x) - \mathrm{P}_n(x) = \frac{\pi_n(x)}{(n+1)!} f^{(n+1)}(\xi_x). \tag{4.9}$$

Indication : *Etudier les zéros de la fonction* $F(t) = f(t) - \mathrm{P}_n(t) - \dfrac{f(x) - \mathrm{P}_n(x)}{\pi_n(x)}\pi_n(t)$.

Correction Exercice

Q. 1 S'il existe $i \in [\![0,n]\!]$ tel que $x = x_i$ alors l'équation (4.9) est immédiatement vérifiée.
Soit $x \in [a,b]$ distinct de tous les x_i. Comme $f \in \mathcal{C}^{n+1}([a;b];\mathbb{R})$, $\mathrm{P}_n \in \mathbb{R}_n[X]$ et $\pi_n \in \mathbb{R}_{n+1}[X]$, on en déduit que la fonction F est dans $\mathcal{C}^{n+1}([a;b];\mathbb{R})$. La fonction F admet aussi $n+2$ zéros : $x, x_0, \cdots, x_n$. On note $\xi_{x,1}^{[0]}, \cdots, \xi_{x,n+2}^{[0]}$ ces $n+2$ zéros ordonnés $\xi_{x,1}^{[0]} < \cdots < \xi_{x,n+2}^{[0]}$. La fonction F étant continue sur $[a,b]$ et dérivable sur $]a,b[$, le théorème de Rolle dit qu'entre deux zéros consécutifs de F, il existe au moins un zéro de $F' = F^{(1)}$. Plus précisemment on a

$$\forall i \in [\![1, n+1]\!], \quad \exists \xi_{x,i}^{[1]} \in]\xi_{x,i}^{[0]}, \xi_{x,i+1}^{[0]}[\ \text{ tels que } \ F^{(1)}(\xi_{x,i}^{[1]}) = 0$$

et on en déduit que la fonction $F^{(1)}$ admet $n+1$ zéros $\xi_{x,1}^{[1]}, \cdots, \xi_{x,n+1}^{[1]}$ et l'on a $\xi_{x,1}^{[0]} < \xi_{x,1}^{[1]} < \cdots < \xi_{x,n+1}^{[1]} < \xi_{x,n+2}^{[0]}$. Il faut noter la dépendance en x des zéros de F' d'où la notation un peu "lourde".
Montrons par récurrence finie que $(\mathcal{P}_k)$ est vraie pour tout $k \in [\![1, n+1]\!]$

$$(\mathcal{P}_k) \ : \ \exists \xi_{x,i}^{[k]}, \ i \in [\![1, n+2-k]\!], \ \xi_{x,1}^{[0]} < \xi_{x,1}^{[k]} < \cdots < \xi_{x,n+2-k}^{[k]} < \xi_{x,n+2}^{[0]} \ \text{ tels que } \ F^{(k)}(\xi_{x,i}^{[k]}) = 0$$

Initialisation : Pour $k = 1$, la preuve a déjà été faites.

Hérédité : Soit $1 < k-1 < n+1$, on suppose $(\mathcal{P}_{k-1})$ vérifiée. La fonction $F^{(k-1)}$ étant continue sur $[a,b]$ et dérivable sur $]a,b[$, le théorème de Rolle dit qu'entre deux zéros consécutifs de $F^{(k-1)}$, il existe au moins un zéro de $F^{(k)}$. Par hypothèse $F^{(k-1)}$ admet $n+2-(k-1)$ zéros vérifiant

$$\xi_{x,1}^{[0]} < \xi_{x,1}^{[k-1]} < \cdots < \xi_{x,n+2-(k-1)}^{[k-1]} < \xi_{x,n+2}^{[0]}$$

La fonction $F^{(k-1)}$ est continue sur $[a,b]$ et dérivable sur $]a,b[$ puisque $F \in \mathcal{C}^{n+1}([a;b];\mathbb{R})$. Par application du théorème de Rolle, entre deux zéros de $F^{(k-1)}$, il existe au moins un zéro de $F^{(k)}$. Plus précisemment pour tout $i \in [\![1, n+2-k]\!]$ on a

$$\exists \xi_{x,i}^{[k]} \in]\xi_{x,i}^{[k-1]}, \xi_{x,i+1}^{[k-1]}[, \quad F^{(k)}(\xi_{x,i}^{[k]}) = 0$$

De plus, par construction, $\xi_{x,1}^{[0]} < \xi_{x,1}^{[k]} < \cdots < \xi_{x,n+2-k}^{[k]} < \xi_{x,n+2}^{[0]}$. et donc $(\mathcal{P}_k)$ est vraie.

Avec $k = n+1$ on obtient

$$(\mathcal{P}_{n+1}) \ : \ \exists \xi_{x,1}^{[n+1]} \in]\xi_{x,1}^{[0]}, \xi_{x,n+2}^{[0]}[\ \text{ tel que } \ F^{(n+1)}(\xi_{x,1}^{[n+1]}) = 0$$

et donc

$$0 = F^{(n+1)}(\xi_{x,1}^{[n+1]}) = f^{(n+1)}(\xi_{x,1}^{[n+1]}) - \mathrm{P}_n^{(n+1)}(\xi_{x,1}^{[n+1]}) - \frac{f(x) - \mathrm{P}_n(x)}{\pi_n(x)}\pi_n^{(n+1)}(\xi_{x,1}^{[n+1]})$$

Comme $\mathrm{P}_n \in \mathbb{R}_n[X]$, on a $\mathrm{P}_n^{(n+1)} = 0$. De plus $\pi_n \in \mathbb{R}_{n+1}[X]$, et comme $\pi_n(x) = x^{n+1} + Q(x)$ avec $Q \in \mathbb{R}_n[X]$ (i.e. son monône de puissance $n+1$ à pour coefficient 1) on obtient $\pi_n^{(n+1)}(x) = (n+1)!$ On a alors

$$f^{(n+1)}(\xi_{x,1}^{[n+1]}) = \frac{f(x) - \mathrm{P}_n(x)}{\pi_n(x)}(n+1)!$$

◇

Le résultat suivant est du à Cauchy (1840)

Théorème 4.4

Soient $n \in \mathbb{N}^*$ et $x_0, \cdots, x_n$ $n+1$ points distincts de l'intervalle $[a,b]$. Soient $f \in \mathcal{C}^{n+1}([a;b];\mathbb{R})$ et $\mathcal{P}_n$ le polynôme d'interpolation de Lagrange de degré n passant par $(x_i, f(x_i))$, $\forall i \in [\![0,n]\!]$. Alors, $\forall x \in [a,b]$, $\exists \xi_x \in (min(x_i,x), max(x_i,x))$,

$$f(x) - \mathcal{P}_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{i=0}^{n}(x - x_i) \tag{4.10}$$

4.1.2 Points de Chebyshev

Pour minimiser l'erreur commise lors de l'interpolation d'une fonction f par un polynôme d'interpolation de Lagrange, on peut, pour un n donné, "jouer" sur le choix des points x_i :
Trouver $(\bar{x}_i)_{i=0}^n$, $\bar{x}_i \in [a,b]$, distincts deux à deux, tels que

$$\max_{t\in[a,b]} \prod_{i=0}^{n} |t - \bar{x}_i| \leqslant \max_{t\in[a,b]} \prod_{i=0}^{n} |t - x_i|, \ \forall (x_i)_{i=0}^n, \ x_i \in [a,b], \ \text{distincts 2 à 2} \tag{4.11}$$

On a alors le résultat suivant

Théorème 4.5

Les points réalisant (4.11) sont les points de Chebyshev donnés par

$$\bar{x}_i = \frac{a+b}{2} + \frac{b-a}{2}\cos\left(\frac{(2i+1)\pi}{2n+2}\right), \ \forall i \in [\![0,n]\!]. \tag{4.12}$$

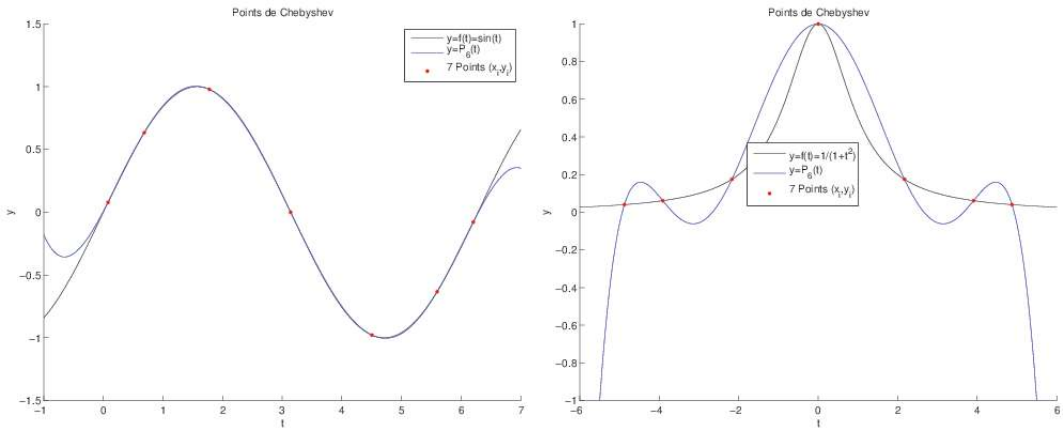


Figure 4.6: Erreurs d'interpolation avec $n = 6$

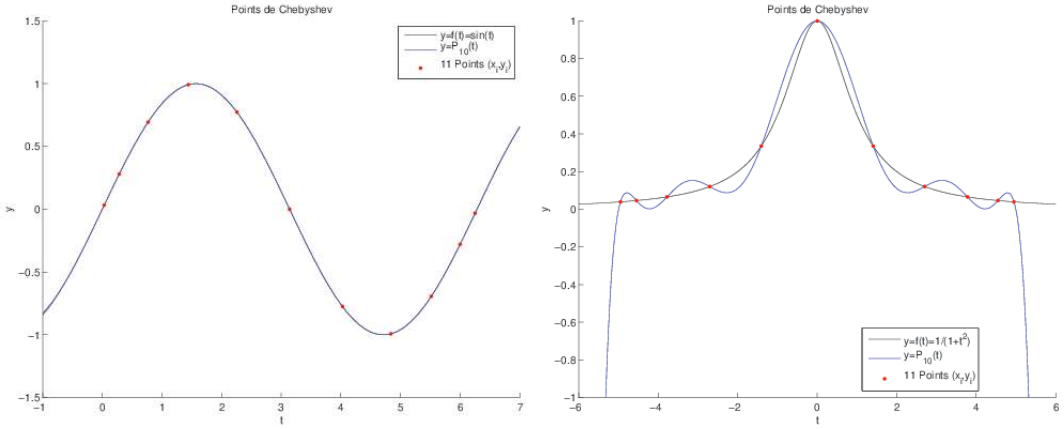


Figure 4.7: Erreurs d'interpolation avec $n = 10$

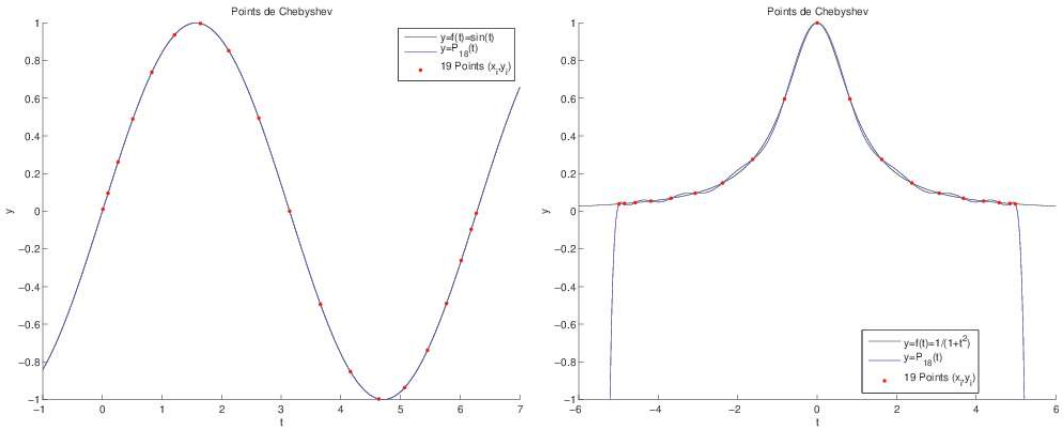


Figure 4.8: Erreurs d'interpolation avec $n = 18$

4.1.3 Stabilité

Inspiré du polycopié de G. Barles : ici

On suppose que l'on commet des erreurs lors du calcul des $f(x_i)$ et l'on note $f_i \approx f(x_i)$ les valeurs numériques obtenues. Dans ce cadre, on a deux polynômes d'interpolation de Lagrange l'un "exact" avec les couples de points $(x_i, y_i) = (x_i, f(x_i))$ noté P_n, et l'autre "approché" avec les couples de points (x_i, f_i) noté $\widehat{\mathrm{P}}_n$. On a donc

$$\mathrm{P}_n(x) = \sum_{i=0}^{n} f(x_i)\mathrm{L}_i(x) \quad \text{et} \quad \widehat{\mathrm{P}}_n(x) = \sum_{i=0}^{n} f_i\mathrm{L}_i(x)$$

L'erreur commise est alors majorée par :

$$\begin{aligned} |\widehat{\mathrm{P}}_n(x) - \mathrm{P}_n(x)| &= |\sum_{i=0}^{n}(f_i - f(x_i))\mathrm{L}_i(x)| \\ &\leqslant \sum_{i=0}^{n} |f_i - f(x_i)||\mathrm{L}_i(x)| \\ &\leqslant \max_{i\in[\![0,n]\!]} |f_i - f(x_i)| \sum_{i=0}^{n} |\mathrm{L}_i(x)|. \end{aligned}$$

On note $\Lambda_n = \max\limits_{x\in[a,b]} \sum\limits_{i=0}^{n} |\mathrm{L}_i(x)|$, dites *Constante de Lebesgue*. Cette constante est bien définie car l'application $x \mapsto \sum_{i=0}^{n} |\mathrm{L}_i(x)|$ est continue sur $[a,b]$ intervalle fermé borné de $\mathbb{R}$ et donc le maximum est bien atteint : il est donc fini.

On obtient alors

$$\left\|\widehat{\mathrm{P}}_n - \mathrm{P}_n\right\|_\infty \leqslant \Lambda_n \max_{i\in[\![0,n]\!]} |f_i - f(x_i)|. \tag{4.13}$$

Proposition 4.6

Soient $n \in N^*$ et $x_0, \cdots, x_n$ des points distincts de $[a,b]$. L'application $\mathcal{L}_n : \mathcal{C}^0([a,b];\mathbb{R}) \longrightarrow \mathbb{R}_n[X]$ qui a toute fonction $f \in \mathcal{C}^0([a,b];\mathbb{R})$ donne le polynôme d'interpolation de Lagrange P_n associés aux couples de $(x_i, f(x_i))_{i\in[\![0,n]\!]}$ est bien définie et linéaire. De plus on a

$$\|\mathcal{L}_n\| \overset{\text{def}}{=} \sup_{\substack{f\in\mathcal{C}^0([a,b];\mathbb{R}) \\ f\neq 0}} \frac{\|\mathcal{L}_n(f)\|_\infty}{\|f\|_\infty} = \Lambda_n. \tag{4.14}$$

Preuve. Bien définie et linéaire : facile mais à écrire

On montre tout d'abord que $\|\mathcal{L}_n\| \leqslant \Lambda_n$. En effet, on a pour tout $x \in [a,b]$ et pour tout $f \in \mathcal{C}^0([a,b];\mathbb{R})$,

$$\begin{aligned} |\mathcal{L}_n(f)(x)| &= |\sum_{i=0}^{n} f(x_i)\mathrm{L}_i(x)| \\ &\leqslant \sum_{i=0}^{n} |f(x_i)\mathrm{L}_i(x)| \leqslant \|f\|_\infty \sum_{i=0}^{n} |\mathrm{L}_i(x)| \\ &\leqslant \Lambda_n \|f\|_\infty . \end{aligned}$$

On obtient alors

$$\|\mathcal{L}_n(f)\|_\infty \leqslant \Lambda_n \|f\|_\infty .$$

et donc

$$\sup_{\substack{f\in\mathcal{C}^0([a,b];\mathbb{R}) \\ f\neq 0}} \frac{\|\mathcal{L}_n(f)\|_\infty}{\|f\|_\infty} \leqslant \Lambda_n.$$

Pour obtenir l'égalité (4.14), il suffit donc, en reprenant les calculs précédents, de regarder si l'on peut trouver $\bar{x} \in [a,b]$ et $\bar{f} \in \mathcal{C}^0([a,b];\mathbb{R})$ vérifiant

$$|\mathcal{L}_n(\bar{f})(\bar{x})| = \Lambda_n \|\bar{f}\|_\infty .$$

On détermine $\bar{x}$ pour que la dernière majoration, correspondant à $\sum_{i=0}^{n} |\mathrm{L}_i(x)| \leqslant \Lambda_n$, devienne une égalité. L'application $x \mapsto \sum_{i=0}^{n} |\mathrm{L}_i(x)|$ étant continue sur le fermé borné $[a,b]$, il existe alors $\bar{x} \in [a,b]$ tel que

$$\Lambda_n = \max_{x\in[a,b]} \sum_{i=0}^{n} |\mathrm{L}_i(x)| = \sum_{i=0}^{n} |\mathrm{L}_i(\bar{x})|.$$

On va maintenant regarder s'il est possible de construire une fonction $\bar{f} \in \mathcal{C}^0([a,b];\mathbb{R})$ telle que

$$|\sum_{i=0}^{n} \bar{f}(x_i)\mathrm{L}_i(\bar{x})| = \sum_{i=0}^{n} |\bar{f}(x_i)\mathrm{L}_i(\bar{x})| = \|\bar{f}\|_\infty \sum_{i=0}^{n} |\mathrm{L}_i(\bar{x})|.$$

Si c'est le cas, on aura bien le résultat souhaité.
Sans déroger à la généralité, on peut supposer les points x_i ordonnés : $x_0 < x_1 < \cdots < x_n$. Pour avoir la première égalité, il faut que tous les $\bar{f}(x_i)\mathrm{L}_i(\bar{x})$ soient de même signe. On choisi le signe positif, et on impose par exemple les valeurs de $\bar{f}(x_i)$, $\forall i \in [\![0,n]\!]$,

$$\begin{cases} \bar{f}(x_i) &= 1, & \text{si } \mathrm{L}_i(\bar{x}) \geqslant 0, \\ \bar{f}(x_i) &= -1, & \text{si } \mathrm{L}_i(\bar{x}) < 0. \end{cases}$$

Il existe une multitude de fonctions de $\mathcal{C}^0([a,b];\mathbb{R})$ mais il faut controler son maximum pour qu'il vaille 1 et avoir ainsi $|\bar{f}(x_i)| = \|\bar{f}\|_\infty$, $\quad \forall i \in [\![0,n]\!]$. On peut alors choisir $\bar{f}$ affine sur chacun des intervalles $[x_k, x_{k+1}]$, $\forall k \in [\![0, n-1]\!]$, pusique l'on connait les valeurs aux extrémités de chaque intervalle ($+1$ ou -1). En dehors de ces intervalles, on prend par exemple $\bar{f}(x) = \bar{f}(x_0)$, $\forall x \in [a, x_0]$, et $\bar{f}(x) = \bar{f}(x_n)$, $\forall x \in [x_n, b]$. Cette fonction est par construction continue sur $[a,b]$ et vérifie

$$\begin{aligned} |\mathcal{L}_n(\bar{f})(\bar{x})| &= |\sum_{i=0}^{n} \bar{f}(x_i)\mathrm{L}_i(\bar{x})| \\ &= \sum_{i=0}^{n} |\bar{f}(x_i)\mathrm{L}_i(\bar{x})| = \|\bar{f}\|_\infty \sum_{i=0}^{n} |\mathrm{L}_i(\bar{x})| \\ &= \Lambda_n \|\bar{f}\|_\infty . \end{aligned}$$

Or on a

$$\|\mathcal{L}_n(\bar{f})\|_\infty = \sup_{x\in[a,b]} |\mathcal{L}_n(\bar{f})(x)| \geqslant |\mathcal{L}_n(\bar{f})(\bar{x})| = \Lambda_n \|\bar{f}\|_\infty .$$

ce qui donne

$$\sup_{\substack{f\in\mathcal{C}^0([a,b];\mathbb{R}) \\ f\neq 0}} \frac{\|\mathcal{L}_n(f)\|_\infty}{\|f\|_\infty} \geqslant \Lambda_n.$$

□

Théorème 4.7

Pour toute fonction $f \in \mathcal{C}^0([a,b];\mathbb{R})$, on a

$$\|f - \mathcal{L}_n(f)\|_\infty \leqslant (1+\Lambda_n) \inf_{\mathrm{Q}\in\mathbb{R}_n[X]} \|f - \mathrm{Q}\|_\infty \tag{4.15}$$

Preuve. Soit $\mathrm{Q} \in \mathbb{R}_n[X]$. Par unicité du théorème d'interpolation on a $\mathcal{L}_n(\mathrm{Q}) = \mathrm{Q}$ et alors

$$\begin{aligned} \|f - \mathcal{L}_n(f)\|_\infty &= \|f - \mathrm{Q} + \mathcal{L}_n(\mathrm{Q}) - \mathcal{L}_n(f)\|_\infty \\ &\leqslant \|f - \mathrm{Q}\|_\infty + \|\mathcal{L}_n(\mathrm{Q} - f)\|_\infty \quad \text{par linéarité de } \mathcal{L}_n \\ &\leqslant \|f - \mathrm{Q}\|_\infty + \Lambda_n \|f - \mathrm{Q}\|_\infty \end{aligned}$$

d'où le résultat. □

- Pour les points équidistants $x_i = a + ih$, $i \in [\![0,n]\!]$ et $h = (b-a)/n$, on a la minoration suivante (voir [3] p. 49)

$$\Lambda_n \geqslant \frac{2^n}{4n^2} \tag{4.16}$$

et le comportement asymptotique

$$\Lambda_n \approx \frac{2^{n+1}}{e.n\ln(n)} \quad \text{quand } n \to +\infty \tag{4.17}$$

- Pour les points de Tchebychev, on a la majoration suivante : il existe $C > 0$ tel que

$$\Lambda_n \leqslant C\ln(n) \tag{4.18}$$

et le comportement asymptotique

$$\Lambda_n \approx \frac{2}{\pi}\ln(n) \quad \text{quand } n \to +\infty \tag{4.19}$$

Proposition 4.8: admis

Pour toute famille de points d'interpolation, il existe une fonction $f \in \mathcal{C}^0([a,b];\mathbb{R})$ telle que la suite des polynômes d'interpolation associés ne converge pas uniformément.

Proposition 4.9: admis

Soit f une fonction lipschitzienne sur $[a,b]$ à valeurs réelles, i.e. il existe une constante $K \geqslant 0$ telle que $\forall (x,y) \in [a,b]^2$, on ait $|f(x) - f(y)| \leqslant K|x-y|$. Soient $n \in N^*$ et $x_0, \cdots, x_n$ les points de Tchebychev $[a,b]$. On note $\mathcal{L}_n(f)$ le polynôme d'interpolation de Lagrange associés aux couples de $(x_i, f(x_i))_{i\in[\![0,n]\!]}$.
Alors la suite $(\mathcal{L}_n(f))_{n\geqslant 1}$ des polynômes d'interpolation converge uniformémént vers f sur $[a,b]$.

Pour conclure, l'interpolation de Lagrange en des points équidistants n'est à utiliser qu'avec un nombre de points assez faible : des phénomènes d'instabilités pouvant apparaître.

4.2 Polynôme d'interpolation de Lagrange-Hermite

Exercice 4.2.1

Soient $(x_i, y_i, z_i)_{i\in[\![0,n]\!]}$ $n+1$ triplets de $\mathbb{R}^3$, où les x_i sont des points distincts deux à deux de l'intervalle $[a,b]$. Le polynôme d'interpolation de **Lagrange-Hermite**, noté H_n, associé aux $n+1$ triplets $(x_i, y_i, z_i)_{i\in[\![0,n]\!]}$, est défini par

$$\mathrm{H}_n(x_i) = y_i \quad \text{et} \quad \mathrm{H}_n'(x_i) = z_i, \ \forall i \in [\![0,n]\!] \tag{4.20}$$

Q. 1 *Quel est a priori le degré de* H_n *?*

On défini le polynôme P_n par

$$\mathrm{P}_n(x) = \sum_{i=0}^{n} y_i A_i(x) + \sum_{i=0}^{n} z_i B_i(x) \tag{4.21}$$

avec, pour $i \in [\![0,n]\!]$, A_i et B_i polynômes de degré au plus $2n+1$ indépendants des valeurs y_i et z_i.

Q. 2 *1. Déterminer des conditions suffisantes sur* A_i *et* B_i *pour que* P_n *vérifie* (4.20).

2. En déduire les expressions de A_i *et* B_i *en fonction de* L_i *et de* $L_i'(x_i)$ *où*

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

Q. 3 *Démontrer qu'il existe un unique polynôme d'interpolation de Lagrange-Hermite de degré au plus $2n+1$ défini par (4.20).*

Correction Exercice

Q. 1 On a $2n+2$ équations, donc à priori H_n est de degré $2n+1$.

Q. 2 1. D'après (4.21) on a pour tout $j \in [\![0,n]\!]$

$$\mathrm{P}_n(x_j) = \sum_{i=0}^{n} y_i A_i(x_j) + \sum_{i=0}^{n} z_i B_i(x_j)$$

Pour avoir $\mathrm{P}_n(x_j) = y_j$ il suffit d'avoir

$$A_i(x_j) = \delta_{i,j} \quad \text{et} \quad B_i(x_j) = 0, \quad \forall i \in [\![0,n]\!]. \tag{4.22}$$

De même, on a

$$\mathrm{P}'_n(x_j) = \sum_{i=0}^{n} y_i A'_i(x_j) + \sum_{i=0}^{n} z_i B'_i(x_j)$$

et donc pour avoir $\mathrm{P}'_n(x_j) = z_j$ il suffit d'avoir

$$A'_i(x_j) = 0 \quad \text{et} \quad B'_i(x_j) = \delta_{i,j}, \quad \forall i \in [\![0,n]\!]. \tag{4.23}$$

2. Soit $i \in [\![0,n]\!]$. On commence par déterminer le polynôme $A_i \in \mathbb{R}_{2n+1}[X]$ vérifiant

$$A_i(x_j) = \delta_{i,j} \quad \text{et} \quad A'_i(x_j) = 0, \quad \forall j \in [\![0,n]\!].$$

Les points $(x_j)_{j\in[\![0,n]\!]\setminus\{i\}}$ sont racines doubles de A_i. Le polynôme $L_i \in \mathbb{R}_n[X]$ admet les mêmes racines (simples) que A_i et donc $L_i^2 \in \mathbb{R}_{2n}[X]$ admet les mêmes racines doubles que A_i. On peut alors écrire

$$A_i(x) = \alpha_i(x) L_i^2(x) \quad \text{avec } \alpha_i(x) \in \mathbb{R}_1[X].$$

Il reste à déterminer le polynôme α_i. Or on a

$$A_i(x_i) = 1 \quad \text{et} \quad A'_i(x_i) = 0.$$

Comme $L_i(x_i) = 1$, on obtient

$$A_i(x_i) = \alpha_i(x_i) L_i^2(x_i) = \alpha_i(x_i) = 1$$

et

$$A'_i(x_i) = \alpha'_i(x_i) L_i^2(x_i) + 2\alpha_i(x_i) L'_i(x_i) L_i(x_i) = \alpha'_i(x_i) + 2\alpha_i(x_i) L'_i(x_i) = 0$$

c'est à dire

$$\alpha_i(x_i) = 1 \quad \text{et} \quad \alpha'_i(x_i) = -2L'_i(x_i).$$

Comme α_i est un polynôme de degré 1 on en déduit

$$\alpha_i(x) = 1 - 2L'_i(x_i)(x - x_i)$$

et donc

$$A_i(x) = (1 - 2L'_i(x_i)(x - x_i)) L_i^2(x). \tag{4.24}$$

On détermine ensuite le polynôme $B_i \in \mathbb{R}_{2n+1}[X]$ vérifiant

$$B_i(x_j) = 0 \quad \text{et} \quad B'_i(x_j) = \delta_{i,j}, \quad \forall j \in [\![0,n]\!].$$

Les points $(x_j)_{j\in[\![0,n]\!]\setminus\{i\}}$ sont racines doubles de B_i et le point x_i est racine simple. Le polynôme $L_i^2 \in \mathbb{R}_{2n}[X]$ admet les mêmes racines doubles. On peut alors écrire

$$B_i(x) = C(x - x_i) L_i^2(x) \quad \text{avec } C \in \mathbb{R}.$$

Il reste à déterminer la constante C. Or $L_i(x_i) = 1$ et comme $B'_i(x_i) = 1$ on obtient

$$B'_i(x_i) = C L_i^2(x_i) + 2C(x_i - x_i) L'_i(x_i) L_i(x_i) = C = 1$$

ce qui donne

$$B_i(x) = (x - x_i) L_i^2(x). \tag{4.25}$$

On vient de démontrer l'existence en construisant un polynôme de degré $2n+1$ vérifiant (4.20).

Q. 3 Deux démonstrations pour l'unicité sont proposées (la deuxième donne aussi l'existence).

dém. 1: Soient P et Q deux polynômes de $\mathbb{R}_{2n+1}[X]$ vérifiant (4.20). Le polynôme $R = P-Q \in \mathbb{R}_{2n+1}[X]$ admet alors $n+1$ racines doubles distinctes :$(x_0, \cdots, x_n)$. Or le seul polynôme de $\mathbb{R}_{2n+1}[X]$ ayant $n+1$ racines doubles est le polynôme nul et donc $R = 0$, i.e. $P = Q$.

dém. 2: Soit $\Phi : \mathbb{R}_{2n+1}[X] \longrightarrow \mathbb{R}^{2n+2}$ définie par

$$\forall P \in \mathbb{R}_{2n+1}[X], \quad \Phi(P) = (P(x_0), \cdots, P(x_n), P'(x_0), \cdots, P'(x_n)).$$

L'existence et l'unicité du polynôme H_n est équivalente à la bijectivité de l'application Φ. Or celle-ci est une application linéaire entre deux espaces de dimension $2n+2$. Elle est donc bijective si et seulement si elle injective (ou surjective). Pour vérifier l'injectivité de Φ il est nécessaire et suffisant de vérifier que son noyau est réduit au polynôme nul.
Soit $P \in \ker \Phi$. On a alors $\Phi(P) = \mathbf{0}_{2n+2}$ et donc $(x_0, \cdots, x_n)$ sont $n+1$ racines doubles distinctes de P. Or le seul polynôme de $\mathbb{R}_{2n+1}[X]$ ayant $n+1$ racines doubles est le polynôme nul et donc $P = 0$.

◇

Definition 4.10

Soient $n \in \mathbb{N}^*$ et $(x_i, y_i, z_i)_{i\in[\![0,n]\!]}$ $n+1$ triplets de $\mathbb{R}^3$, où les x_i sont des points distincts deux à deux de l'intervalle $[a, b]$. Le **polynôme d'interpolation de Lagrange-Hermite**, noté H_n, associé aux $n+1$ triplets $(x_i, y_i, z_i)_{i\in[\![0,n]\!]}$, est défini par

$$H_n(x) = \sum_{i=0}^{n} y_i A_i(x) + \sum_{i=0}^{n} z_i B_i(x) \tag{4.26}$$

avec

$$A_i(x) = (1 - 2L_i'(x_i)(x - x_i))L_i^2(x) \quad \text{et} \quad B_i(x) = (x - x_i)L_i^2(x) \tag{4.27}$$

où

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

Théorème 4.11

Le **polynôme d'interpolation de Lagrange-Hermite**, H_n, associé aux $n+1$ triplets $(x_i, y_i, z_i)_{i\in[\![0,n]\!]}$, est l'unique polynôme de degré au plus $2n+1$, vérifiant

$$H_n(x_i) = y_i \quad \text{et} \quad H_n'(x_i) = z_i, \ \forall i \in [\![0, n]\!] \tag{4.28}$$

Exercice 4.2.2

Soit $f \in \mathcal{C}^{2n+2}([a, b]; \mathbb{R})$. On suppose de plus que, $\forall i \in [\![0, n]\!]$, $x_i \in [a, b]$, $y_i = f(x_i)$ et $z_i = f'(x_i)$. On note

$$\pi_n^2(x) = \prod_{i=0}^{n}(x - x_i)^2$$

et H_n le polynôme d'interpolation de Lagrange-Hermite associé aux triplets $(x_i, f(x_i), f'(x_i))_{i\in[\![0,n]\!]}$.

Q. 1 *Montrer que*

$$|f(x) - H_n(x)| \leqslant \frac{\left\|f^{(2n+2)}\right\|_\infty}{(2n+2)!}\pi_n^2(x). \tag{4.29}$$

Indications : *Etudier les zéros de la fonction* $F(y) = f(y) - \mathrm{H}_n(y) - \dfrac{f(x)-\mathrm{H}_n(x)}{\pi_n^2(x)}\pi_n^2(y)$ *et appliquer le théorème de Rolle.*

Correction Exercice

Q. 1 Soit $i \in [\![1,n]\!]$, on a $f(x_i) - \mathrm{H}_n(x_i) = 0$ et l'inégalité (4.29) est donc vérifiée pour $x = x_i$.
Soit $x \in [a,b]$ tel que $x \neq x_i$, $\forall i \in [\![1,n]\!]$. On a alors $\pi_n^2(x) \neq 0$. Comme $f \in \mathcal{C}^{2n+2}([a;b];\mathbb{R})$, $\mathrm{H}_n \in \mathbb{R}_{2n+1}[X]$ et $\pi_n \in \mathbb{R}_{n+1}[X]$, on en déduit que

$$F \in \mathcal{C}^{2n+2}([a;b];\mathbb{R}).$$

On note que π_n^2 admet $(x_0,\cdots,x_n)$ comme racines doubles distinctes. Par construction $f - \mathrm{H}_n$ admet les mêmes racines doubles. On en déduit alors que F admet aussi $(x_0,\cdots,x_n)$ comme racines doubles. De plus, on a $F(x) = 0$ (i.e. x est racine simple) et donc

$$F \text{ admet au moins } 2n+3 \text{ racines (comptées avec leurs multiplicités).}$$

Les points $x, x_0, \cdots, x_n$ étant distincts, la fonction F' admet par le théorème de Rolle $n+1$ zeros distincts entre eux et distincts des points $x, x_0, \cdots, x_n$. De plus les points $x_0, \cdots, x_n$ sont racines de F' puisque racines doubles de F. On en déduit alors que

$$F' \text{ admet au moins } 2n+2 \text{ racines distinctes deux à deux.}$$

Par applications successives du théorème de Rolle, on abouti a :

$$F^{(2n+2)} \text{ admet au moins une racine notée } \xi_x \in]a,b[.$$

On a alors

$$F^{(2n+2)}(\xi_x) = 0 = f^{(2n+2)}(\xi_x) - \mathrm{H}_n^{(2n+2)}(\xi_x) - \frac{f(x)-\mathrm{H}_n(x)}{\pi_n^2(x)}\frac{d^{2n+2}\pi_n^2}{dx^{2n+2}}(\xi_x)$$

Comme $\mathrm{H}_n \in \mathbb{R}_{2n+1}[X]$ on a $\mathrm{H}_n^{(2n+2)} \equiv 0$. De plus comme $\pi_n^2(x) = \prod_{i=0}^{n}(x-x_i)^2 \in \mathbb{R}_{2n+2}[X]$ sa dérivée d'ordre $2n+2$ est constante et

$$\frac{d^{2n+2}\pi_n^2}{dx^{2n+2}} = (2n+2)!$$

On en déduit alors

$$f^{(2n+2)}(\xi_x) = \frac{f(x)-\mathrm{H}_n(x)}{\pi_n^2(x)}(2n+2)!$$

On a donc montrer que $\forall x \in [a,b]$ $\exists \xi_x \in]a,b[$ tels que

$$f(x) - \mathrm{H}_n(x) = \frac{\pi_n^2(x)}{(2n+2)!}f^{(2n+2)}(\xi_x).$$

Comme $\pi_n^2(x) \geqslant 0$ on obtient bien (4.29).

◇

Théorème 4.12

Soient $n \in \mathbb{N}^*$ et $x_0, \cdots, x_n$, $n+1$ points distincts de l'intervalle $[a,b]$. Soient $f \in \mathcal{C}^{2n+2}([a;b];\mathbb{R})$ et H_n le polynôme d'interpolation de Lagrange-Hermite associé aux $n+1$ triplets $(x_i, f(x_i), f'(x_i))_{i\in[\![0,n]\!]}$. On a alors $\forall x \in [a,b]$, $\exists \xi_x \in (\min(x_i,x), \max(x_i,x))$, tels que

$$f(x) - \mathrm{H}_n(x) = \frac{f^{(2n+2)}(\xi_x)}{(2n+2)!}\prod_{i=0}^{n}(x-x_i)^2 \tag{4.30}$$

Exercice 4.2.3

Ecrire une fonction algorithmique HERMITE permettant de calculer H_n (polynôme d'interpolation de Lagrange-Hermite associé aux $n+1$ triplets $(x_i, y_i, z_i)_{i\in[\![0,n]\!]}$) en $t \in \mathbb{R}$.

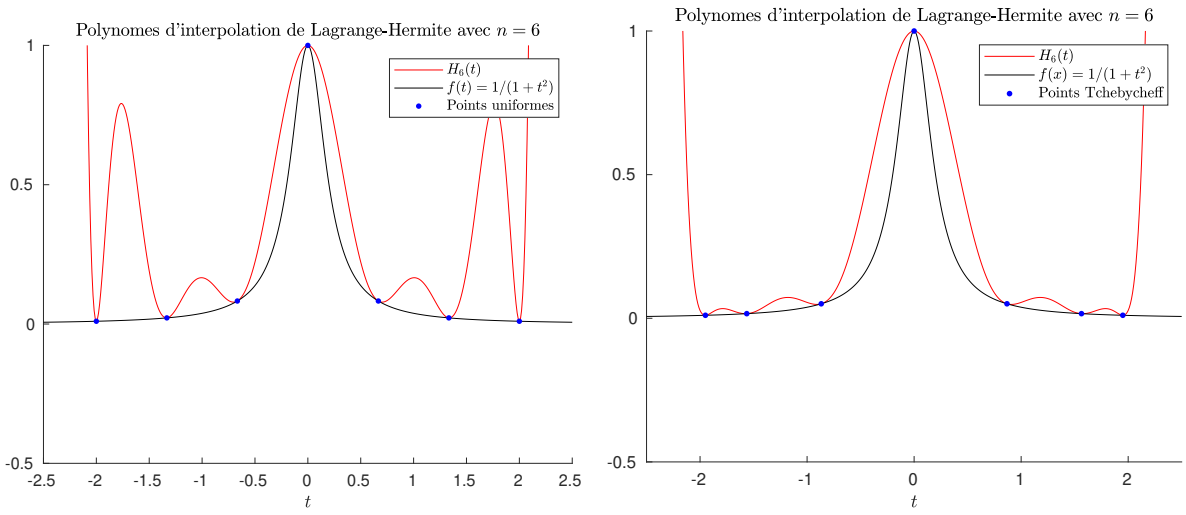


Figure 4.9: Polynôme d'interpolation de lagrange-Hermite avec $n = 6$ (7 points) pour la fonction $f : x \longrightarrow 1/(1 + 25x^2)$. A gauche avec des points uniformément répartis et à droite avec des points de Tchebychev

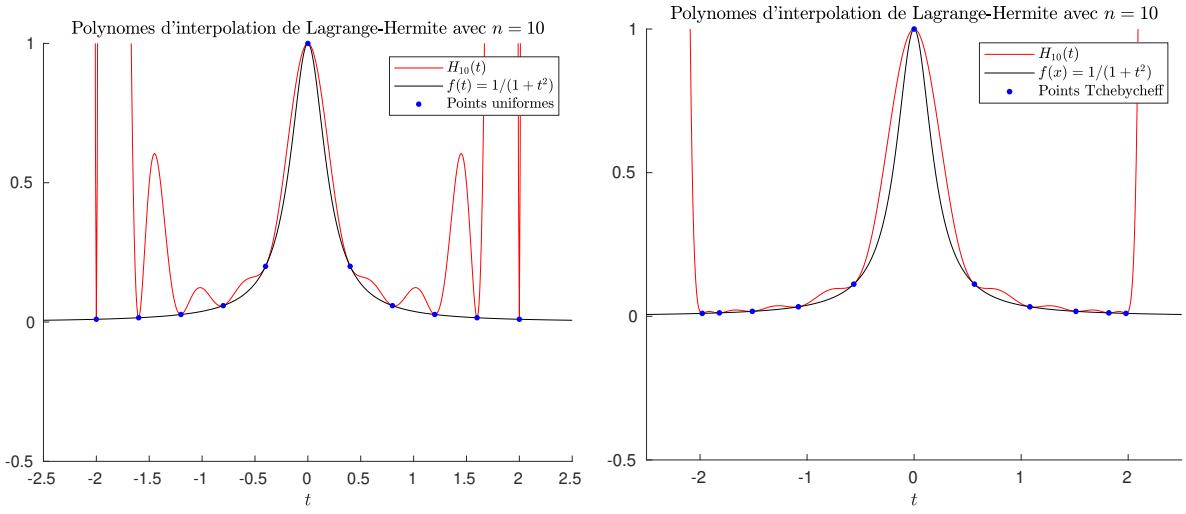


Figure 4.10: Polynôme d'interpolation de lagrange-Hermite avec $n = 10$ (11 points) pour la fonction $f : x \longrightarrow 1/(1 + 25x^2)$. A gauche avec des points uniformément répartis et à droite avec des points de Tchebychev

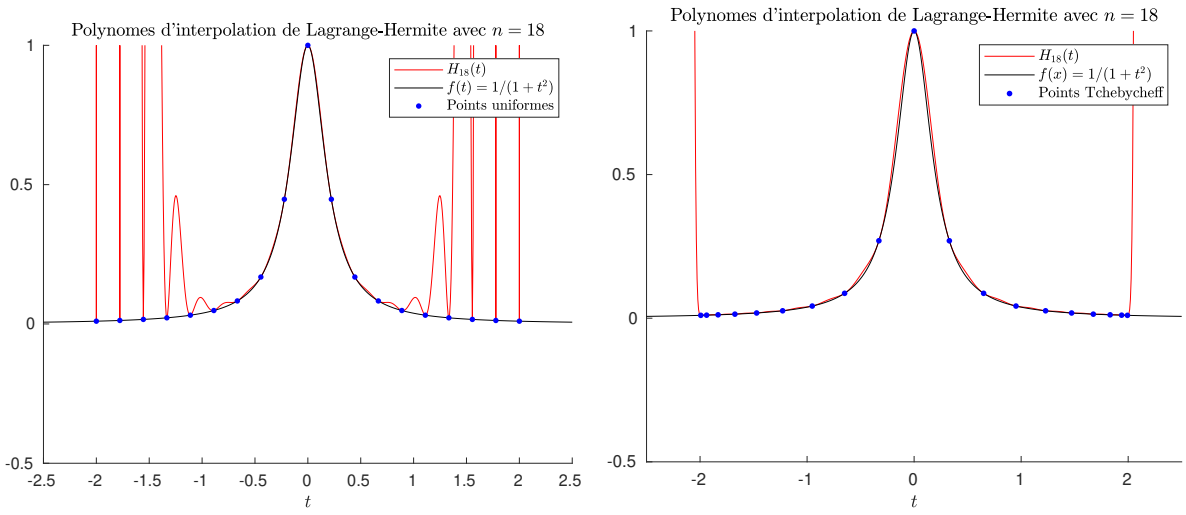


Figure 4.11: Polynôme d'interpolation de lagrange-Hermite avec $n = 18$ (19 points) pour la fonction $f : x \longrightarrow 1/(1 + 25x^2)$. A gauche avec des points uniformément répartis et à droite avec des points de Tchebychev

Correction Exercice

But : Calculer le polynôme $\mathrm{H}_n(t)$ définit par (4.26)

Données : $\boldsymbol{X}$: vecteur/tableau de $\mathbb{R}^{n+1}$, $X(i) = x_{i-1}\ \forall i \in [\![1, n+1]\!]$ et $X(i) \neq X(j)$ pour $i \neq j$,
$\boldsymbol{Y}$: vecteur/tableau de $\mathbb{R}^{n+1}$, $Y(i) = y_{i-1}\ \forall i \in [\![1, n+1]\!]$,
$\boldsymbol{Z}$: vecteur/tableau de $\mathbb{R}^{n+1}$, $Z(i) = z_{i-1}\ \forall i \in [\![1, n+1]\!]$,
t : un réel.

Résultat : pH : le réel $\mathrm{pH} = \mathrm{H}_n(t)$.

D'après la Définition 4.10, on a

$$\mathrm{H}_n(t) = \sum_{i=0}^{n} y_i A_i(t) + \sum_{i=0}^{n} z_i B_i(t) = \sum_{i=0}^{n} (y_i A_i(t) + z_i B_i(t))$$

avec

$$A_i(t) = (1 - 2L_i'(x_i)(t - x_i))L_i^2(t) \quad \text{et} \quad B_i(t) = (t - x_i)L_i^2(t)$$

où

$$L_i(t) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{t - x_j}{x_i - x_j}.$$

Pour rendre effectif le calcul de $\mathrm{H}_n(t)$, il reste à déterminer $L_i'(x_i)$. On a

$$L_i'(x) = \sum_{\substack{k=0 \\ k \neq i}}^{n} \frac{1}{x_i - x_k} \prod_{\substack{j=0 \\ j \neq i \\ j \neq k}}^{n} \frac{t - x_j}{x_i - x_j}$$

d'où

$$L_i'(x_i) = \sum_{\substack{k=0 \\ k \neq i}}^{n} \frac{1}{x_i - x_k}. \tag{4.31}$$

La fonction que l'on va écrire use (et certains diront abuse) de fonctions.

Algorithme 4.2 Fonction HERMITE permettant de calculer le polynôme d'interpolation de Lagrange-Hermite $\mathrm{H}_n(t)$ définit par (4.26)

```
1: Fonction pH ← HERMITE ( X, Y, Z, t )
2:     pH ← 0
3:     Pour i ← 0 à n faire
4:         pH ← pH + polyA(i, X, t) * Y(i + 1) + polyB(i, X, t) * Z(i + 1)
5:     Fin Pour
6: Fin Fonction
```

Les différentes fonctions utilisées pour la fonction HERMITE (directement ou indirectement) sont les suivantes :

polyA : calcul du polynôme A_i en t, (données i, X, t)

polyB : calcul du polynôme B_i en t, (données i, X, t)

polyL : calcul du polynôme L_i en t, (données i, X, t)

polyLp : calcul de $L_i'(x_i)$, (données i, X)

Algorithme 4.3 Fonction polyA permettant de calculer le polynôme A_i en $t \in \mathbb{R}$ donné par $A_i(t) = (1 - 2L_i'(x_i)(t - x_i))L_i^2(t)$

```
1: Fonction y ← polyA ( i, X, t )
2:     y ← (1 − 2 * polyLp(i, X) * (t − X(i + 1))) * (polyL(i, X, t))^2
3: Fin Fonction
```

Algorithme 4.4 Fonction polyB permettant de calculer le polynôme B_i en $t \in \mathbb{R}$ donné par $B_i(t) = (t - x_i)L_i^2(t)$

```
1: Fonction y ← polyB ( i, X, t )
2:     y ← (t − X(i + 1)) * (polyL(i, X, t))^2
3: Fin Fonction
```

Algorithme 4.5 Fonction polyL permettant de calculer le polynôme L_i en $t \in \mathbb{R}$ donné par $L_i(t) = \prod_{j=0, j\neq i}^{n} \frac{t - x_j}{x_i - x_j}$

```
1: Fonction y ← polyL ( i, X, t )
2:   y ← 1
3:   Pour j ← 0 à n, (j ~= i) faire
4:     y ← y * (t − X(j + 1))/(X(i + 1) − X(j + 1))
5:   Fin Pour
6: Fin Fonction
```

Algorithme 4.6 Fonction polyLp permettant de calculer $L_i'(x_i) = \sum_{k=0, k\neq i}^{n} \frac{1}{x_i - x_k}$

```
1: Fonction y ← polyLp ( i, X )
2:   y ← 0
3:   Pour k ← 0 à n, (k ~= i) faire
4:     y ← y + 1/(X(i + 1) − X(k + 1))
5:   Fin Pour
6: Fin Fonction
```

Bien évidemment une telle écriture est loin d'être optimale mais elle a l'avantage d'être facile à programmer et facile à lire car elle "colle" aux formules mathématiques.

On laisse le soin au lecteur d'écrire des fonctions plus performantes... ◇

4.3 Exercices

Exercice 4.3.1

Q. 1 *1. Ecrire explicitement un polynôme* P *de degré* 2 *passant par les points* $A = (1; 2)$, $B = (2; 6)$ *et* $C = (3; 12)$.

2. Démontrer que le polynôme P *est l'unique polynôme de degré* 2 *passant par les points* A, B *et* C.

Q. 2 *Ecrire explicitement un polynôme* Q *de degré* 3 *tel que*

$$\begin{aligned} &\mathrm{Q}(1) = 4, && \mathrm{Q}(2) = 5, \\ &\mathrm{Q}'(1) = 3, && \mathrm{Q}'(2) = 2. \end{aligned}$$

Exercice 4.3.2

Q. 1 *Construire les polynômes* h_{00}, h_{10}, h_{01} *et* h_{11} *de degré* 3 *vérifiant*

$$h_{00}(0) = 1, h'_{00}(0) = h_{00}(1) = h'_{00}(1) = 0, \tag{4.32}$$
$$h_{10}(1) = 1, h_{10}(0) = h'_{00}(0) = h'_{10}(1) = 0, \tag{4.33}$$
$$h'_{01}(0) = 1, h_{01}(0) = h_{01}(1) = h'_{01}(1) = 0, \tag{4.34}$$
$$h'_{11}(1) = 1, h_{11}(0) = h'_{11}(0) = h_{11}(1) = 0; \tag{4.35}$$

On pose

$$\mathrm{P}(x) = \alpha h_{00}(x) + \beta h_{10}(x) + \gamma h_{01}(x) + \delta h_{11}(x). \tag{4.36}$$

Q. 2 *Quelles sont les particularités de* P *?*

Soient a et b deux réels, $a < b$ et Q le polynôme de degré 3 vérifiant

$$\mathrm{Q}(a) = u_a, \ \mathrm{Q}'(a) = v_a, \ \mathrm{Q}(b) = u_b \ \text{ et } \mathrm{Q}'(b) = v_b.$$

Q. 3 *Exprimer le polynôme* Q *avec les fonctions* h_{00}, h_{10}, h_{01} *et* h_{11}.

Exercice 4.3.3

Soit $t_0 < t_1$ deux nombres réels et soit ε un réel tel que $0 < \varepsilon < \frac{t_1 - t_0}{2}$.

Q. 1 *Expliciter un polynôme* P_ε *de degré* 3 *tel que*

$$\mathrm{P}_\varepsilon(t_0) = \mathrm{P}_\varepsilon(t_0+\varepsilon) = 1, \tag{4.37}$$
$$\mathrm{P}_\varepsilon(t_1) = \mathrm{P}_\varepsilon(t_1+\varepsilon) = 0. \tag{4.38}$$

On note $\Phi_0(t) = \lim_{\varepsilon\to 0} \mathrm{P}_\varepsilon(t)$.

Q. 2 *1. Montrer que* $\Phi_0(t_0) = 1$, $\Phi_0'(t_0) = 0$, $\Phi_0(t_1) = 0$ *et* $\Phi_0'(t_1) = 0$ *(i.e.* Φ_0 *est une fonction de base des polynômes de degré* 3 *pour l'interpolation de Hermite).*

2. Peut-on obtenir toutes les fonctions de base de Hermite par des procédés analogues. Si oui, expliquer comment!

Exercice 4.3.4: ★★★★★

Soient $(x_i)_{i\in\mathbb{N}}$ une suite de points distincts de l'intervalle $[a,b]$ et f une fonction définie sur $[a,b]$ à valeurs réelles.
On désigne par $f[]$ les **différences divisées** de la fonction f définie par

$$f[x_i] = f(x_i),\ \forall i \in \mathbb{N},\ \text{(ordre 0)} \tag{4.39}$$

et

$$f[x_k,\dots,x_{k+r}] = \frac{f[x_{k+1},\dots,x_{k+r}] - f[x_k,\dots,x_{k+r-1}]}{x_{k+r}-x_k},\ \forall k \in \mathbb{N},\ \forall r \in \mathbb{N}^*,\ \text{(ordre r)} \tag{4.40}$$

Q. 1 *Montrer que*

$$f[x_k,\dots,x_{k+r}] = \sum_{i=k}^{k+r} \frac{f(x_i)}{\prod\limits_{\substack{j=k\\ j\neq i}}^{k+r}(x_i-x_j)},\ \forall k \in \mathbb{N},\ \forall r \in \mathbb{N}^*. \tag{4.41}$$

Q. 2 *Soit σ une permutation des entiers $\{k,\dots,k+r\}$. Montrer que*

$$f[x_k,\dots,x_{k+r}] = f[x_{\sigma(1)},\dots,x_{\sigma(r+1)}]. \tag{4.42}$$

On note $\mathrm{Q}_{k,r}$ le polynôme d'interpolation associé aux $r+1$ couples $(x_{k+i}, f(x_{k+i}))_{i\in[\![0,r]\!]}$.

Q. 3
1. *Exprimer le polynôme $\mathrm{Q}_{k,1}$ en fonction de $\mathrm{Q}_{k,0}$.*
2. *Exprimer le polynôme $\mathrm{Q}_{k,1}$ en fonction de $\mathrm{Q}_{k,0}$ et $\mathrm{Q}_{k+1,0}$.*
3. *En déduire que*

$$\mathrm{Q}_{k,1}(x) = \mathrm{Q}_{k,0}(x) + f[x_k,x_{k+1}](x-x_k). \tag{4.43}$$

Q. 4
1. *Exprimer le polynôme $\mathrm{Q}_{k,2}$ en fonction de $\mathrm{Q}_{k,1}$.*
2. *Exprimer le polynôme $\mathrm{Q}_{k,2}$ en fonction de $\mathrm{Q}_{k,1}$ et $\mathrm{Q}_{k+1,1}$.*
3. *En déduire que*

$$\mathrm{Q}_{k,2}(x) = \mathrm{Q}_{k,1}(x) + f[x_k,x_{k+1},x_{k+2}](x-x_k)(x-x_{k+1}). \tag{4.44}$$

Q. 5
1. *Montrer que*

$$\mathrm{Q}_{k,r}(x) = \mathrm{Q}_{k,r-1}(x) + f[x_k,\dots,x_{k+r}]\prod_{j=0}^{r-1}(x-x_{k+j}) \tag{4.45}$$

Indication : *Effectuer une démonstration par récurrence en écrirant le polynôme $\mathrm{Q}_{k,r}$ sous deux formes : l'une en fonction de $\mathrm{Q}_{k,r-1}$ et l'autre en fonction de $\mathrm{Q}_{k,r-1}$ et $\mathrm{Q}_{k+1,r-1}$.*

2. *En déduire*

$$\mathrm{Q}_{k,r}(x) = f[x_k] + \sum_{i=1}^{r} f[x_k,\dots,x_{k+i}]\prod_{j=0}^{i-1}(x-x_{k+j}) \tag{4.46}$$

Q. 6 *On suppose que $f \in \mathcal{C}^r([a,b];\mathbb{R})$. Montrer qu'il existe $\xi \in]\min\limits_{i\in[\![0,r]\!]} x_{k+i}; \max\limits_{i\in[\![0,r]\!]} x_{k+i}[$ tel que*

$$f[x_k,\dots,x_{k+r}] = \frac{f^{(r)}(\xi)}{r!}. \tag{4.47}$$

Q. 7 *On suppose $f \in \mathcal{C}^{r+1}([a;b];\mathbb{R})$. Montrer que, $\forall x \in [a;b]$, il existe ξ_x appartenant au plus petit intervalle fermé contenant $x, x_k,\dots,x_{k+r}$ tel que*

$$f(x) - \mathrm{Q}_{k,r}(x) = \frac{\prod\limits_{j=0}^{r}(x-x_{k+j})}{(r+1)!} f^{(r+1)}(\xi_x). \tag{4.48}$$

Exercice 4.3.5

Soit $\boldsymbol{X} = (x_i)_{i\in[\![0,n]\!]}$ $n+1$ points deux à deux distincts de l'intervalle $[a,b]$. On note s le changement de variables $s : t \longrightarrow a+(b-a)t$ de $[0,1]$ à valeurs dans $[a,b]$. Pour tout $i \in [\![0,n]\!]$, on note $t_i = s^{-1}(x_i) = \frac{x_i-a}{b-a}$ et $\boldsymbol{T} = (t_i)_{i\in[\![0,n]\!]}$.
Les polynômes d'interpolation de Lagrange $\mathcal{L}_n(f)$ et $\mathcal{L}_n(g)$ associés respectivement aux points

$(x_i, f(x_i))_{i\in[\![0,n]\!]}$ et $(t_i, g(t_i))_{i\in[\![0,n]\!]}$ sont définis par

$$\mathcal{L}_n(f)(x) = \sum_{i=0}^{n} L_i^{\boldsymbol{X}}(x) f(x_i) \text{ avec } L_i^{\boldsymbol{X}}(x) = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x - x_j}{x_i - x_j}$$

$$\mathcal{L}_n(g)(t) = \sum_{i=0}^{n} L_i^{\boldsymbol{T}}(t) g(t_i) \text{ avec } L_i^{\boldsymbol{T}}(t) = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t - t_j}{t_i - t_j}$$

Q. 1 *Montrer que* $L_i^{\boldsymbol{X}} \circ s = L_i^{\boldsymbol{T}}$.

Q. 2 *En déduire que* $\mathcal{L}_n(f \circ s) = \mathcal{L}_n(f) \circ s = \mathcal{L}_n(g)$.

Correction Exercice

Q. 1

$$\begin{aligned} L_i^{\boldsymbol{X}} \circ s(t) &= \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{s(t) - x_j}{x_i - x_j} = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{s(t) - s(t_j)}{s(t_i) - s(t_j)} \\ &= \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{a + t(b-a) - (a + t_j(b-a))}{a + t_i(b-a) - (a + t_j(b-a))} \\ &= \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t - t_j}{t_i - t_j} = L_i^{\boldsymbol{T}}(t) \end{aligned}$$

Q. 2 On a

$$\begin{aligned} \mathcal{L}_n(f \circ s)(t) &\overset{\text{def}}{=} \sum_{i=0}^{n} L_i^{\boldsymbol{T}}(t) f \circ s(t_i) \\ &= \sum_{i=0}^{n} L_i^{\boldsymbol{T}}(t) g(t_i) = \mathcal{L}_n(g) \end{aligned}$$

et

$$\begin{aligned} \mathcal{L}_n(f) \circ s(t) &\overset{\text{def}}{=} \sum_{i=0}^{n} L_i^{\boldsymbol{X}} \circ s(t) f(x_i) \\ &= \sum_{i=0}^{n} L_i^{\boldsymbol{T}}(t) f \circ s(t_i) = \mathcal{L}_n(g) \end{aligned}$$

◇

Exercice 4.3.6: Spline cubique

Soient $n \geqslant 3$ un entier et $a = x_0 < x_1 \ldots < x_{n-1} < x_n = b$ une discrétisation régulière de l'intervalle $[a, b]$. On note $h = x_{k+1} - x_k$. Une fonction s définie sur $[a; b]$ à valeurs réelles s'appelle **spline cubique** si elle est deux fois continûment différentiable et si, sur chaque intervalle $[x_{k-1}; x_k]$, elle est polynomiale de degré inférieur ou égal à 3.

Soit $f \in \mathcal{C}^2([a; b]; \mathbb{R})$ et s une spline cubique vérifiant

$$s(x_i) = f(x_i) = f_i, \ \forall i \in [\![0, n]\!]. \tag{4.49}$$

Q. 1 *Montrer que si*

$$s''(b)(f'(b) - s'(b)) = s''(a)(f'(a) - s'(a)) \tag{4.50}$$

alors

$$\int_a^b (s''(x))^2 dx \leqslant \int_a^b (f''(x))^2 dx. \tag{4.51}$$

Indications : Poser $r = f - s$ et montrer par intégrations par parties que $\int_a^b s''(x)r''(x)dx = 0$.

Soient $k \in [\![1, n]\!]$ et S_k un polynôme de degré inférieur ou égal à 3 vérifiant

$$\begin{cases} S_k(x_{k-1}) &= f_{k-1} & (4.52a)\\ \mathrm{S}_k(x_k) &= f_k & (4.52b)\\ \mathrm{S}_k''(x_{k-1}) &= m_{k-1} & (4.52c)\\ \mathrm{S}_k''(x_k) &= m_k. & (4.52d) \end{cases}$$

Q. 2 *1. Montrer l'existence et l'unicité du polynôme* S_k.

2. Montrer que polynôme S_k *peut s'écrire sous la forme*

$$\mathrm{S}_k(x) = a_k(x_k - x)^3 + b_k(x - x_{k-1})^3 + \alpha_k(x_k - x) + \beta_k(x - x_{k-1}) \tag{4.53}$$

en explicitant les coefficients $(a_k, b_k, \alpha_k, \beta_k)$ *en fonction de* $(f_{k-1}, f_k, m_{k-1}, m_k)$ *et* h.

On note g la fonction dont la restriction à chaque intervalle $[x_{k-1}; x_k]$, $k \in [\![1, n]\!]$, est S_k.

Q. 3 *1. Vérifier que* g *est bien définie sur* $[a; b]$.

2. Montrer que g *est une spline cubique si et seulement si,* $\forall k \in [\![1, n-1]\!]$,

$$m_{k+1} + 4m_k + m_{k-1} = \frac{6}{h^2}(f_{k+1} - 2f_k + f_{k-1}). \tag{4.54}$$

Q. 4 *1. Montrer qu'une condition nécessaire et suffisante pour que* g *soit une spline cubique et vérifie* $g''(a) = 0$, $g''(b) = 0$, *est que le vecteur* $\boldsymbol{M} \in \mathbb{R}^{n+1} = (m_0, m_1, \ldots, m_n)^t$ *soit solution d'un système linéaire de la forme*

$$\mathbb{A}\boldsymbol{M} = \boldsymbol{b} \tag{4.55}$$

que l'on précisera.

2. Montrer que la matrice $\mathbb{A}$ *est inversible.*

Correction Exercice

Théorème 4.13: Intégration par parties

Soient $u \in \mathcal{C}^1([a; b]; \mathbb{R})$ et $v \in \mathcal{C}^1([a; b]; \mathbb{R})$ alors

$$\int_a^b u'(x)v(x)dx = [u(x)v(x)]_a^b - \int_a^b u(x)v'(x)dx.$$

Q. 1 On pose $r = f - s$. On a alors $r \in \mathcal{C}^2([a; b]; \mathbb{R})$ et,

$$\forall i \in [\![0, n]\!],\ r(x_i) = 0. \tag{4.56}$$

De plus

$$\begin{aligned} \int_a^b (f''(x))^2 dx &= \int_a^b (s''(x) + r''(x))^2 dx \\ &= \int_a^b (s''(x))^2 dx + 2\int_a^b s''(x)r''(x)dx + \int_a^b (r''(x))^2 dx \end{aligned} \tag{4.57}$$

Montrons que $\int_a^b s''(x)r''(x)dx = 0$.

On ne peut effectuer une intégration par partie pour $\int_a^b s''(x)r''(x)dx$ car r'' et s'' ne sont pas dérivables. Par contre, on a

$$\int_a^b s''(x)r''(x)dx = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} s''(x)r''(x)dx$$

et, sur chaque intervalle $[x_{i-1}, x_i]$, s'' est un polynôme de degré au plus 1. On a donc $s'' \in \mathcal{C}^1([x_{i-1}, x_i]; \mathbb{R})$ et $r' \in \mathcal{C}^1([x_{i-1}, x_i]; \mathbb{R})$ et il est alors possible de faire une intégration par partie avec $u = r'$ et $v = s''$

sur $[x_{i-1}, x_i]$:

$$\int_{x_{i-1}}^{x_i} s''(x)r''(x)dx = [s''(x)h'(x)]_{x_{i-1}}^{x_i} - \int_{x_{i-1}}^{x_i} s'''(x)r'(x)dx. \tag{4.58}$$

Or, sur $[x_{i-1}, x_i]$, $s \in \mathbb{R}^3[X]$ et donc s''' est constante, ce qui donne, en utilisant (4.56),

$$\int_{x_{i-1}}^{x_i} s'''(x)r'(x)dx = s'''(x_i)\int_{x_{i-1}}^{x_i} r'(x)dx = s'''(x_i)(r(x_i) - r(x_{i-1})) = 0.$$

De (4.58), On a alors, $\forall i \in [\![1, n]\!]$,

$$\int_{x_{i-1}}^{x_i} s'''(x)r'(x)dx = s''(x_i)h'(x_i) - s''(x_{i-1})h'(x_{i-1}).$$

En sommant, on abouti a

$$\int_a^b s''(x)r''(x)dx = s''(x_n)h'(x_n) - s''(x_0)h'(x_0) = s''(b)h'(b) - s''(a)h'(a).$$

Sous l'hypothèse (4.50) on a bien $\int_a^b s''(x)r''(x)dx = 0$. L'équation (4.57) devient alors

$$\int_a^b (f''(x))^2dx = \int_a^b (s''(x))^2dx + \int_a^b (r''(x))^2dx.$$

D'où

$$\int_a^b (f''(x))^2dx \geqslant \int_a^b (s''(x))^2dx.$$

Q. 2 1. Soit $\Phi_k : \mathbb{R}^3[X] \longrightarrow \mathbb{R}^4$ définie par

$$\Phi_k(P) = (P(x_{k-1}), P(x_k), P''(x_{k-1}), P''(x_k))^t.$$

L'existence et l'unicité du polynôme S_k est équivalente à la bijectivité de Φ_k. Cette dernière étant une application entre deux espaces vectoriels de même dimension finie 4, elle est bijective si et seulement si elle est injective. Pour établir l'injectivité de Φ_k il faut montrer que son noyau est réduit au polynôme nul.
Soit $P \in \ker \Phi_k$ alors $\Phi_k(P) = 0_{\mathbb{R}^4}$. On en déduit que x_{k-1} et x_k sont racines de P, et P s'écrit alors sous la forme

$$P(x) = (x - x_{k-1})(x - x_k)Q(x)$$

avec $Q(x) = \alpha x + \beta$ polynôme de degré 1.
On a

$$P'(x) = (x - x_{k-1})Q(x) + (x - x_k)Q(x) + \alpha(x - x_{k-1})(x - x_k)$$

et

$$P''(x) = 2(Q(x) + \alpha(x - x_{k-1}) + \alpha(x - x_k).$$

Comme $P''(x_{k-1}) = P''(x_k) = 0$, on obtient

$$\begin{cases} Q(x_{k-1}) + \alpha(x_{k-1} - x_k) &= 0, \\ Q(x_k) + \alpha(x_k - x_{k-1}) &= 0, \end{cases} \iff \begin{cases} \alpha(x_{k-1} - h) + \beta &= 0, \\ \alpha(x_k + h) + \beta &= 0, \end{cases}$$

En soustrayant la première équation à la deuxième, on obtient $3\alpha h = 0$. Comme $h \neq 0$, on obtient $\alpha = \beta = 0$. D'où $Q \equiv 0$ et donc $P \equiv 0$.

2. On a $S_k''(x) = 6a_k(x_k - x) + 6b_k(x - x_{k-1})$. Pour déterminer a_k et b_k, on utilise les équations (4.52c) et (4.52d) qui deviennent respectivement $6ha_k = m_{k-1}$ et $6hba_k = m_k$. On obtient

$$a_k = \frac{m_{k-1}}{6h} \text{ et } b_k = \frac{m_k}{6h}.$$

Pour déterminer α_k et β_k on utilise les équations (4.52c) et (4.52d) qui deviennent respectivement

$$a_k h^3 + \alpha_k h = f_{k-1} \text{ et } b_k h^3 + \beta_k h = f_k.$$

En remplaçant a_k et b_k par leurs valeurs, on obtient

$$\alpha_k = \frac{f_{k-1}}{h} - \frac{h}{6}m_{k-1} \text{ et } \beta_k = \frac{f_k}{h} - \frac{h}{6}m_k.$$

Q. 3 1. On a par définition $\forall k \in [\![1,n]\!]$, $g(x) = S_k(x)$, $\forall x \in [x_{k-1}, x_k]$. Le problème de définition de g provient du fait que g est définie deux fois en x_k, $k \in [\![1, n-1]\!]$. En effet, on a

$$g(x_k) = S_k(x_k) \text{ et } g(x_k) = S_{k+1}(x_k).$$

Or par construction des S_k, on a $S_k(x_k) = S_{k+1}(x_k) = f_k$ et donc la fonction g est bien définie sur $[a,b]$.

2. Par construction, sur chaque intervalle $[x_{k-1}; x_k]$, la fonction g est polynomiale de degré inférieur ou égal à 3. Pour quelle soit un spline cubique, il reste à démontrer qu'elle est deux fois continûment différentiable sur $[a,b]$. Il suffit pour celà de vérifier qu'en chaque point x_k, $k \in [\![1, n-1]\!]$, la fonction g est continue et admet des dérivées premières et secondes.
La continuité est immédiate puisque $g(x_k) = f_k$. Pour les dérivées premières et secondes, il faut que leurs limites à gauche et à droite soient égales, c'est à dire

$$\forall k \in [\![1, n-1]\!],\ S'_k(x_k) = S'_{k+1}(x_k) \text{ et } S''_k(x_k) = S''_{k+1}(x_k).$$

Par construction des S_k, la seconde équation est immédiate : $S''_k(x_k) = S''_{k+1}(x_k) = m_k$. On a $S'_k(x) = -3a_k(x_k - x)^2 + 3b_k(x - x_{k-1})^2 - \alpha_k + \beta_k$ et donc

$$S'_k(x_k) = 3b_k h^2 - \alpha_k + \beta_k = \frac{h}{2}m_k + \frac{1}{h}(f_k - fk - 1) - \frac{h}{6}(m_k - m_{k-1})$$

De même, on obtient

$$S'_{k+1}(x_k) = -3a_{k+1}h^2 - \alpha_{k+1} + \beta_{k+1} = -\frac{h}{2}m_k + \frac{1}{h}(f_{k+1} - fk) - \frac{h}{6}(m_{k+1} - m_k)$$

Donc g sera dérivable en x_k si $S'_k(x_k) = S'_{k+1}(x_k)$, c'est à dire si

$$\frac{h}{2}m_k + \frac{1}{h}(f_k - fk - 1) - \frac{h}{6}(m_k - m_{k-1}) = -\frac{h}{2}m_k + \frac{1}{h}(f_{k+1} - fk) - \frac{h}{6}(m_{k+1} - m_k)$$

ce qui s'écrit encore, $\forall k \in [\![1, n-1]\!]$,

$$m_{k+1} + 4m_k + m_{k-1} = \frac{6}{h^2}(f_{k+1} - 2f_k + f_{k-1})$$

Q. 4 1. La condition $g''(a) = 0$ se traduit par $S''_1(x_0) = 0$ or par (4.52c) avec $k = 1$ on a $S''_1(x_0) = m_0$ d'où $m_0 = 0$.
La condition $g''(b) = 0$ se traduit par $S''_n(x_n) = 0$ or par (4.52d) avec $k = n$ on a $S''_n(x_n) = m_n$ d'où $m_n = 0$.
Pour déterminer les m_k, $k \in [\![0,n]\!]$, on a $n+1$ équations linéaires qui s'écrivent sous la forme matricielle $\mathbb{A}\boldsymbol{M} = \boldsymbol{b}$ avec

$$\mathbb{A} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 4 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & 4 & 1 \\ 0 & \dots & 0 & 0 & 1 \end{pmatrix} \text{ et } \boldsymbol{b} = \frac{6}{h^2}\begin{pmatrix} 0 \\ f_0 - 2f_1 + f_2 \\ \vdots \\ f_{n-2} - 2f_{n-1} + f_n \\ 0 \end{pmatrix}$$

2. La matrice $\mathbb{A}$ est à diagonale strictement dominante : elle est donc inversible.

◇

Chapitre 5

Intégration numérique

Soit f une fonction définie et intégrable sur un intervalle $[a, b]$ donné. On propose de chercher des approximations de

$$I = \int_a^b f(x)dx$$

dans le cas où l'on ne connait pas de primitive de f.

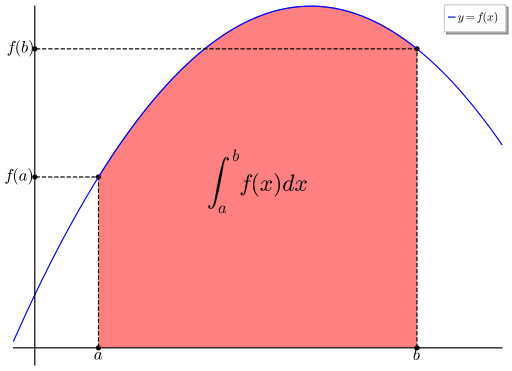


Figure 5.1: Représentation de $\int_a^b f(x)dx$ (aire de la surface colorée)

Definition 5.1

Soient $f \in \mathcal{C}^0([a, b]; \mathbb{R})$ et $\mathcal{Q}_n(f, a, b)$ la **formule de quadrature élémentaire** donnée par :

$$\mathcal{Q}_n(f, a, b) \overset{\text{def}}{=} (b - a) \sum_{j=0}^{n} w_j f(x_j) \tag{5.1}$$

avec $\forall j \in [\![0, n]\!]$ $w_j \in \mathbb{R}$ et $x_j \in [a, b]$ distincts deux à deux. L'erreur associée à cette formule de

quadrature, notée $\mathcal{E}_{a,b}(f)$, est définie par

$$\mathcal{E}_{a,b}(f) = \int_a^b f(x)dx - \mathcal{Q}_n(f,a,b), \quad \forall f \in \mathcal{C}^0([a,b];\mathbb{R}) \tag{5.2}$$

Definition 5.2

On dit qu'une formule d'intégration (ou formule de quadrature) est d'ordre p ou a pour **degré d'exactitude** p si elle est exacte pour les polynômes de degré inférieur ou égal à p.

5.1 Méthodes de quadrature élémentaires

On suppose que les points x_j de la formule de quadrature (5.1) sont deux à deux disticts.

5.1.1 Méthodes simplistes

On peut approcher f par un polynôme constant. Les trois formules usuels sont

Méthode du rectangle à gauche : En figure 5.2, on représente l'approximation de $\int_a^b f(x)dx$ lorsque f est approché par le polynôme constant $P(x) = f(a)$.

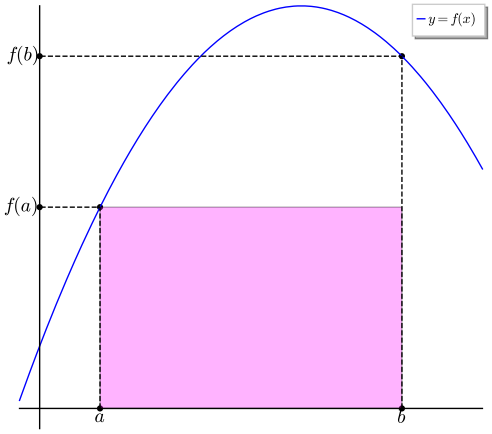


Figure 5.2: Formule du rectangle à gauche : $\int_a^b f(x)dx \approx (b-a)f(a)$ (aire de la surface colorée)

On a alors

$$\int_a^b f(x)dx \approx \mathcal{Q}_0(f,a,b) = (b-a)f(a), \text{ formule du rectangle (à gauche)}$$

et son degré d'exactitude est 0.

Méthode du rectangle à droite : En figure 5.3, on représente l'approximation de $\int_a^b f(x)dx$ lorsque f est approché par le polynôme constant $P(x) = f(b)$.

On a alors

$$\int_a^b f(x)dx \approx \mathcal{Q}_0(f,a,b) = (b-a)f(b), \text{ formule du rectangle (à droite)}$$

et son degré d'exactitude est 0.

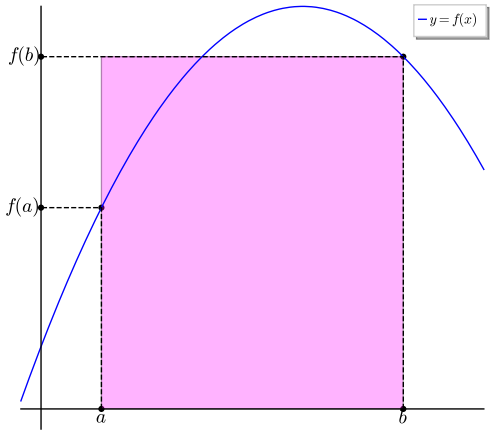


Figure 5.3: Formule du rectangle à droite : $\int_a^b f(x)dx \approx (b-a)f(b)$ (aire de la surface colorée)

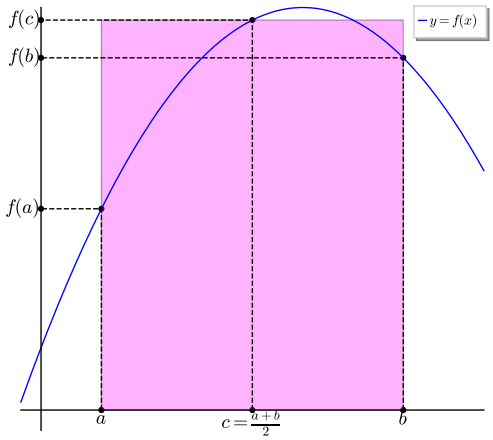


Figure 5.4: Formule du point milieu : $\int_a^b f(x)dx \approx (b-a)f(\frac{a+b}{2})$ (aire de la surface colorée)

Méthode du point milieu : En figure 5.4, on représente l'approximation de $\int_a^b f(x)dx$ lorsque f est approché par le polynôme constant $P(x) = f((a+b)/2)$. On a alors

$$\int_a^b f(x)dx \approx \mathcal{Q}_0(f,a,b) = (b-a)f\left(\frac{a+b}{2}\right), \text{ formule du point milieu}$$

et son degré d'exactitude est 1.

La précision de ces formules est toute relative!

Nous allons maintenant voir comment *généraliser* en approchant la fonction f par des polynômes d'interpolation de Lagrange de degré plus élevés.

5.1.2 Quelques résultats théoriques

Pour obtenir certains résultats associés à une formule de quadrature sur l'intervalle $[a;b]$, il serait souvent plus simple de les déterminer sur les intervalles d'intégration $[0;1]$ ou $[-1;1]$. La proposition suivante va nous permettre d'affirmer qu'étudier le degré d'exactitude d'une formule de quadrature sur l'intervalle $[a;b]$ est équivalent à l'étudier sur l'intervalle $[0;1]$ ou sur l'intervalle $[-1;1]$ moyennant respectivemment les changements de variables

$$x = \varphi(t) = a + (b-a)t \quad \text{ou} \quad x = \varphi(t) = \frac{a+b}{2} + \frac{b-a}{2}t$$

Proposition 5.3

Soit $\mathcal{Q}_n(f,a,b)$ definie en (5.1), une formule de quadrature élémentaire à $n+1$ points $(x_i)_{i\in[\![0,n]\!]}$ (distincts deux à deux dans $[a,b]$).
On note $x = \varphi(t) = \alpha + \beta t$, $\beta \in \mathbb{R}^*$, le changement de variable affine, $t_i = \varphi^{-1}(x_i)$, $\forall i \in [\![0,n]\!]$, et

$$\mathcal{Q}_n(g, \varphi^{-1}(a), \varphi^{-1}(b)) = \left(\varphi^{-1}(b) - \varphi^{-1}(a)\right) \sum_{i=0}^{n} w_i g(t_i). \tag{5.3}$$

Alors $\mathcal{Q}_n(f,a,b)$ est de degré d'exactitude k si et seulement si $\mathcal{Q}_n(g, \varphi^{-1}(a), \varphi^{-1}(b))$ est de degré d'exactitude k.

Preuve. On a $\varphi^{-1}(x) = \frac{x-\alpha}{\beta}$.

$\boxed{\Rightarrow}$ On suppose que $\mathcal{Q}_n(f,a,b)$ est de degré d'exactitude k. Soit $\mathrm{Q} \in \mathbb{R}_k[X]$. On a

$$\int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} \mathrm{Q}(t)dt = \frac{1}{\beta}\int_a^b \mathrm{Q}\circ\varphi^{-1}(x)dx.$$

Or φ^{-1} est un polynôme de degré 1 et $\mathrm{Q}\circ\varphi^{-1}$ est la composé de deux polynômes: c'est donc un polynôme de degré le produit des degrés des deux polynômes, i.e. $\mathrm{Q}\circ\varphi^{-1} \in \mathbb{R}_k[X]$. Comme $\mathcal{Q}_n(f,a,b)$ est de degré d'exactitude k, on en déduit que

$$\int_a^b \mathrm{Q}\circ\varphi^{-1}(x)dx = \mathcal{Q}_n(\mathrm{Q}\circ\varphi^{-1},a,b) = (b-a)\sum_{i=0}^{n} w_i \mathrm{Q}\circ\varphi^{-1}(x_i) = (b-a)\sum_{i=0}^{n} w_i \mathrm{Q}(t_i).$$

On a alors

$$\int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} \mathrm{Q}(t)dt = \frac{b-a}{\beta}\sum_{i=0}^{n} w_i \mathrm{Q}(t_i)$$

or

$$\varphi^{-1}(b) - \varphi^{-1}(a) = \frac{b-\alpha}{\beta} - \frac{a-\alpha}{\beta} = \frac{b-a}{\beta}.$$

On en conclu donc que $\mathcal{Q}_n(g, \varphi^{-1}(a), \varphi^{-1}(b))$ est de degré d'exactitude k.

$\boxed{\Leftarrow}$ On suppose que $\mathcal{Q}_n(g, \varphi^{-1}(a), \varphi^{-1}(b))$ est de degré d'exactitude k. Soit $\mathrm{P} \in \mathbb{R}_k[X]$. On a

$$\int_a^b \mathrm{P}(x)dx = \beta \int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} \mathrm{P} \circ \varphi(t)dt.$$

Or φ est un polynôme de degré 1 et $\mathrm{P} \circ \varphi^{-1}$ est la composé de deux polynômes: c'est donc un polynôme de degré le produit des degrés des deux polynômes, i.e. $\mathrm{P} \circ \varphi \in \mathbb{R}_k[X]$. Comme $\mathcal{Q}_n(g, \varphi^{-1}(a), \varphi^{-1}(b))$ est de degré d'exactitude k, on en déduit que

$$\begin{aligned}\int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} \mathrm{P} \circ \varphi(t)dt &= \mathcal{Q}_n(\mathrm{P} \circ \varphi, \varphi^{-1}(a), \varphi^{-1}(b)) \\ &= \big(\varphi^{-1}(b) - \varphi^{-1}(a)\big) \sum_{i=0}^{n} w_i \mathrm{P} \circ \varphi(t_i) \\ &= \big(\varphi^{-1}(b) - \varphi^{-1}(a)\big) \sum_{i=0}^{n} w_i \mathrm{P}(x_i).\end{aligned}$$

On a alors

$$\int_a^b \mathrm{P}(x)dx = \beta \big(\varphi^{-1}(b) - \varphi^{-1}(a)\big) \sum_{i=0}^{n} w_i \mathrm{P}(x_i)$$

et comme $\varphi^{-1}(b) - \varphi^{-1}(a) = \frac{b-a}{\beta}$, on en déduit que $\mathcal{Q}_n(f, a, b)$ est de degré d'exactitude k.

□

Proposition 5.4

Soit $\mathcal{Q}_n(f, a, b)$ definie en (5.1), une formule de quadrature élémentaire à $n+1$ points.
L'application $f \longmapsto \mathcal{Q}_n(f, a, b)$ définie de $f \in \mathcal{C}^0([a, b]; \mathbb{R})$, muni de la norme infini, à valeurs dans $\mathbb{R}$ est linéaire continue.

Preuve. On commence par démontrer la linéarité. Soient f et g dans $\in \mathcal{C}^0([a, b]; \mathbb{R})$, et λ et μ deux réels. Alors $\lambda f + \mu g \in \mathcal{C}^0([a, b]; \mathbb{R})$, et on a

$$\begin{aligned}\mathcal{Q}_n(\lambda f + \mu g, a, b) &= (b-a) \sum_{j=0}^{n} w_j (\lambda f + \mu g)(x_j) \\ &= (b-a) \sum_{j=0}^{n} w_j \big(\lambda f(x_j) + \mu g(x_j)\big) \\ &= \lambda (b-a) \sum_{j=0}^{n} w_j f(x_j) + \mu (b-a) \sum_{j=0}^{n} w_j g(x_j) \\ &= \lambda \mathcal{Q}_n(f, a, b) + \mu \mathcal{Q}_n(g, a, b).\end{aligned}$$

L'application $f \longmapsto \mathcal{Q}_n(f, a, b)$ est donc linéaire. Pour démontrer qu'elle est continue, il suffit alors de démontrer que

$$\exists C > 0, \quad \text{tel que } |\mathcal{Q}_n(f, a, b)| \leqslant C \|f\|_\infty, \ \forall f \in \mathcal{C}^0([a, b]; \mathbb{R}).$$

Or, on a, pour tout $f \in \mathcal{C}^0([a, b]; \mathbb{R})$,

$$\begin{aligned}|\mathcal{Q}_n(f, a, b)| &= |(b-a) \sum_{j=0}^{n} w_j f(x_j)| \\ &\leqslant (b-a) \sum_{j=0}^{n} |w_j||f(x_j)| \\ &\leqslant C \|f\|_\infty, \quad \text{avec} \quad C = (b-a) \sum_{j=0}^{n} |w_j| \text{ indépendant de } f.\end{aligned}$$

□

Proposition 5.5

La formule de quadrature élémentaire (5.1) à $n+1$ points est de degré d'exactitude k (au moins) si et seulement si

$$(b-a)\sum_{i=0}^{n} w_i x_i^r = \frac{b^{r+1}-a^{r+1}}{r+1}, \quad \forall r \in [\![0,k]\!]. \tag{5.4}$$

Preuve.

- $\boxed{\Rightarrow}$ Si la formule (5.1) est de degré d'exactitude k, elle est donc exacte pour tout polynôme de $\mathbb{R}_k[X]$ et plus particulièrement pour tous les monômes $1, X, X^2, \ldots, X^k$. Soit $r \in [\![0,k]\!]$. En prenant $f(x) = x^r$, la formule (5.1) étant exacte par hypothèse, on obtient

$$\mathcal{Q}_n(x \mapsto x^r, a, b) = (b-a)\sum_{i=0}^{n} w_i x_i^r \overset{\text{hyp}}{=} \int_a^b x^r dx = \frac{b^{r+1}-a^{r+1}}{r+1}$$

- $\boxed{\Leftarrow}$ On suppose que l'on a (5.4). Soit $\mathrm{P} \in \mathbb{R}_k[X]$. On va montrer que la formule de quadrature (5.1) est alors exacte.
Le polynôme P peut s'écrire comme combinaison linéaire des monômes de $\{1, X, X^2, \ldots, X^k\}$, base de $\mathbb{R}_k[X]$.
1ère démonstration: On a donc

$$\mathrm{P}(x) = \sum_{j=0}^{k} \alpha_j x^j, \ \forall x \in \mathbb{R}.$$

En prenant $f = \mathrm{P}$, la formule de quadrature (5.1) donne

$$\int_a^b \mathrm{P}(x)dx \approx (b-a)\sum_{i=0}^{n} w_i \mathrm{P}(x_i) = (b-a)\sum_{i=0}^{n} w_i \sum_{j=0}^{k} \alpha_j x_i^j$$

De plus, par linéarité de l'intégrale, on a

$$\int_a^b \mathrm{P}(x)dx = \sum_{j=0}^{k} \alpha_j \int_a^b x^j dx = \sum_{j=0}^{k} \alpha_j \frac{b^{j+1}-a^{j+1}}{j+1}$$

et en utilisant (5.4) on obtient

$$\int_a^b \mathrm{P}(x)dx = (b-a)\sum_{j=0}^{k} \alpha_j \sum_{i=0}^{n} w_i x_i^j$$

Ce qui donne

$$\int_a^b \mathrm{P}(x)dx = (b-a)\sum_{i=0}^{n} w_i \mathrm{P}(x_i).$$

La formule de quadrature est donc de degré d'exactitude k.
2ème démonstration: On a donc

$$\mathrm{P} = \sum_{j=0}^{k} \alpha_j X^j$$

et par linéarité de l'application $f \longmapsto \mathcal{Q}_n(f,a,b)$ (voir Proposition 5.4) on obtient

$$\begin{aligned}\mathcal{Q}_n(\mathrm{P},a,b) &= \mathcal{Q}_n(\sum_{j=0}^{k} \alpha_j X^j, a, b) \\ &= \sum_{j=0}^{k} \alpha_j \mathcal{Q}_n(X^j,a,b).\end{aligned}$$

Par hypothèse, on a (5.4) et, comme par définition X^j est le polynôme $x \mapsto x^j$, on obtient

$$\forall j \in [\![0,k]\!], \ \mathcal{Q}_n(X^j,a,b) \overset{\text{hyp}}{=} \int_a^b X^j(x)dx = \int_a^b x^j dx = \frac{b^{j+1}-a^{j+1}}{j+1}.$$

De plus, par linéarité de l'intégrale, on a

$$\int_a^b \mathrm{P}(x)dx = \sum_{j=0}^{k} \alpha_j \int_a^b x^j dx = \sum_{j=0}^{k} \alpha_j \frac{b^{j+1}-a^{j+1}}{j+1}$$

et donc

$$\int_a^b \mathrm{P}(x)dx = \sum_{j=0}^{k} \alpha_j \mathcal{Q}_n(X^j, a, b) = \mathcal{Q}_n(\mathrm{P}, a, b).$$

□

En appliquant la Proposition 5.5 avec $k = 0$, on obtient immédiatemment le corolaire suivant.

Corollaire 5.6

La formule de quadrature élémentaire (5.1) à $n+1$ points est de degré d'exactitude 0 au moins si et seulement si

$$\sum_{i=0}^{n} w_i = 1.$$

Proposition 5.7

Soient $(x_i)_{i\in[\![0,n]\!]}$ des points deux à deux distincts de l'intervalle $[a, b]$ donnés. Il existe alors une unique formule de quadrature élémentaire (5.1) à $n+1$ points de degré d'exactitude n au moins.

Preuve. En fixant les points $(x_i)_{i\in[\![0,n]\!]}$ deux à deux distincts, pour obtenir explicitement la formule de quadrature de type (5.1) il faut déterminer les $n+1$ poids $(w_i)_{i\in[\![0,n]\!]}$. Or, de (5.4), en prenant $k = n$, on obtient exactement $n+1$ équations linéaires en les (w_i) s'écrivant matriciellement sous la forme :

$$(b-a)\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix}\begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} b-a \\ \frac{b^2-a^2}{2} \\ \vdots \\ \frac{b^{n+1}-a^{n+1}}{n+1} \end{pmatrix}$$

La matrice intervenant dans le système précédent s'appelle **la matrice de Vandermonde** et elle est inversible (car les (x_i) sont deux à deux distincts, voir Exercice B.3.13). Ceci établi donc l'existence et l'unicité de poids $(w_i)_{i\in[\![0,n]\!]}$ tels que la formule de quadrature élémentaire (5.1) soit d'ordre (au moins) n.
Il est aussi possible de démontrer l'unicité classiquement. Supposons qu'il existe $(w_i)_{i\in[\![0,n]\!]}$ et $(\tilde{w}_i)_{i\in[\![0,n]\!]}$ tels que pour tout $\mathrm{P} \in \mathbb{R}_n[X]$, on ait

$$\int_a^b \mathrm{P}(x)dx = (b-a)\sum_{i=0}^{n} w_i \mathrm{P}(x_i) = (b-a)\sum_{i=0}^{n} \tilde{w}_i \mathrm{P}(x_i).$$

On a alors $\forall \mathrm{P} \in \mathbb{R}_n[X]$,

$$\sum_{i=0}^{n} (w_i - \tilde{w}_i)\mathrm{P}(x_i) = 0. \tag{5.5}$$

On rappelle que les fonctions de base de Lagrange associées aux $(n+1)$ points $(x_i)_{i\in[\![0,n]\!]}$ définies en (4.5), notées L_i, sont dans $\mathbb{R}_n[X]$ et vérifient

$$L_i(x_j) = \delta_{i,j}, \ \forall j \in [\![0, n]\!]$$

Soit $j \in [\![0, n]\!]$. En choisissant $\mathrm{P} = L_j$ dans (5.5), on obtient

$$0 = \sum_{i=0}^{n} (w_i - \tilde{w}_i)L_j(x_i) = (w_j - \tilde{w}_j)$$

ce qui prouve l'unicité. □

Exercice 5.1.1

Soient $(x_i)_{i=0}^n$ $(n+1)$ points donnés et distincts 2 à 2 d'un intervalle $[a,b]$ $(a<b)$. Ecrire une fonction algorithmique WeightsFromPoints permettant de déterminer les poids $(w_i)_{i=0}^n$ de telle sorte que la formule de quadrature élémentaire associée soit de degré d'exactitude n au moins en s'inspirant de résultats obtenus dans la démonstration de la Proposition 5.7. On pourra utiliser la fonction algorithmique $\boldsymbol{x} \leftarrow$ Solve$(\mathbb{A}, \boldsymbol{b})$ permettant de résoudre le système linéaire $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$.

Correction Exercice Nous avons vu, dans la Proposition 5.7, que pour avoir une formule de quadrature élémentaire de degré d'exactitude n, il est nécessaire et suffisant que les $(n+1)$ poids $(w_i)_{i=0}^n$ soient solution du système linéaire suivant:

$$(b-a)\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix}\begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} b-a \\ \frac{b^2-a^2}{2} \\ \vdots \\ \frac{b^{n+1}-a^{n+1}}{n+1} \end{pmatrix}$$

Algorithme 5.1 Fonction WeightsFromPoints retournant le tableau des poids $\boldsymbol{w}$ associé à un tableau de points $\boldsymbol{x}$ donnés (points 2 à 2 distincts) appartenant à un intervalle $[a,b]$.

Données : $\boldsymbol{x}$: tableau de $\mathbb{R}^{n+1}$ contenant $(n+1)$ points distincts deux à deux dans un intervalle $[a,b]$ avec la convention $\boldsymbol{x}(i) = x_{i-1}$, $\forall i \in [\![1, n+1]\!]$

$a,\ b$: deux réels, $a<b$.

Résultat : $\boldsymbol{w}$: vecteur de $\mathbb{R}^{n+1}$ avec $\boldsymbol{w}(i) = w_{i-1}$, $\forall i \in [\![1, n+1]\!]$

```
Fonction w ← WeightsFromPoints ( x, a, b )
    b ← O_{n+1}
    𝔸 ← 𝕆_{n+1,n+1}
    Pour i ← 1 à n + 1 faire
        Pour j ← 1 à n + 1 faire
            𝔸(i, j) ← x(j)^(i − 1)
        Fin Pour
        b(i) ← (b^i − a^i)/(i ∗ (b − a))
    Fin Pour
    w ← Solve(𝔸, b)
Fin Fonction
```

◇

Proposition 5.8: symétrie

Soit $\mathcal{Q}_n(f,a,b)$ definie en (5.1), une formule de quadrature élémentaire à $n+1$ points (distincts deux à deux et ordonnés). On dit qu'elle est **symétrique** si

$$\forall i \in [\![0,n]\!],\quad \frac{x_i + x_{n-i}}{2} = \frac{a+b}{2} \quad \text{et} \quad w_i = w_{n-i}. \tag{5.6}$$

Dans ce cas si cette formule est exacte pour les polynômes de degré $2m$ alors elle est nécessairement exacte pour les polynômes de degré $2m+1$.

Preuve. Soit $\mathrm{P} \in \mathbb{R}_{2m+1}[X]$. Il peut alors s'écrire sous la forme

$$\mathrm{P}(x) = C\left(x - \frac{a+b}{2}\right)^{2m+1} + \mathrm{R}(x)$$

avec C une constante réelle et $\mathrm{R} \in \mathbb{R}_{2m}[X]$. On a alors

$$\int_a^b \mathrm{P}(x)dx = C\int_a^b \left(x - \frac{a+b}{2}\right)^{2m+1} dx + \int_a^b \mathrm{R}(x)dx$$

et en appliquant la formule de quadrature au polynôme P on obtient

$$\sum_{i=0}^{n} w_i \mathrm{P}(x_i) = C\sum_{i=0}^{n} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} + \sum_{i=0}^{n} w_i \mathrm{R}(x_i)$$

On veut donc démontrer que

$$\int_a^b \mathrm{P}(x)dx = (b-a)\sum_{i=0}^{n} w_i \mathrm{P}(x_i)$$

c'est à dire

$$C\int_a^b \left(x - \frac{a+b}{2}\right)^{2m+1} dx + \int_a^b \mathrm{R}(x)dx = (b-a)C\sum_{i=0}^{n} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} + (b-a)\sum_{i=0}^{n} w_i \mathrm{R}(x_i)$$

Comme la formule de quadrature est supposée exacte pour les polynôme de degré $2m$, on a

$$\int_a^b \mathrm{R}(x)dx = (b-a)\sum_{i=0}^{n} w_i \mathrm{R}(x_i).$$

Il reste donc à démontrer que

$$\int_a^b \left(x - \frac{a+b}{2}\right)^{2m+1} dx = (b-a)\sum_{i=0}^{n} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1}.$$

Or en effectuant le changement de variable $t \mapsto \frac{a+b}{2} + t\frac{b-a}{2}$ on obtient

$$\int_a^b \left(x - \frac{a+b}{2}\right)^{2m+1} dx = \frac{b-a}{2}\int_{-1}^{1} t^{2m+1}dt = 0.$$

Des propriétés de symétrie de la formule, on déduit

$$x_i + x_{n-i} = a+b \quad \Leftrightarrow \quad x_i - \frac{a+b}{2} = -\left(x_{n-i} - \frac{a+b}{2}\right)$$

- Si $n = 2k$, (n paire), on a alors un nombre **impair** de points avec nécessairement $x_k = x_{n-k} = \frac{a+b}{2}$ et

$$\begin{aligned}
\sum_{i=0}^{n} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} &= \sum_{i=0}^{k-1} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} + 0 \times w_k + \sum_{i=k+1}^{2k} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} \\
&= \sum_{i=0}^{k-1} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} - \sum_{i=k+1}^{2k} w_{n-i} \left(x_{n-i} - \frac{a+b}{2}\right)^{2m+1} \\
&= \sum_{i=0}^{k-1} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} - \sum_{j=0}^{k-1} w_j \left(x_j - \frac{a+b}{2}\right)^{2m+1} \\
&= 0.
\end{aligned}$$

- Si $n = 2k-1$, (n impaire), on alors un nombre **pair** de points (avec $x_i \neq \frac{a+b}{2}$, $\forall i \in [\![0, n]\!]$) et

$$\begin{aligned}
\sum_{i=0}^{n} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} &= \sum_{i=0}^{k-1} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} + \sum_{i=k}^{2k-1} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} \\
&= \sum_{i=0}^{k-1} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} - \sum_{i=k}^{2k-1} w_{n-i} \left(x_{n-i} - \frac{a+b}{2}\right)^{2m+1} \\
&= \sum_{i=0}^{k-1} w_i \left(x_i - \frac{a+b}{2}\right)^{2m+1} - \sum_{j=0}^{k-1} w_j \left(x_j - \frac{a+b}{2}\right)^{2m+1} \\
&= 0.
\end{aligned}$$

□

Le résultat suivant fait le lien avec les polynômes d'interpolation de Lagrange pour établir une majoration de l'erreur associée à une formule de quadrature élémentaire.

Proposition 5.9

Soit $\mathcal{Q}_n(f,a,b)$ definie en (5.1), une formule de quadrature élémentaire à $n+1$ points $(x_i)_{i\in[\![0,n]\!]}$ (distincts deux à deux).
La formule de quadrature est de degré d'exactitude n au moins si et seulement si pour tout $i \in [\![0,n]\!]$, les poids w_i sont donnés par

$$w_i \overset{\text{def}}{=} \frac{1}{b-a}\int_a^b \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}dx = \int_0^1 \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-t_j}{t_i-t_j}dx, \quad \forall i \in [\![0,n]\!] \tag{5.7}$$

avec $t_i = (x_i-a)/(b-a)$.
Si $f \in \mathcal{C}^{n+1}([a,b];\mathbb{R})$ alors on a

$$|\mathcal{E}_{a,b}(f)| \leqslant \frac{1}{(n+1)!}\left\|f^{(n+1)}\right\|_\infty \int_a^b |\prod_{i=0}^{n}(x-x_i)|dx \tag{5.8}$$

Preuve. Montrons tout d'abord que l'on a l'égalité suivante

$$\frac{1}{b-a}\int_a^b \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}dx = \int_0^1 \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-t_j}{t_i-t_j}dx.$$

Par le changement de variables $s : t \longrightarrow a+(b-a)t$ on obtient

$$\int_a^b L_i(x)dx = (b-a)\int_0^1 L_i \circ s(t)dt$$

et l'on a $x_i = s(t_i) = a+(b-a)t_i$ où $t_i = (x_i-a)/(b-a)$. On en déduit

$$\begin{aligned}\int_0^1 L_i\circ s(t)dt &= \int_0^1 \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{s(t)-s(t_j)}{s(t_i)-s(t_j)}dt = \int_0^1 \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{(b-a)(t-t_j)}{(b-a)(t_i-t_j)}dt\\ &= \int_0^1 \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-t_j}{t_i-t_j}dt\end{aligned}$$

L'égalité est donc démontré.

Pour démontrer la proposition, on note $\mathcal{L}_n(f)$ le polynôme d'interpolation de Lagrange associés aux points $(x_i, f(x_i))_{i\in[\![0,n]\!]}$:

$$\mathcal{L}_n(f)(x) = \sum_{i=0}^{n} L_i(x)f(x_i) \text{ avec } L_i(x) = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}$$

On a alors

$$\int_a^b \mathcal{L}_n(f)(x)dx = \sum_{i=0}^{n} f(x_i)\int_a^b L_i(x)dx.$$

$\boxed{\Rightarrow}$ On suppose que la formule de quadrature est de degré d'exactitude n. Soit $i \in [\![0,n]\!]$, $L_i \in \mathbb{R}_n[X]$ et on a alors par hypothèse

$$\mathcal{Q}_n(L_i,a,b) \overset{\text{hyp}}{=} \int_a^b L_i(x)dx.$$

Or comme $L_i(x_j) = \delta_{i,j}$ on a

$$\mathcal{Q}_n(L_i, a, b) = (b-a)\sum_{j=0}^{n} w_j L_i(x_j) = (b-a)w_i.$$

On obtient donc

$$w_i = \frac{1}{b-a}\int_a^b L_i(x)dx.$$

$\boxed{\Leftarrow}$ On suppose les poids $(w_i)_{i=0}^n$ donnés par (5.7). La formule de quadrature s'écrit alors

$$\mathcal{Q}_n(f, a, b) \overset{\text{hyp}}{=} \sum_{i=0}^{n} f(x_i)\int_a^b L_i(x)dx.$$

Soit $\mathrm{P} \in \mathbb{R}_n[X]$. Par unicité du polynôme d'interpolation de Lagrange, on a $\mathrm{P} = \mathcal{L}_n(P)$ et

$$\begin{aligned}\int_a^b \mathrm{P}(x)dx &= \int_a^b \mathcal{L}_n(P)(x)dx \\ &= \int_a^b \sum_{i=0}^{n} L_i(x)\mathrm{P}(x_i)dx \\ &= \sum_{i=0}^{n} \mathrm{P}(x_i)\int_a^b L_i(x)dx \\ &= (b-a)\sum_{i=0}^{n} w_i \mathrm{P}(x_i) = \mathcal{Q}_n(\mathrm{P}, a, b).\end{aligned}$$

La formule de quadrature est donc exacte pour tout les polynômes de degré n au moins.

Pour démontrer l'inégalité (5.8), comme $f \in \mathcal{C}^{n+1}([a,b];\mathbb{R})$, on peut appliquer le théorème 4.4 pour obtenir

$$\forall x \in [a,b],\ \exists \xi_x \in [a,b],\quad f(x) - \mathcal{L}_n(f)(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{i=0}^{n}(x - x_i)$$

On en déduit

$$\begin{aligned}|f(x) - \mathcal{L}_n(f)(x)| &= \left|\frac{f^{(n+1)}(\xi_x)}{(n+1)!}\pi_n(x)\right| \\ &\leqslant \frac{\left\|f^{(n+1)}\right\|_\infty}{(n+1)!}|\pi_n(x)|\end{aligned}$$

L'application $f - \mathcal{L}_n(f)$ étant intégrable sur $[a,b]$, l'application $|f - \mathcal{L}_n(f)|$ l'est aussi. De même $|\pi_n(x)|$ est intégrable sur $[a,b]$. On obtient alors

$$\int_a^b |f(x) - \mathcal{L}_n(f)(x)|dx \leqslant \frac{\left\|f^{(n+1)}\right\|_\infty}{(n+1)!}\int_a^b |\pi_n(x)|dx.$$

De plus

$$\left|\int_a^b f(x) - \mathcal{L}_n(f)(x)dx\right| \leqslant \int_a^b |f(x) - \mathcal{L}_n(f)(x)|dx$$

ce qui donne

$$\left|\int_a^b f(x)dx - \int_a^b \mathcal{L}_n(f)(x)dx\right| \leqslant \frac{\left\|f^{(n+1)}\right\|_\infty}{(n+1)!}\int_a^b |\prod_{i=0}^{n}(x - x_i)|dx.$$

La formule de quadrature est de degré d'exactitude n au moins et le polynôme d'interpolation de Lagrange $\mathcal{L}_n(f)$ est de degré n donc on a

$$\begin{aligned}\int_a^b \mathcal{L}_n(f)(x)dx &= \mathcal{Q}_n(\mathcal{L}_n(f),a,b)\\ &= (b-a)\sum_{i=0}^{n} w_i \mathcal{L}_n(f)(x_i)\\ &= (b-a)\sum_{i=0}^{n} w_i f(x_i) \text{ car } \mathcal{L}_n(f)(x_i)=f(x_i)\\ &= \mathcal{Q}_n(f,a,b)\end{aligned}$$

ce qui donne

$$\begin{aligned}|\mathcal{E}_{a,b}(f)| &= \left|\int_a^b f(x)dx - \mathcal{Q}_n(f,a,b)\right|\\ &= \left|\int_a^b f(x)dx - \int_a^b \mathcal{L}_n(f)(x)dx\right|\\ &\leqslant \frac{\left\|f^{(n+1)}\right\|_\infty}{(n+1)!}\int_a^b |\prod_{i=0}^{n}(x-x_i)|dx.\end{aligned}$$

□

Exercice 5.1.2

Soient $(x_i)_{i=0}^n$ des points distincts 2 à 2 de l'intervalle $[a,b]$ vérifiant

$$\forall i \in [\![0,n]\!],\quad \frac{x_i + x_{n-i}}{2} = \frac{a+b}{2}.$$

On note $\mathrm{L}_i \in \mathbb{R}_n[X]$, $i \in [\![0,n]\!]$ les $(n+1)$ polynômes de base de Lagrange définis par

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

et vérifiant $L_i(x_j) = \delta_{i,j}$, $\forall (i,j) \in [\![0,n]\!]^2$

Q. 1 *Soit $i \in [\![0,n]\!]$. Montrer que*

$$\forall x \in \mathbb{R},\ \mathrm{L}_i((a+b)-x) = \mathrm{L}_{n-i}(x).$$

Q. 2 *Soient $(w_i)_{i=0}^n$ définis par*

$$w_i = \frac{1}{b-a}\int_a^b \mathrm{L}_i(t)dt,\quad \forall i \in [\![0,n]\!]$$

Montrer que l'on a alors

$$\forall i \in [\![0,n]\!],\quad w_i = w_{n-i}$$

Correction Exercice

Q. 1 Soit $i \in [\![0,n]\!]$. On note $\varphi(x) = (a+b) - x$ le polynôme de degré 1 et $\mathrm{P} = \mathrm{L}_i \circ \varphi$ le polynôme de $\mathbb{R}_n[X]$ (la composé de 2 polynômes est de degré le produit des degrés des 2 polynômes).

On a

$$\forall j \in [\![0,n]\!],\quad x_{n-j} = (a+b) - x_j$$

et

$$\mathrm{P}(x_j) = \mathrm{L}_i((a+b)-x_j) = \mathrm{L}_i(x_{n-j}) = \delta_{i,n-j} = \begin{cases} 1, & \text{si } i = n-j \\ 0, & \text{sinon} \end{cases} = \begin{cases} 1, & \text{si } j = n-i \\ 0, & \text{sinon} \end{cases} = \delta_{n-i,j}.$$

C'est à dire

$$\forall j \in [\![0,n]\!],\ \mathrm{P}(x_j) = \delta_{n-i,j}.$$

Or L_{n-i} est l'unique polynôme de $\mathbb{R}_n[X]$ vérifiant la relation précédente dont $\mathrm{P} = \mathrm{L}_{n-i}$ (voir Exercice 4.1.1, page 126).

Q. 2 Soit $i \in [\![0,n]\!]$. On note $t = \varphi(x) = (a+b) - x$ le changement de variable affine. On a alors $\varphi^{-1}(t) = (a+b) - t$ et

$$\begin{aligned} w_i &= \frac{1}{b-a}\int_a^b \mathrm{L}_i(t)dt \\ &= \frac{1}{b-a}\int_{\varphi^{-1}(a)}^{\varphi^{-1}(b)} \mathrm{L}_i \circ \varphi(x)\varphi'(x)dt \\ &= \frac{1}{b-a}\int_b^a \mathrm{L}_i((a+b)-x)(-1)dx \\ &= \frac{1}{b-a}\int_a^b \mathrm{L}_i((a+b)-x)dx \\ &= \frac{1}{b-a}\int_a^b \mathrm{L}_{n-i}(x)dx \quad \text{d'après la question précédente} \\ &= w_{n-i}. \end{aligned}$$

◇

De l'Exercice 5.1.2 et de la Proposition 5.8, on obtient le lemme suivant

Lemme 5.10

Soient $(x_i)_{i=0}^n$ des points distincts 2 à 2 de l'intervalle $[a,b]$ vérifiant

$$\forall i \in [\![0,n]\!],\quad \frac{x_i + x_{n-i}}{2} = \frac{a+b}{2}.$$

Soient $(w_i)_{i=0}^n$ définis par

$$w_i = \frac{1}{b-a}\int_a^b \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}dx,\quad \forall i \in [\![0,n]\!]$$

On a alors

$$\forall i \in [\![0,n]\!],\quad w_i = w_{n-i}$$

et la formule de quadrature élémentaire associée est de degré d'exactitude au moins n si n est impaire et au moins $n+1$ sinon.

Proposition 5.11: Degré maximal d'exactitude

Soit $\mathcal{Q}_n(f,a,b)$ défini par (5.1) une formule de quadrature élémentaire de degré d'exactitude au moins n. Elle est alors de degré d'exactitude $n+m$, $m \in \mathbb{N}^*$, au moins si et seulement si

$$\int_a^b \pi_n(x)\mathrm{Q}(x)dx = 0,\ \forall \mathrm{Q} \in \mathbb{R}_{m-1}[X] \tag{5.9}$$

où π_n est le polynôme de degré $n+1$ défini par

$$\pi_n(x) = \prod_{i=0}^{n}(x - x_i). \tag{5.10}$$

Le degré maximal d'exactitude d'une formule de quadrature élémentaire à $n+1$ points est $2n+1$. De plus, on a

$$(5.9) \iff \int_a^b \pi_n(x)x^k dx = 0, \ \forall k \in [\![0, m-1]\!]. \tag{5.11}$$

Preuve. On a par définition

$$\mathcal{Q}_n(f, a, b) = (b-a)\sum_{j=1}^{n} w_j f(x_j)$$

où les $n+1$ points x_j sont distincts deux à deux dans l'intervalle $[a, b]$.

La formule de quadrature est exacte pour les polynômes de degré $n+m$ si et seulement si

$$\int_a^b \mathrm{P}(x)dx = \mathcal{Q}_n(\mathrm{P}, a, b), \ \forall \mathrm{P} \in \mathbb{R}_{n+m}[X]$$

Soit $\mathrm{P} \in \mathbb{R}_{n+m}[X]$, on peut effectuer la division euclidienne de P par $\pi_n \in \mathbb{R}_{n+1}[X]$. Il existe donc $\mathrm{Q} \in \mathbb{R}_{m-1}[X]$ (quotient) et $\mathrm{R} \in \mathbb{R}_n[X]$ (reste) tels que

$$\mathrm{P} = \mathrm{Q}\pi_n + \mathrm{R}.$$

On a alors par linéarite de l'intégrale

$$\int_a^b \mathrm{P}(x)dx = \int_a^b \mathrm{Q}(x)\pi_n(x)dx + \int_a^b \mathrm{R}(x)dx$$

et par linéarite de $\mathcal{Q}_n$

$$\mathcal{Q}_n(\mathrm{P}, a, b) = \mathcal{Q}_n(\mathrm{Q}\pi_n, a, b) + \mathcal{Q}_n(\mathrm{R}, a, b).$$

Par hypothèse, la formule de quadrature a pour degré d'exactitude n et comme $\mathrm{R} \in \mathbb{R}_n[X]$ on obtient

$$\int_a^b \mathrm{R}(x)dx = \mathcal{Q}_n(\mathrm{R}, a, b).$$

On en déduit alors que

$$\int_a^b \mathrm{P}(x)dx - \mathcal{Q}_n(\mathrm{P}, a, b) = \int_a^b \mathrm{Q}(x)\pi_n(x)dx - \mathcal{Q}_n(\mathrm{Q}\pi_n, a, b).$$

Par construction $\pi_n(x_j) = 0$, $\forall j \in [\![0, n]\!]$, ce qui donne

$$\mathcal{Q}_n(\mathrm{Q}\pi_n, a, b) = (b-a)\sum_{j=0}^{n} w_j \mathrm{Q}(x_j)\pi_n(x_j) = 0$$

et donc

$$\int_a^b \mathrm{P}(x)dx - \mathcal{Q}_n(\mathrm{P}, a, b) = \int_a^b \mathrm{Q}(x)\pi_n(x)dx. \tag{5.12}$$

$\boxed{\Leftarrow}$ si (5.9) est vérifié alors,

$$\int_a^b \mathrm{P}(x)dx - \mathcal{Q}_n(\mathrm{P}, a, b) = 0.$$

La formule de quadrature est donc de degré d'exactitude $n+m$.

$\boxed{\Rightarrow}$ si la formule de quadrature est de degré d'exactitude $n+m$ alors pour tout $\mathrm{Q} \in \mathbb{R}_{m-1}[X]$, le polynôme $\mathrm{P} = \mathrm{Q}\pi_n \in \mathbb{R}_{n+m}[X]$ et donc

$$\int_a^b \mathrm{P}(x)dx - \mathcal{Q}_n(\mathrm{P}, a, b) = 0.$$

En utilisant (5.12), on obtient alors

$$\int_a^b \mathrm{Q}(x)\pi_n(x)dx = 0.$$

- La plus grande valeur que puisse prendre m est $n+1$ (i.e. $m+n=2n+1$). En effet si $m=n+2$ alors $\mathrm{Q}\in\mathbb{R}_{n+1}[X]$ dans (5.9). Or $\pi_n\in\mathbb{R}_{n+1}[X]$ et en prenant $\mathrm{Q}=\pi_n$ on obtient

$$\int_a^b \pi_n(x)\mathrm{Q}(x)dx=\int_a^b \mathrm{Q}^2(x)dx>0.$$

- Dans l'équivalence (5.11), $\boxed{\Rightarrow}$ est immédiat car $x\mapsto x^k\in\mathbb{R}_{m-1}[X]$.
Pour établir l'implication $\boxed{\Leftarrow}$, on utilise la linéarite de l'intégrale. En effet, soit $\mathrm{P}\in\mathbb{R}_{m-1}[X]$, il existe $(\alpha_0,\dots,\alpha_{m-1})\in\mathbb{R}^m$ tel que

$$\forall x\in\mathbb{R},\quad \mathrm{P}(x)=\sum_{k=0}^{m-1}\alpha_k x^k.$$

On a alors

$$\begin{aligned}\int_a^b \mathrm{P}(x)\pi_n(x)dx&=\int_a^b \pi_n(x)\sum_{k=0}^{m-1}\alpha_k x^k dx\\&=\sum_{k=0}^{m-1}\alpha_k\int_a^b \pi_n(x)x^k dx\\&\overset{\text{hyp}}{=}\sum_{k=0}^{m-1}\alpha_k\times 0=0.\end{aligned}$$

□

Cette proposition sera utilisée pour déterminer les points de quadrature de la méthode de Gauss-Legendre. Une fois connu les points de quadrature x_i, les poids w_i pourront être calculés par (5.7).

Si l'on choisi comme points de quadrature les $n+1$ points de la discrétisation régulière de l'intervalle $[a,b]$, on obtient la méthode de Newton-Cotes.

Bien d'autres méthodes peuvent être obtenues (avec d'autres points), certaines permettant le calcul d'intégrales avec poids de la forme $\int_a^b w(x)f(x)dx$:

- méthode de Newton-Cotes ouvertes,
- méthode de Gauss-Legendre,
- méthode de Gauss-Jacobi,
- méthode de Gauss-Tchebychev,
- méthode de Gauss-Laguerre,
- méthode de Gauss-Lobatto,
- méthode de Romberg...

5.1.3 Formules élémentaires de Newton-Cotes

Soit $(x_i)_{i\in[\![0,n]\!]}$ une discrétisation régulière de l'intervalle $[a,b]$:

$$\forall i\in[\![0,n]\!],\quad x_i=a+ih \text{ avec } h=(b-a)/n.$$

On a alors

$$\forall i\in[\![0,n]\!],\quad \frac{x_i+x_{n-i}}{2}=\frac{a+b}{2}.$$

Proposition 5.12

Soient $f\in\mathcal{C}^0([a,b];\mathbb{R})$ et $(x_i)_{i\in[\![0,n]\!]}$ une discrétisation régulière de l'intervalle $[a,b]$: $x_i=a+ih$ avec $h=(b-a)/n$.

Les formules de quadrature élémentaires de Newton-Cotes s'écrivent sous la forme

$$\int_a^b f(x)dx \approx (b-a)\sum_{i=0}^{n} w_i f(x_i)$$

où les poids $(w_i)_{i=0}^n$ sont donnés par (5.7).
Elles sont symétriques et leur degré d'exactitude (d.e. dans le tableau suivant) est égal à n si n est impair et à $n+1$ sinon.

n	**d.e.**	w_i (poids)									**nom**
1	1	$\frac{1}{2}$	$\frac{1}{2}$								trapèze
2	3	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$							Simpson
3	3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$						Newton
4	5	$\frac{7}{90}$	$\frac{16}{45}$	$\frac{2}{15}$	$\frac{16}{45}$	$\frac{7}{90}$					Villarceau
5	5	$\frac{19}{288}$	$\frac{25}{96}$	$\frac{25}{144}$	$\frac{25}{144}$	$\frac{25}{96}$	$\frac{19}{288}$				?
6	7	$\frac{41}{840}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{34}{105}$	$\frac{9}{280}$	$\frac{9}{35}$	$\frac{41}{840}$			Weddle
7	7	$\frac{751}{17280}$	$\frac{3577}{17280}$	$\frac{49}{640}$	$\frac{2989}{17280}$	$\frac{2989}{17280}$	$\frac{49}{640}$	$\frac{3577}{17280}$	$\frac{751}{17280}$		?
8	9	$\frac{989}{28350}$	$\frac{2944}{14175}$	$-\frac{464}{14175}$	$\frac{5248}{14175}$	$-\frac{454}{2835}$	$\frac{5248}{14175}$	$-\frac{464}{14175}$	$\frac{2944}{14175}$	$\frac{989}{28350}$	?

Table 5.1: Méthodes de Newton-Cotes

Par exemple, la formule de Simpson ($n=2$) est

$$\int_a^b f(x)dx \approx \frac{b-a}{6}\left(f(a)+4f(\frac{a+b}{2})+f(b)\right) \tag{5.13}$$

Pour un n donné, déterminer les coefficients w_i de la formule de Newton-Cotes est assez simple.

Démonstration 1: En effet, il suffit de remarquer que la formule doit être exacte si f est un polynôme de degré au plus n. Ensuite par linéarité de la formule, on est ramené à résoudre une système linéaire à $n+1$ inconnues (les $(w_i)_{i\in[\![0,n]\!]}$) en écrivant que pour chaque monôme

$$(b-a)\sum_{i=0}^{n} w_i f(x_i) = \int_a^b f(x)dx,\ \forall f \in \{1, X, X^2, \ldots, X^n\} \subset \mathbb{R}_n[X]. \tag{5.14}$$

Par exemple, pour $n=2$, on a $b=a+2h$. A partir de (5.14) on obtient les trois équations :

$$\begin{cases} w_0+w_1+w_2 &= 1, & (f(x)=1)\\ aw_0+(a+h)w_1+(a+2h)w_2 &= \frac{1}{b-a}\int_a^b x dx = a+h, & (f(x)=x)\\ a^2w_0+(a+h)^2w_1+(a+2h)^2w_2 &= \frac{1}{b-a}\int_a^b x^2dx = \frac{1}{3}(3a^2+6ah+4h^2), & (f(x)=x^2). \end{cases}$$

On a alors

$$\begin{cases} w_0 = -w_1-w_2+1,\\ a(-w_1-w_2+1)+(a+h)w_1+(a+2h)w_2 = a+h,\\ a^2(-w_1-w_2+1)+(a+h)^2w_1+(a+2h)^2w_2 = a^2+2ah+\frac{4}{3}h^2. \end{cases}$$

c'est à dire

$$\begin{cases} w_0 = -w_1-w_2+1,\\ w_1+2w_2 = 1,\\ 2a(w_1+2w_2)+h(w_1+4w_2) = 2a+4h/3, \end{cases}$$

Par substitution, de la deuxième équation dans la troisième, on obtient enfin

$$\begin{cases} w_0 = -w_1-w_2+1,\\ w_1+2w_2 = 1,\\ w_1+4w_2 = 4/3, \end{cases}$$

ce qui donne $w_0=w_2=\frac{1}{6}$ et $w_1=\frac{2}{3}$.

Démonstration 2: En utilisant la Proposition 5.9, on a

$$w_i \overset{\text{def}}{=} \frac{1}{b-a}\int_a^b \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{x-x_j}{x_i-x_j}dx = \int_0^1 \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-t_j}{t_i-t_j}dx, \quad \forall i \in [\![0,n]\!]$$

On a donc dans le cadre des points équidistants sur $[0,1]$, $t_i = i/n$, $\forall i \in [\![0,n]\!]$ et

$$\prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-t_j}{t_i-t_j} = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j/n}{(i-j)/n} = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{nt-j}{i-j}$$

Par exemple, pour $n=2$, on a

$$\begin{aligned} L_0(t) &= \frac{2t-1}{-1}\frac{2t-2}{-2} = (2\,t-1)(t-1) \\ L_1(t) &= \frac{2t}{1}\frac{2t-2}{-1} = -4\,(t-1)t \\ L_2(t) &= \frac{2t}{2}\frac{2t-1}{1} = (2\,t-1)t \end{aligned}$$

et

$$w_0 = \int_0^1 L_0(t) = \frac{1}{6}, \quad w_1 = \int_0^1 L_1(t) = \frac{2}{3}, \quad w_2 = \int_0^1 L_2(t) = \frac{1}{6}$$

On sait par le Lemme 5.10 que, $n=2$ étant paire, la formule de quadrature élémentaire de Simpson est au moins de degré d'exactitude 3. Pour montrer qu'elle est de degré d'exactitude 3, il suffit de vérifier qu'elle n'est plus exacte pour le monôme x^4 sur l'intervalle $[0,1]$.
En effet, on a

$$\frac{1}{6}\left(0^4 + 4(\frac{1}{2})^4 + 1^4\right) = \frac{5}{24} \neq \int_0^1 x^4 dx = \frac{1}{5}$$

Exercice 5.1.3

Q. 1 *Ecrire une fonction algorithmique* WeightsPointsNC *retournant les* $(n+1)$ *points et les* $(n+1)$ *poids de la formule de quadrature élémentaire de Newton-Cotes à* $(n+1)$ *points.*

Q. 2 *Ecrire une fonction algorithmique* QuadElemNC *retournant la valeur de* $\mathcal{Q}_n(f,a,b)$ *correspondant à la formule de quadrature élémentaire de Newton-Cotes à* $(n+1)$ *points.*

Correction Exercice

Algorithme 5.2 Fonction WeightsPointsNC retournant le tableau de points $\boldsymbol{x}$ donnés correspondant à la discrétisation régulière intervalle $[a,b]$. et le tableau des poids $\boldsymbol{w}$ associé à un

Données : $a,\ b$: deux réels, $a<b$,
n : $n \in \mathbb{N}^*$.
Résultat : $\boldsymbol{x}$: vecteur de $\mathbb{R}^{n+1}$ avec $\boldsymbol{x}(i) = x_{i-1}$, $\forall i \in [\![1,n+1]\!]$ et $x_{i-1} = a + (i-1)h$, $h = (b-a)/n$,
$\boldsymbol{w}$: vecteur de $\mathbb{R}^{n+1}$ avec $\boldsymbol{w}(i) = w_{i-1}$, $\forall i \in [\![1,n+1]\!]$

Q. 1
1: **Fonction** $[\boldsymbol{x},\boldsymbol{w}] \leftarrow$ WeightsPointsNC (a,b,n)
2: $\boldsymbol{x} \leftarrow a : (b-a)/n : b$
3: $\boldsymbol{w} \leftarrow$ WeightsFromPoints$(\boldsymbol{x},a,b)$
4: **Fin Fonction**

Q. 2 On a de manière générique l'algorithme suivant:

Algorithme 5.3 Fonction QuadElemGen retourne la valeur de $I = (b-a)\sum_{j=0}^{n} w_j f(x_j)$.

Données : f : une fonction définie de $[a,b]$ dans $\mathbb{R}$,
a, b : deux réels avec $a < b$
$\boldsymbol{x}$: vecteur de $\mathbb{R}^{n+1}$ contenant $(n+1)$ points distincts deux à deux dans un intervalle $[a,b]$ avec la convention $\boldsymbol{x}(i) = x_{i-1}$, $\forall i \in [\![1, n+1]\!]$
$\boldsymbol{w}$: vecteur de $\mathbb{R}^{n+1}$ tel que $\boldsymbol{w}(i) = w_{i-1}$, $\forall i \in [\![1, n+1]\!]$
Résultat : I : un réel

```
1: Fonction I ← QuadElemGen ( f, a, b, x, w )
2:   I ← 0
3:   Pour i ← 1 à Length(x) faire
4:     I ← I + w(i) * f(x(i))
5:   Fin Pour
6:   I ← (b − a) * I
7: Fin Fonction
```

On peut noter que si l'on dispose de la fonction $s \leftarrow \text{Dot}(\boldsymbol{u}, \boldsymbol{v})$ correpondant au produit scalaire de deux vecteurs du même espace alors on a directement

$$I \leftarrow (b-a) * \text{Dot}(\boldsymbol{w}, f(\boldsymbol{x})).$$

Algorithme 5.4 Fonction QuadElemGen retourne la valeur de $I = (b-a)\sum_{j=0}^{n} w_j f(x_j)$ où les poids w_i et les points x_i sont ceux définis par la formule de quadrature élémentaire de Newton-Cotes

Données : f : une fonction définie de $[a,b]$ dans $\mathbb{R}$,
a, b : deux réels avec $a < b$,
n : $n \in \mathbb{N}^*$
Résultat : I : un réel

```
1: Fonction I ← QuadElemNC ( f, a, b, n )
2:   [x, w] ← WeightsPointsNC(a, b, n)
3:   I ← QuadElemGen(f, a, b, x, w)
4: Fin Fonction
```

◇

Remarque 5.13 Pour les méthode de Newton-Cotes, il ne faut pas trop "monter" en ordre car le phénomène de Runge (forte oscillation possible du polynôme d'interpolation sur les bords de l'intervalle) peut conduire à de très grande erreurs. Au delà de $n = 7$, des poids négatifs apparaissent dans les formules et les rendent beaucoup plus sensibles aux erreurs d'arrondis.

Sur un exemple très simple, il est possible d'illustrer ce phénomène. Soit $f(x) = 3x+2$. On a démontré que toutes les formules de Newton-Cotes à $n+1$ points, $n \geqslant 1$, sont exactes pour le calcul de $\int_a^b f(x)dx$ car f est un polynôme de degré 1. De plus les poids $(w_i)_{i\in[\![0,n]\!]}$ peuvent être calculés sous forme fractionnaire : il est immédiat à partir de la formule (5.7) que $w_i \in \mathbb{Q}$.
En choissant, par exemple, $a = 0$ et $b = 1$ les $n+1$ points x_i de la formule de Newton-Cotes sont aussi des nombres rationnels. La fonction f étant polynomiale, on obtient $f(x_i) \in \mathbb{Q}$. On en déduit que la formule de Newton-Cotes à $n+1$ points donne dans ce cas un résultat (exacte) sous forme fractionnaire. Numériquement, les poids $w_i \in \mathbb{Q}$ et les points $x_i \in \mathbb{Q}$ vont être approchés en commettant de très petites erreurs. Sous `Sage`, logiciel gratuit de calcul formel (entre autres), on peut programmer à la fois le calcul des méthodes de Newton-Cotes exactes (en restant dans le corps $\mathbb{Q}$) et le calcul "approché" correspondant aux méthodes de Newton-Cotes *approchées* (i.e. les w_i et les x_i sont approchés avant le calcul de la somme). On représente en Figure 5.5 les erreurs obtenues en fonction de n par ces deux approches. Ces courbes en échelle logarithmique en ordonnées ont été obtenues en ajoutant aux erreurs le nombre $1e-16$ (pour éviter de prendre le log de zéro).

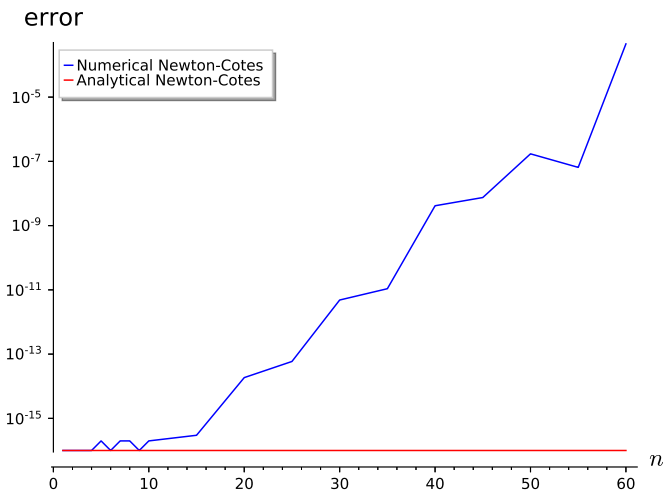


Figure 5.5: Instabilité des méthodes de Newton-Cotes élémentaires

Pour pallier ce problème, on étudiera les méthodes de quadrature composées.

5.1.4 Formules élémentaires de Gauss-Legendre

Exercice 5.1.4

Q. 1 *Déterminer les points* t_0, t_1 *de l'intervalle* $[-1,1]$ *et les poids* w_0, w_1 *tel que la formule de quadrature*

$$\int_{-1}^{1} g(t)dt \approx 2\sum_{i=0}^{1} w_i g(t_i)$$

soit de degré d'exactitude 3.

Q. 2 *En déduire une formule de quadrature pour le calcul de* $\int_a^b f(x)dx$ *qui soit de degré d'exactitude* 3.

Correction Exercice

Q. 1 D'après la Proposition 5.9, si t_0 et t_1 sont distincts et dans l'intervalle $[-1,1]$ alors avec

$$w_0 = \frac{1}{2}\int_{-1}^{1} \frac{t-t_1}{t_0-t_1}dt \text{ et } w_1 = \frac{1}{2}\int_{-1}^{1} \frac{t-t_0}{t_1-t_0}dt$$

la formule de quadrature est de degré d'exactitude 1.

Pour déterminer les points t_0 et t_1, on utilise la Proposition 5.11 (degré maximal d'exactitude) avec $m = n+1 = 2$: la formule de quadrature est de degré $2n+1 = 3$ ssi

$$\int_{-1}^{1} \pi_1(t)Q(t)dt = 0, \forall Q \in \mathbb{R}_1[X]$$

avec $\pi_1(t) = (t-t_0)(t-t_1)$. Par linéarité ceci est équivalent à

$$\int_{-1}^{1} \pi_1(t)t^k dt = 0, \forall k \in [\![0,1]\!]$$

c'est à dire

$$\int_{-1}^{1} \pi_1(t)dt = 0 \text{ et } \int_{-1}^{1} \pi_1(t)tdt = 0.$$

Or on a

$$\int_{-1}^{1} \pi_1(t)dt = \int_{-1}^{1} t^2 - (t_0+t_1)t + t_0t_1 dt = \frac{2}{3} + 2t_0t_1$$

et

$$\int_{-1}^{1} \pi_1(t)tdt = \int_{-1}^{1} t^3 - (t_0+t_1)t^2 + t_0t_1tdt = -\frac{2}{3}(t_0+t_1).$$

On est amené à résoudre le système non linéaire

$$t_0t_1 = -\frac{1}{3} \text{ et } -\frac{2}{3}(t_0+t_1) = 0$$

ce qui donne $t_0 = -\sqrt(3)/3$ et $t_1 = \sqrt(3)/3$.

Il reste à calculer w_0 et w_1. On a

$$w_0 = \frac{1}{2}\int_{-1}^{1} \frac{t-\sqrt(3)/3}{-2\sqrt(3)/3}dt = \frac{1}{2}\int_{-1}^{1} \frac{-\sqrt(3)/3}{-2\sqrt(3)/3}dt = 1/2$$

et

$$w_1 = \frac{1}{2}\int_{-1}^{1} \frac{t+\sqrt(3)/3}{2\sqrt(3)/3}dt = \frac{1}{2}\int_{-1}^{1} \frac{\sqrt(3)/3}{2\sqrt(3)/3}dt = 1/2.$$

La formule de quadrature

$$\int_{-1}^{1} g(t)dt \approx g(-\frac{\sqrt{3}}{3}) + g(\frac{\sqrt{3}}{3})$$

est donc d'ordre 3.

Q. 2 On effectue le changement de variable $x = \varphi(t) = \frac{a+b}{2} + \frac{b-a}{2}t$ en appliquant la Proposition 5.3 (changement de variable affine) pour obtenir que

$$(b-a)\left(\frac{1}{2}f(x_0) + \frac{1}{2}f(x_1)\right)$$

est une formule de quadrature approchant $\int_a^b f(x)dx$ avec un degré d'exactitude de 3 où $x_0 = \varphi(t_0)$ et $x_1 = \varphi(t_1)$.

◇

Avant d'établir des résultats plus généraux, on donne quelques résultats classiques sur les polynômes de Legendre [].

Les polynômes de Legendre peuvent être définis par la formule de récurrence de Bonnet

$$(n+1)\mathrm{P}_{n+1}(t) = (2n+1)t\mathrm{P}_n(t) - n\mathrm{P}_{n-1}(t),\ \forall n \geqslant 1 \tag{5.15}$$

avec $\mathrm{P}_0(t) = 1$ et $\mathrm{P}_1(t) = t$.

On a les propriétés suivantes:

prop.1 le polynôme de Legendre P_n est de degré n,

prop.2 la famille $\{\mathrm{P}_k\}_{k=0}^n$ est une base de $\mathbb{R}_n[X]$,

prop.3 pour tout $(m,n) \in \mathbb{N}^2$, on a

$$\int_{-1}^{1} \mathrm{P}_m(t)\mathrm{P}_n(t)dt = \frac{2}{2n+1}\delta_{m,n}, \tag{5.16}$$

ce qui correspond à l'orthogonalité des polynômes de Legendre pour le produit scalaire

$$\langle \mathrm{P}_m, \mathrm{P}_n \rangle = \int_{-1}^{1} \mathrm{P}_m(t)\mathrm{P}_n(t)dt.$$

prop.4 Pour $n \geqslant 1$, P_n est scindé sur $\mathbb{R}$ et ses n racines sont simples dans $]-1,1[$, c'est à dire

$$\mathrm{P}_n(t) = C\prod_{i=0}^{n-1}(t-t_i),\ C \in \mathbb{R}^*$$

où les t_i sont 2 à 2 distincts (et ordonnés). Les $(n+1)$ racines simples de P_{n+1} sont alors chacunes dans l'un des $(n+1)$ intervalles $]-1,t_0[$, $]t_0,t_1[$, $\ldots$, $]t_{n-2},t_{n-1}[$, $]t_{n-1},1[$.

Proposition 5.14

Soit $(t_i)_{i=0}^n$ les $n+1$ racines distinctes du polynôme de Legendre de degré $(n+1)$. On note $x_i = \frac{a+b}{2} + \frac{b-a}{2}t_i$, $\forall i \in [\![0,n]\!]$ et w_i les poids donnés par (5.7). La formule de quadrature élémentaire

$$\int_a^b f(x)dx \approx (b-a)\sum_{i=0}^n w_i f(x_i)$$

est appelée la formule de quadrature de Gauss-Legendre. C'est l'unique formule de quadrature élémentaire à $(n+1)$ points ayant pour degré d'exactitude $2n+1$.

n	**exactitude**	w_i (poids)	t_i (points)
0	1	1	0
1	3	$1/2, 1/2$	$-\sqrt{1/3}, \sqrt{1/3}$
2	5	$5/18, 8/18, 5/18$	$-\sqrt{3/5}, 0, \sqrt{3/5}$

Table 5.2: Méthodes de Gauss-Legendre sur $[-1,1]$

Preuve. D'après la Proposition 5.3, il suffit de démontrer ce résultat sur l'intervalle $[-1,1]$ à l'aide du changement de variable affine $x = \varphi(t) = \frac{a+b}{2} + \frac{b-a}{2}t$.

Montrons donc que la formule de quadrature élémentaire

$$\int_{-1}^1 g(t)dt \approx 2\sum_{i=0}^n w_i g(t_i)$$

est de degré d'exactitude $2n+1$.

Comme les poids w_i sont donnés par (5.7), cette formule est de degré d'exactitude au moins n. Pour établir qu'elle est de degré d'exactitude $2n+1$, il faut, d'après la Proposition 5.11 avec $m = n+1$, que

$$\int_{-1}^1 \pi_n(t)\mathrm{Q}(t)dt = 0, \ \forall \mathrm{Q} \in \mathbb{R}_n[X]$$

avec $\pi_n(t) = \prod_{i=0}^n (t-t_i)$. D'après les propriétés des polynômes de Legendre P_n, on a $\mathrm{P}_{n+1}(t) = C\pi_n(t)$ avec $C \in R^*$. On doit donc avoir de manière équivalente

$$\int_{-1}^1 \mathrm{P}_{n+1}(t)\mathrm{Q}(t)dt = 0, \ \forall \mathrm{Q} \in \mathbb{R}_n[X].$$

Or, la famille des les polynômes de Legendre $\{\mathrm{P}_0, \ldots, \mathrm{P}_n\}$ est une base de $\mathbb{R}_n[X]$ et comme les polynômes de Legendre sont orthogonaux, la relation précedente est vérifiée.

Pour démontrer l'unicité de la formule, il suffit d'établir l'unicité des points de quadrature, les poids étant calculés à partir de ces points.

L'unicité du $(n+1)$-uplet $(x_0, \ldots, x_n)$ revient à établir l'unicité du $(n+1)$-uplet $(t_0, \ldots, t_n)$. Supposons qu'il existe $(t_0, \ldots, t_n)$ et $(\tilde{t}_0, \ldots, \tilde{t}_n)$. Notons $\pi_n(t) = \prod_{i=0}^n (t-t_i)$ et $\tilde{\pi}_n(t) = \prod_{i=0}^n (t-\tilde{t}_i)$. On a donc

$$\int_{-1}^1 \pi_n(t)\mathrm{Q}(t)dt = \int_{-1}^1 \tilde{\pi}_n(t)\mathrm{Q}(t)dt = 0, \ \forall \mathrm{Q} \in \mathbb{R}_n[X]$$

Le polynôme $\mathrm{R} = \pi_n - \tilde{\pi}$ est dans $\mathbb{R}_n[X]$ car les polynômes π_n et $\tilde{\pi}_n$ de $\mathbb{R}_{n+1}[X]$ sont unitaires. On a alors

$$\int_{-1}^1 \mathrm{R}(t)\mathrm{Q}(t)dt = 0, \ \forall \mathrm{Q} \in \mathbb{R}_n[X]$$

En choisissant $\mathrm{Q} = \mathrm{R}$, on obtient

$$\int_{-1}^1 \mathrm{R}^2(t)dt = 0$$

ce qui entraine $\mathrm{R} = 0$. □

Théorème 5.15

Soient $f \in \mathcal{C}^{2n+2}([a,b];\mathbb{R})$ et $\mathcal{Q}_n(f,a,b)$ la formule de quadrature de Gauss-Legendre définie dans la Proposition 5.14. Alors on a

$$|\int_a^b f(x)dx - \mathcal{Q}_n(f,a,b)| \leqslant \frac{\|f^{(2n+2)}\|_\infty}{(2n+2)!}\int_a^b \pi_n(x)^2 dx \tag{5.17}$$

où $\pi_n(x) = \prod_{i=0}^n (x-x_i)$, les x_i étant les points de la formule de quadrature.

Preuve. On va utiliser H_n, le polynôme d'interpolation de Lagrange-Hermite aux $(n+1)$ points de la quadrature. D'après le Théorème 4.11, c'est l'unique polynôme de degré au plus $2n+1$ vérifiant

$$\mathrm{H}_n(x_i) = f(x_i) \text{ et } \mathrm{H}_n'(x_i) = f'(x_i),\ \forall i \in [\![0,n]\!].$$

La fonction f étant dans $\mathcal{C}^{2n+2}([a,b];\mathbb{R})$, on peut appliquer le Théorème 4.12 et pour tout $x \in [a,b]$, $\exists \xi_x \in]a,b[$ tel que

$$f(x) - \mathrm{H}_n(x) = \frac{f^{(2n+2)}(\xi_x)}{(2n+2)!}\prod_{i=0}^n (x-x_i)^2.$$

Ensuite on intégre cette relation:

$$\int_a^b (f(x) - \mathrm{H}_n(x))dx = \int_a^b \frac{f^{(2n+2)}(\xi_x)}{(2n+2)!}\prod_{i=0}^n (x-x_i)^2 dx.$$

Le polynôme H_n étant de degré $2n+1$, la formule de quadrature est donc exacte et on a:

$$\begin{aligned}\int_a^b \mathrm{H}_n(x)dx &= \mathcal{Q}_n(\mathrm{H}_n,a,b)\\ &= (b-a)\sum_{i=0}^n w_i \mathrm{H}_n(x_i)\\ &= (b-a)\sum_{i=0}^n w_i f(x_i)\\ &= \mathcal{Q}_n(f,a,b).\end{aligned}$$

On a donc

$$\int_a^b f(x)dx - \mathcal{Q}_n(f,a,b) = \int_a^b \frac{f^{(2n+2)}(\xi_x)}{(2n+2)!}(\pi_n(x))^2 dx.$$

Ce qui donne

$$|\int_a^b f(x)dx - \mathcal{Q}_n(f,a,b)| \leqslant \frac{\|f^{(2n+2)}\|_\infty}{(2n+2)!}\int_a^b \pi_n(x)^2 dx.$$

□

Exercice 5.1.5

L'objectif de cet exercice est de calculer les points et les poids de la formule de quadrature de Gauss-Legendre à $(n+1)$ points. La formule de quadrature de Gauss-Legendre à $(n+1)$ points sur $[-1,1]$ est donnée par

$$\int_{-1}^1 g(t)dt \approx 2\sum_{i=0}^n w_i g(t_i)$$

où les $(t_i)_{i=0}^n$ sont les $n+1$ racines du polynôme de Legendre $\mathrm{P}_{n+1}(t)$. Cette formule à pour degré d'exactitude $2n+1$.

Soient $\langle \mathrm{P},\mathrm{Q}\rangle = \int_{-1}^1 \mathrm{P}(t)\mathrm{Q}(t)dt$ le produit scalaire sur $\mathbb{R}[X]$ et $\|\mathrm{P}\| = \langle \mathrm{P},\mathrm{P}\rangle^{1/2}$ la norme associée. Soit M_n le polynôme de Legendre normalisé de degré $(n+1)$, $\mathrm{M}_n = \frac{\mathrm{P}_n}{\|\mathrm{P}_n\|}$. On utilisera les résultats sur les polynômes de Legendre rappelés en cours.

Q. 1 *Montrer que*

$$c_{n+1}\mathrm{M}_{n+1}(t) = t\mathrm{M}_n(t) - c_n\mathrm{M}_{n-1}(t), \quad n > 1 \tag{5.18}$$

avec

$$\mathrm{M}_0(t) = \sqrt{\frac{1}{2}},\ \mathrm{M}_1(t) = \sqrt{\frac{3}{2}}t \text{ et } c_n = \sqrt{\frac{n^2}{4n^2-1}}$$

On définit le vecteur $\boldsymbol{M}(t)$ de $\mathbb{R}^{n+1}$ par

$$\boldsymbol{M}(t) = (\mathrm{M}_0(t), \ldots, \mathrm{M}_n(t))^t.$$

Q. 2 *Montrer que l'on a*

$$t\boldsymbol{M}(t) = \mathbb{A}\boldsymbol{M}(t) + c_{n+1}\mathrm{M}_{n+1}(t)\boldsymbol{e}_{n+1} \tag{5.19}$$

où l'on explictera la matrice tridiagonale $\mathbb{A} \in \mathcal{M}_{n+1}(\mathbb{R})$ en fonction des coefficients $c_1, \ldots, c_n$. Le vecteur $\boldsymbol{e}_{n+1}$ étant le $(n+1)$-ème vecteur de la base canonique de $\mathbb{R}^{n+1}$.

Q. 3 *En déduire que les $(n+1)$ racines distinctes de $\mathrm{M}_{n+1} \in \mathbb{R}_{n+1}[X]$ sont les $(n+1)$ valeurs propres de $\mathbb{A}$.*

Q. 4 *Montrer que*

$$\sum_{k=0} w_k\mathrm{M}_i(t_k)\mathrm{M}_j(t_k) = \delta_{i,j},\ \forall (i,j) \in [\![0,n]\!]^2 \tag{5.20}$$

où $\delta_{i,j} = 0$, si $i \neq j$ et $\delta_{i,i} = 1$.

On note $\mathbb{W} \in \mathcal{M}_{n+1}(\mathbb{R})$ la matrice diagonale, de diagonale $(w_0, \ldots, w_n)$ et $\mathbb{P} \in \mathcal{M}_{n+1}(\mathbb{R})$ la matrice définie par $\mathbb{P}_{i+1,j+1} = \mathrm{M}_j(t_i)$, $\forall (i,j) \in [\![0,n]\!]^2$.

Q. 5
1. *Montrer que $2\mathbb{P}^t\mathbb{W}\mathbb{P} = \mathbb{I}$.*
2. *En déduire que $\mathbb{W}^{-1} = 2\mathbb{P}\mathbb{P}^t$.*
3. *En déduire que $\frac{1}{w_i} = 2\sum_{k=0}^{n}\left(\mathrm{M}_k(t_i)\right)^2$, $\forall i \in [\![0,n]\!]$.*

On suppose que l'on dispose de la fonction **algorithmique** EIG($\mathbb{A}$) retournant l'ensemble des valeurs propres d'une matrice symétrique $\mathbb{A} \in \mathcal{M}_{n+1}(\mathbb{R})$ dans l'ordre croissant sous la forme d'un vecteur de $\mathbb{R}^{n+1}$.

Q. 6
1. *Ecrire la fonction $[\boldsymbol{t}, \boldsymbol{w}] \leftarrow$ GaussLegendre(n) retournant le tableau des points $\boldsymbol{t}$ et le tableau des poids $\boldsymbol{w}$ en utilisant les résultats obtenus dans cet exercice.*
2. *Ecrire la fonction $I \leftarrow$ QuadElemGaussLegendre(f, a, b, n) retournant une approximation de $\int_a^b f(x)dx$ en utilisant la formule de quadrature de Gauss-Legendre à $(n+1)$ points sur l'intervalle $[a,b]$.*

Correction Exercice On rappelle les résultats suivants (donnés en cours).
Les polynômes de Legendre peuvent être définis par la formule de récurrence de Bonnet

$$(n+1)\mathrm{P}_{n+1}(t) = (2n+1)t\mathrm{P}_n(t) - n\mathrm{P}_{n-1}(t),\ \forall n \geqslant 1$$

avec $\mathrm{P}_0(t) = 1$ et $\mathrm{P}_1(t) = t$.
On a les propriétés suivantes:

1. le polynôme de Legendre P_n est de degré n,
2. la famille $\{\mathrm{P}_k\}_{k=0}^n$ est une base de $\mathbb{R}_n[X]$,
3. pour tout $(m,n) \in \mathbb{N}^2$, on a

$$\int_{-1}^{1} \mathrm{P}_m(t)\mathrm{P}_n(t)dx = \frac{2}{2n+1}\delta_{m,n},$$

ce qui correspond à l'orthogonalité des polynômes de Legendre pour le produit scalaire

$$\langle \mathrm{P}_m, \mathrm{P}_n \rangle = \int_{-1}^{1} \mathrm{P}_m(t)\mathrm{P}_n(t)dx.$$

4. Pour $n \geqslant 1$, P_n est scindé sur $\mathbb{R}$ et ses n racines sont simples dans $]-1,1[$, c'est à dire

$$\mathrm{P}_n(t) = C\prod_{i=0}^{n-1}(t-t_i),\ C \in \mathbb{R}^*$$

où les t_i sont 2 à 2 distincts (et ordonnés). Les $n+1$ racines simples de P_{n+1} sont alors chacunes dans l'un des $n+1$ intervalles $]a,t0[$, $]t_0,t_1[$, $\ldots$, $]t_{n-2},t_{n-1}[$, $]t_{n-1},b[$.

Q. 1 Par définition, on a $\mathrm{M}_n = \frac{\mathrm{P}_n}{\|\mathrm{P}_n\|}$ et

$$\|\mathrm{P}_n\|^2 = \langle \mathrm{P}_n, \mathrm{P}_n\rangle = \int_{-1}^{1}\mathrm{P}_n(t)\mathrm{P}_n(t)dx = \frac{2}{2n+1}.$$

On en déduit que

$$\mathrm{M}_n = \sqrt{\frac{2n+1}{2}}\mathrm{P}_n.$$

De plus, par la formule de récurrence de Bonnet, on obtient

$$\mathrm{M}_0(t) = \sqrt{\frac{1}{2}} \text{ et } \mathrm{M}_1(t) = \sqrt{\frac{3}{2}}t$$

ainsi que, $\forall n \geqslant 1$,

$$\begin{aligned}(n+1)\sqrt{\frac{2}{2n+3}}\mathrm{M}_{n+1}(t) &= (2n+1)\sqrt{\frac{2}{2n+1}}t\mathrm{M}_n(t) - n\sqrt{\frac{2}{2n-1}}\mathrm{M}_{n-1}\\ &= \sqrt{2(2n+1)}t\mathrm{M}_n(t) - n\sqrt{\frac{2}{2n-1}}\mathrm{M}_{n-1}\end{aligned}$$

En multipliant cette équation par $\sqrt{\frac{1}{2(2n+1)}}$, on a

$$(n+1)\sqrt{\frac{2}{2n+3}}\sqrt{\frac{1}{2(2n+1)}}\mathrm{M}_{n+1}(t) = t\mathrm{M}_n(t) - n\sqrt{\frac{2}{2n-1}}\sqrt{\frac{1}{2(2n+1)}}\mathrm{M}_{n-1}$$

Or on a

$$n\sqrt{\frac{2}{2n-1}}\sqrt{\frac{1}{2(2n+1)}} = \sqrt{\frac{n^2}{(2n-1)(2n+1)}} = c_n$$

et

$$(n+1)\sqrt{\frac{2}{2n+3}}\sqrt{\frac{1}{2(2n+1)}} = \sqrt{\frac{(n+1)^2}{(2(n+1)-1)(2(n+1)+1)}} = c_{n+1}$$

ce qui démontre le résultat voulu.

Q. 2 On déduit de la question précédente que

$$t\mathrm{M}_0(t) = \sqrt{\frac{1}{3}}t = c_1\mathrm{M}_1(t)$$

et

$$t\mathrm{M}_i(t) = c_i\mathrm{M}_{i-1}(t) + c_{i+1}\mathrm{M}_{i+1}(t),\ \forall i \in [\![1,n]\!].$$

Ces $(n+1)$ équations peuvent s'écrire matriciellement sous la forme

$$\begin{aligned}t\begin{pmatrix}\mathrm{M}_0(t)\\ \mathrm{M}_1(t)\\ \vdots\\ \vdots\\ \mathrm{M}_{n-1}(t)\\ \mathrm{M}_n(t)\end{pmatrix} &= \begin{pmatrix}0 & c_1 & 0 & \ldots & \ldots & 0\\ c_1 & 0 & c_2 & 0 & \ldots & 0\\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & \ddots & 0\\ 0 & \ldots & 0 & c_{n-1} & 0 & c_n\\ 0 & \ldots & \ldots & 0 & c_n & 0\end{pmatrix}\begin{pmatrix}\mathrm{M}_0(t)\\ \mathrm{M}_1(t)\\ \vdots\\ \vdots\\ \mathrm{M}_{n-1}(t)\\ \mathrm{M}_n(t)\end{pmatrix} + \begin{pmatrix}0\\ \vdots\\ \vdots\\ \vdots\\ 0\\ c_{n+1}\mathrm{M}_{n+1}(t)\end{pmatrix}\\ &= \mathbb{A}\boldsymbol{M}(t) + c_{n+1}\mathrm{M}_{n+1}(t)\boldsymbol{e}_{n+1}.\end{aligned}$$

Q. 3 Le polynôme de Legendre $\mathrm{P}_{n+1} \in \mathbb{R}_{n+1}[X]$ admet $(n+1)$ racines simples distinctes dans $]-1,1[$ notées $(t_i)_{i=0}^n$. (voir rappels) Donc le polynôme de Legendre normalisé $\mathrm{M}_{n+1} \in \mathbb{R}_{n+1}[X]$ a les mêmes racines et on déduit de la question précédente que

$$\mathbb{A}\boldsymbol{M}(t_i) = t_i \boldsymbol{M}(t_i),\ \forall i \in [\![0,n]\!].$$

Comme les $(n+1)$ racines de P_{n+1} séparent strictement les n racines de P_n (voir rappels), alors $\mathrm{P}_n(t_i) \neq 0$ et donc $\mathrm{M}_n(t_i) \neq 0$. On en déduit que le vecteur $\boldsymbol{M}(t_i)$ est non nul et,

$$\forall i \in [\![0,n]\!],\ (t_i, \boldsymbol{M}(t_i)) \text{ est un mode propre de } \mathbb{A}.$$

On peut noter que $\mathbb{A}$ est symétrique et donc ses valeurs propres sont réelles.
Les $(n+1)$ valeurs propres de la matrice $\mathbb{A}$ sont les $(n+1)$ racines de P_{n+1}, et donc les $(n+1)$ points de la formule de quadrature de Gauss-Legendre.

Q. 4 Par construction, on a

$$\int_{-1}^{1} \mathrm{M}_i(t)\mathrm{M}_j(t)dx = \delta_{i,j},\ \forall (i,j) \in [\![0,n]\!].$$

On a $\mathrm{M}_i \in \mathbb{R}_i[X]$ et $\mathrm{M}_j \in \mathbb{R}_j[X]$, ce qui donne $\mathrm{M}_i\mathrm{M}_j \in \mathbb{R}_{i+j}[X]$ avec $i+j \leqslant n$. Or la formule de quadrature de Gauss-Legendre à $(n+1)$ points a pour degré d'exactitude $2n+1$, elle est donc exacte pour le polynôme $\mathrm{M}_i\mathrm{M}_j$. On en déduit alors

$$\delta_{i,j} = \int_{-1}^{1} \mathrm{M}_i(t)\mathrm{M}_j(t)dt = 2\sum_{k=0}^{n} w_k \mathrm{M}_i(t_k)\mathrm{M}_j(t_k),\ \forall (i,j) \in [\![0,n]\!]$$

Q. 5 1. Soit $(i,j) \in [\![1,n+1]\!]^2$, on a

$$\begin{aligned}(\mathbb{P}^{\mathtt{t}}\mathbb{W}\mathbb{P})_{i,j} &= \sum_{k=1}^{n+1} (\mathbb{P}^{\mathtt{t}})_{i,k}(\mathbb{W}\mathbb{P})_{k,j} \\ &= \sum_{k=1}^{n+1} \mathbb{P}_{k,i}(\mathbb{W}\mathbb{P})_{k,j}\end{aligned}$$

et, comme $\mathbb{W}$ est diagonale,

$$(\mathbb{W}\mathbb{P})_{k,j} = \sum_{l=1}^{n+1} \mathbb{W}_{k,l}\mathbb{P}_{l,j} = \mathbb{W}_{k,k}\mathbb{P}_{k,j}.$$

On obtient donc

$$(\mathbb{P}^{\mathtt{t}}\mathbb{W}\mathbb{P})_{i,j} = \sum_{k=1}^{n+1} \mathbb{P}_{k,i}\mathbb{W}_{k,k}\mathbb{P}_{k,j} = \sum_{k=1}^{n+1} w_{k-1}\mathrm{M}_{i-1}(t_{k-1}\mathrm{M}_{j-1}(t_{k-1}).$$

En utilisant la relation démontré dans la question précédente, on a

$$(\mathbb{P}^{\mathtt{t}}\mathbb{W}\mathbb{P})_{i,j} = \frac{1}{2}\delta_{i-1,j-1} = \frac{1}{2}\delta_{i,j}$$

c'est à dire

$$\mathbb{P}^{\mathtt{t}}\mathbb{W}\mathbb{P} = \frac{1}{2}\mathbb{I}$$

et on en déduit que $\mathbb{P}$ et $\mathbb{W}$ sont inversibles .

2. A partir de $\mathbb{P}^{\mathtt{t}}\mathbb{W}\mathbb{P} = \frac{1}{2}\mathbb{I}$, on déduit

$$2\mathbb{I} = \left(\mathbb{P}^{\mathtt{t}}\mathbb{W}\mathbb{P}\right)^{-1} = 2\mathbb{I} = \mathbb{P}^{-1}\mathbb{W}^{-1}\left(\mathbb{P}^{\mathtt{t}}\right)^{-1}$$

En multipliant par $\mathbb{P}$ à gauche et par $\mathbb{P}^{\mathtt{t}}$ à droite, on obtient

$$\mathbb{W}^{-1} = 2\mathbb{P}\mathbb{P}^{\mathtt{t}}.$$

3. Comme la matrice $\mathbb{W}$ est diagonale inversible, son inverse est diagonale et on a

$$\left(\mathbb{W}^{-1}\right)_{i,i} = \frac{1}{\mathbb{W}_{i,i}} = \frac{1}{w_{i-1}}, \ \forall i \in [\![1, n+1]\!].$$

On a donc

$$\begin{aligned}\frac{1}{w_{i-1}} &= 2\left(\mathbb{P}\mathbb{P}^{\mathsf{t}}\right)_{i,i} \\ &= 2\sum_{j=1}^{n+1} \mathbb{P}_{i,j}\left(\mathbb{P}^{\mathsf{t}}\right)_{j,i} \\ &= 2\sum_{j=1}^{n+1} \mathbb{P}_{i,j}^2 \\ &= 2\sum_{k=0}^{n} \left(\mathrm{M}_k(t_{i-1})\right)^2, \ \forall i \in [\![1, n+1]\!].\end{aligned}$$

Q. 6 1. Les $(n+1)$ points $(t_i)_{i=0}^n$ de la méthode de quadrature de Gauss-Legendre sur $[-1, 1]$, sont les racines du polynôme de Legendre P_{n+1} de degré $n+1$. Pour les calculer, on va utiliser le fait que ce sont les valeurs propres de la matrice symétrique $\mathbb{A} \in \mathcal{M}_{n+1}(\mathbb{R})$

$$\begin{pmatrix} 0 & c_1 & 0 & \dots & \dots & 0 \\ c_1 & 0 & c_2 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & c_{n-1} & 0 & c_n \\ 0 & \dots & \dots & 0 & c_n & 0 \end{pmatrix}$$

avec $c_k = \sqrt{\frac{k^2}{4k^2-1}}$, $\forall k \geqslant 1$.
Pour calculer les poids $(w_i)_{i=0}^n$, on va utiliser la formule

$$\frac{1}{w_i} = 2\sum_{k=0}^{n} \left(\mathrm{M}_k(t_i)\right)^2, \ \forall i \in [\![0, n]\!]$$

conjointement avec la formule de récurrence

$$\mathrm{M}_k(t) = \frac{1}{c_k}\left(t\mathrm{M}_{k-1}(t) - c_{k-1}\mathrm{M}_{k-2}(t)\right), \ k \geqslant 2, \text{ avec } \mathrm{M}_0(t) = \sqrt{\frac{1}{2}}, \ \mathrm{M}_1(t) = \sqrt{\frac{3}{2}}t.$$

Algorithme 5.5 Fonction GaussLegendre retournant le tableau des points $\boldsymbol{t}$ et le tableau des poids $\boldsymbol{w}$

Données : n : $n \in \mathbb{N}$
Résultat : $\boldsymbol{t}$: vecteur de $\mathbb{R}^{n+1}$ avec $\boldsymbol{t}(i) = t_{i-1}$, $\forall i \in [\![1, n+1]\!]$
$\boldsymbol{w}$: vecteur de $\mathbb{R}^{n+1}$ avec $\boldsymbol{w}(i) = w_{i-1}$, $\forall i \in [\![1, n+1]\!]$

Fonction $[\boldsymbol{t}, \boldsymbol{w}] \leftarrow$ GaussLegendre (n)
  $\boldsymbol{c} \leftarrow \boldsymbol{O}_n$
  $\mathbb{A} \leftarrow \mathbb{O}_{n+1,n+1}$
  Pour $k \leftarrow 1$ **à** n **faire**
    $\boldsymbol{c}(k) \leftarrow$ sqrt$(k^2/(4 * k\hat{}2 - 1))$
    $\mathbb{A}(k, k+1) \leftarrow \boldsymbol{c}(k)$
    $\mathbb{A}(k+1, k) \leftarrow \boldsymbol{c}(k)$
  Fin Pour
  $\boldsymbol{t} \leftarrow$ eig$(\mathbb{A})$
  Pour $i \leftarrow 1$ **à** $n+1$ **faire**
    $M0 \leftarrow$ sqrt$(1/2)$
    $M1 \leftarrow$ sqrt$(3/2) * \boldsymbol{t}(i)$
    $S \leftarrow M0\hat{}2 + M1\hat{}2$
    Pour $k \leftarrow 2$ **à** n **faire**
      $M \leftarrow (1/\boldsymbol{c}(k)) * (M1 * \boldsymbol{t}(i) - \boldsymbol{c}(k-1) * M0)$
      $S \leftarrow S + M\hat{}2$
      $M0 \leftarrow M1$
      $M1 \leftarrow M$
    Fin Pour
    $w(i) \leftarrow 1/(2 * S)$
  Fin Pour
Fin Fonction

2. On va utiliser la formule

$$I = (b-a) = \sum_{i=0}^{n} w_i f(x_i)$$

avec $x_i = \frac{a+b}{2} + \frac{b-a}{2} t_i$ où les points $(t_i)_{i=0}^n$ et les poids $(w_i)_{i=0}^n$ sont ceux de la méthode de quadrature de Gauss-Legendre sur $[-1, 1]$.

Algorithme 5.6 Fonction QuadElemGaussLegendre retournant une approximation de $\int_a^b f(x)dx$ en utilisant la formule de quadrature de Gauss-Legendre à $n+1$ points sur l'intervalle $[a, b]$.

Données : f : une fonction de $[a, b]$ à valeurs réels
a, b : deux réels avec $a < b$
n : $n \in \mathbb{N}$
Résultat : I : un réel

Fonction $I \leftarrow$ QuadElemGaussLegendre (f, a, b, n)
  $[\boldsymbol{t}, \boldsymbol{w}] \leftarrow$ GaussLegendre(n)
  $I \leftarrow 0$
  Pour $i \leftarrow 1$ **à** $n+1$ **faire**
    $I \leftarrow I + \boldsymbol{w}(i) * f((a+b)/2 + (b-a)/2 * t(i))$
  Fin Pour
  $I \leftarrow (b-a) * I$
Fin Fonction

◇

5.2 Méthodes de quadrature composées

Soit f une fonction définie et intégrable sur un intervalle $[\alpha, \beta]$ donné. On propose de chercher une approximation de

$$I = \int_\alpha^\beta f(x)dx.$$

Ces méthodes consistent en l'utilisation de la relation de Chasles pour décomposer l'intégrale en une somme d'intégrales sur des domaines plus petits puis à approcher ces dernières par une formule de quadrature élémentaire à $n+1$ points.
La **méthode de quadrature composée associée à** $\mathcal{Q}_n$, notée $\mathcal{Q}_{k,n}^{\text{comp}}$, est alors donnée par

$$\mathcal{Q}_{k,n}^{\text{comp}}(f,\alpha,\beta) = \sum_{i=1}^{k} \mathcal{Q}_n(f,\alpha_{i-1},\alpha_i) \approx \int_\alpha^\beta f(x)dx \tag{5.21}$$

Definition 5.16

Soit $(\alpha_i)_{i\in[\![0,k]\!]}$ une subdivison de l'intervalle $[\alpha,\beta]$:

$$\alpha = \alpha_0 < \alpha_1 < \cdots < \alpha_k = \beta.$$

On a alors

$$\int_\alpha^\beta f(x)dx = \sum_{i=1}^{k} \int_{\alpha_{i-1}}^{\alpha_i} f(x)dx. \tag{5.22}$$

Soit $\mathcal{Q}_n(g,a,b)$ la formule de quadrature élémentaire à $n+1$ points d'ordre p donnée par

$$\mathcal{Q}_n(g,a,b) \overset{\text{def}}{=} (b-a)\sum_{j=0}^{n} w_j g(x_j) \approx \int_a^b g(x)dx.$$

La **méthode de quadrature composée associée à** $\mathcal{Q}_n$, notée $\mathcal{Q}_{k,n}^{\text{comp}}$, est donnée par

$$\mathcal{Q}_{k,n}^{\text{comp}}(f,\alpha,\beta) = \sum_{i=1}^{k} \mathcal{Q}_n(f,\alpha_{i-1},\alpha_i) \approx \int_\alpha^\beta f(x)dx \tag{5.23}$$

La proposition suivante est immédiate

Proposition 5.17

Soit $\mathcal{Q}_n$ une formule de quadrature élémentaire à $n+1$ points. Si $\mathcal{Q}_n$ est d'ordre p alors la méthode de quadrature composée associée est aussi d'ordre p : elle est exacte pour tout polynôme de degré p.

5.2.1 Exemples

Pour les trois formules composites qui suivent on choisit une discrétisation régulière. Soit $(\alpha_i)_{i\in[\![0,k]\!]}$ une discrétisation régulière de l'intervalle $[\alpha;\beta]$: $\alpha_i = \alpha + ih$ avec $h = (\beta-\alpha)/k$ le pas de la discrétisation.

Formule composite des points milieux

La formule élémentaire des points milieux est donnée par

$$\mathcal{Q}_0(g,a,b) \overset{\text{def}}{=} (b-a)g(\frac{a+b}{2})$$

On note $m_j = \frac{\alpha_{j-1}+\alpha_j}{2}$ le point milieu de l'intervalle $[\alpha_{j-1},\alpha_j]$.

$$\int_\alpha^\beta f(x)dx = \sum_{j=1}^{k}\int_{\alpha_{j-1}}^{\alpha_j} f(x)dx \approx \sum_{j=1}^{k} \mathcal{Q}_0(f,\alpha_{j-1},\alpha_j) = h\sum_{j=1}^{k} f(m_j) \tag{5.24}$$

On illustre graphiquement l'approximation d'une intégrale par cette formule en Figure 5.6.

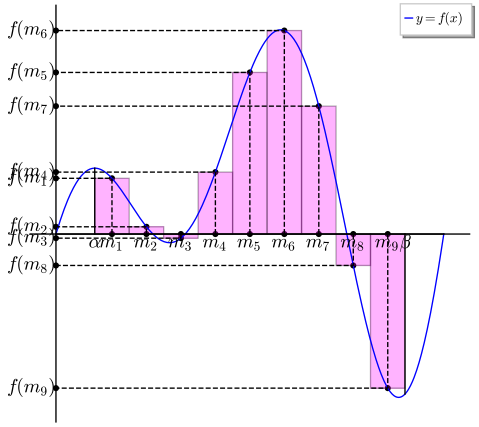


Figure 5.6: Formule composite des points milieux : $\int_\alpha^\beta f(x)dx \approx h \sum_{j=1}^{k} f(m_j)$ (aire de la surface colorée)

Formule composite des trapèzes

La formule élémentaire des points trapèzes est donnée par

$$\mathcal{Q}_1(g,a,b) \stackrel{\text{def}}{=} \frac{b-a}{2}(g(a)+g(b))$$

On obtient alors

$$\begin{aligned}\int_a^b f(x)dx = \sum_{j=1}^{k}\int_{\alpha_{j-1}}^{\alpha_j} f(x)dx &\approx \sum_{j=1}^{k}\mathcal{Q}_0(f,\alpha_{j-1},\alpha_j) = \frac{h}{2}\sum_{j=1}^{k}(f(\alpha_{j-1})+f(\alpha_j)) \\ &\approx \frac{h}{2}\left(f(\alpha_0)+2\sum_{j=1}^{k-1}f(\alpha_j)+f(\alpha_k)\right) \qquad (5.25)\end{aligned}$$

C'est une formule d'ordre 2 par rapport à h.

On illustre graphiquement l'approximation d'une intégrale par cette formule en Figure 5.7.

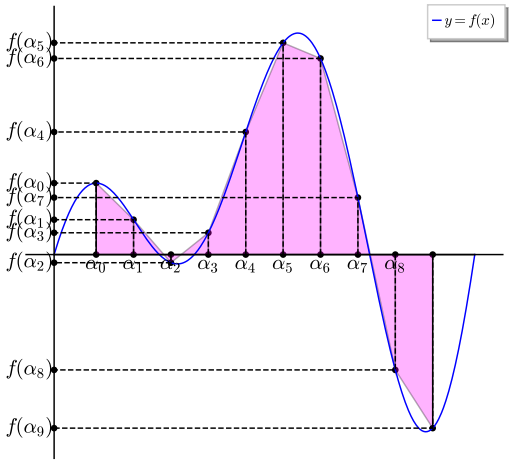


Figure 5.7: Formule composite des trapèzes : $\int_\alpha^\beta f(x)dx \approx \frac{h}{2}\left(f(\alpha_0) + 2\sum_{j=1}^{k-1} f(\alpha_j) + f(\alpha_k)\right)$ (aire de la surface colorée)

Formule composite de Simpson

Exercice 5.2.1

Ecrire une fonction algorithmique QuadSimpson retournant une approximation de l'intégrale d'une fonction f sur l'intervalle $[\alpha, \beta]$ utilisant la méthode de quadrature composée de Simpson en **minimisant** le nombre d'appels à la fonction f. On rappelle que la formule élémentaire de Simpson est donnée par

$$\mathcal{Q}_2(g, a, b) \stackrel{\text{def}}{=} \frac{b-a}{6}(g(a) + 4g(\frac{a+b}{2}) + g(b)).$$

Correction Exercice En notant $m_j = \frac{\alpha_{j-1}+\alpha_j}{2}$ le point milieu de l'intervalle $[\alpha_{j-1}, \alpha_j]$, on obtient

$$\begin{aligned}\int_a^b f(x)dx \sum_{j=1}^{k}\int_{\alpha_{j-1}}^{\alpha_j} f(x)dx &\approx \sum_{j=1}^{k}\mathcal{Q}_2(f, \alpha_{j-1}, \alpha_j) = \frac{h}{6}\sum_{j=1}^{k}(f(\alpha_{j-1}) + 4f(m_j) + f(\alpha_j)) \\ &\approx \frac{h}{6}\left(4\sum_{j=1}^{k} f(m_j) + f(\alpha_0) + 2\sum_{j=1}^{k-1} f(\alpha_j) + f(\alpha_k)\right) \qquad (5.26)\end{aligned}$$

Algorithme 5.7 Fonction QuadSimpson retourne une approximation de l'intégrale d'une fonction f sur l'intervalle $[\alpha, \beta]$ utilisant la méthode de quadrature composée de Simpson en **minimisant** le nombre d'appels à la fonction f.

Données : f : une fonction définie de $[alpha, beta]$ dans $\mathbb{R}$,
$alpha, beta$: deux réels avec $alpha < beta$,
k : $n \in \mathbb{N}^*$

Résultat : I : un réel

1: **Fonction** $I \leftarrow$ QuadSimpson $(\ f, alpha, beta, k\ )$
2: $h \leftarrow (beta - alpha)/k$
3: $\boldsymbol{x} \leftarrow alpha : h : beta$
4: $\boldsymbol{m} \leftarrow alpha + h/2 : h : beta$
5: $S \leftarrow 0$ ▷ Calcul de $\sum_{j=1}^{k} f(m_j)$
6: **Pour** $j \leftarrow 1$ **à** k **faire**
7: $S \leftarrow S + f(\boldsymbol{m}(j))$
8: **Fin Pour**
9: $I \leftarrow 4 * S$
10: $S \leftarrow 0$ ▷ Calcul de $\sum_{j=1}^{k-1} f(\alpha_j) = \sum_{j=2}^{k} f(\boldsymbol{x}_j)$
11: **Pour** $j \leftarrow 2$ **à** k **faire**
12: $S \leftarrow S + f(\boldsymbol{x}(j))$
13: **Fin Pour**
14: $I \leftarrow (h/6) * (I + 2 * S + f(\boldsymbol{x}(1)) + f(\boldsymbol{x}(k+1)))$
15: **Fin Fonction**

◇

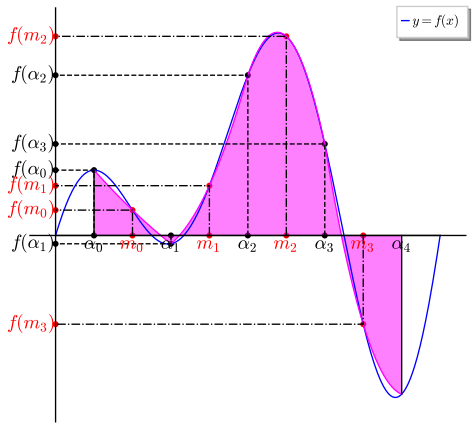


Figure 5.8: Formule composite de Simpson : $\int_\alpha^\beta f(x)dx \approx \frac{h}{6}\left(4\sum_{j=1}^{k} f(m_j) + f(\alpha_0) + 2\sum_{j=1}^{k-1} f(\alpha_j) + f(\alpha_k)\right)$ (aire de la surface colorée)

5.2.2 Erreurs des méthodes de quadrature composées

Soit $\mathcal{Q}_{k,n}^{\text{comp}}$ une méthode de quadrature composée associée à une méthode de quadrature élémentaire $\mathcal{Q}_n$. On note $\mathcal{E}_{\alpha,\beta}^{\text{comp}}(f)$ l'erreur de cette méthode de quadrature composée :

$$\mathcal{E}_{\alpha,\beta}^{\text{comp}}(f) = \int_\alpha^\beta f(x)dx - \mathcal{Q}_{k,n}^{\text{comp}}(f,\alpha,\beta). \tag{5.27}$$

On a alors

$$\mathcal{E}_{\alpha,\beta}^{\text{comp}}(f) = \sum_{j=1}^{k}\left(\int_{\alpha_{j-1}}^{\alpha_j} f(x)dx - \mathcal{Q}_n(f,\alpha_{j-1},\alpha_j)\right) = \sum_{j=1}^{k}\mathcal{E}_{\alpha_{j-1},\alpha_j}(f)$$

Dans le cadre des méthodes composées de Newton-Cotes, on peut démontrer (voir [2]), la majoration suivante

$$\max_{x\in[a,b]}|\prod_{i=0}^{n}(x-x_i)| \leqslant C\frac{e^{-n}}{\sqrt{n}\log(n)}(b-a)^{n+1}. \tag{5.28}$$

où $(x_i)_{i\in[\![0,n]\!]}$ est la discrétisation régulière de $[a,b]$ et $C>0$. On suppose que $f\in\mathcal{C}^{n+1}([a,b];\mathbb{R})$. En notant $h_j=\alpha_j-\alpha_{j-1}$ et en utilisant les majorations (5.8) et (5.28) on obtient

$$\begin{aligned}
|\mathcal{E}_{\alpha,\beta}^{\text{comp}}(f)| &\leqslant \sum_{j=1}^{k}|\mathcal{E}_{\alpha_{j-1},\alpha_j}(f)| \\
&\leqslant \sum_{j=1}^{k}\frac{K_n}{(n+1)!}\max_{x\in[\alpha_{j-1},\alpha_j]}|f^{(n+1)}(x)|h_j^{n+2} \quad \text{avec } K_n = C\frac{e^{-n}}{\sqrt{n}\log(n)} \\
&\leqslant K_n\frac{h^{n+1}}{(n+1)!}\max_{x\in[\alpha,\beta]}|f^{(n+1)}(x)|\sum_{j=1}^{k}h_j \quad \text{avec } h=\max_{j\in[\![1,k]\!]}h_j \\
&\leqslant K_n(\beta-\alpha)\frac{h^{n+1}}{(n+1)!}\left\|f^{(n+1)}\right\|_\infty
\end{aligned}$$

Cette majoration n'est pas optimale et n'est valable que pour les formules composées de Newton-Cotes. A l'aide des **noyaux de Peano**, on peut démontrer le théorème suivant :

Théorème 5.18: [2], page 43 (admis)

Soient $\mathcal{Q}_{k,n}^{\text{comp}}$ une méthode de quadrature composée associée à une méthode de quadrature élémentaire $\mathcal{Q}_n$ de degré d'exactitude $p\geqslant n$ et $f\in\mathcal{C}^{p+1}([\alpha,\beta];\mathbb{R})$. On a alors

$$|\int_\alpha^\beta f(x)dx - \mathcal{Q}_{k,n}^{\text{comp}}(f,\alpha,\beta)| \leqslant C_p(\beta-\alpha)h^{p+1}\left\|f^{(p+1)}\right\|_\infty \tag{5.29}$$

avec $h=\max_{j\in[\![1,k]\!]}(\alpha_j-\alpha_{j-1})$ et $C_p>0$. Ceci s'écrit aussi sous la forme

$$|\int_\alpha^\beta f(x)dx - \mathcal{Q}_{k,n}^{\text{comp}}(f,\alpha,\beta)| = \mathcal{O}(h^{p+1}) \tag{5.30}$$

et son **ordre de convergence** est $p+1$.

On illustre ce théorème pour les méthodes de Newton-Cotes composées $\mathcal{Q}_{k,n}^{\text{comp}}$, pour $n\in[\![1,8]\!]$, par les Figures 5.9 et 5.10.

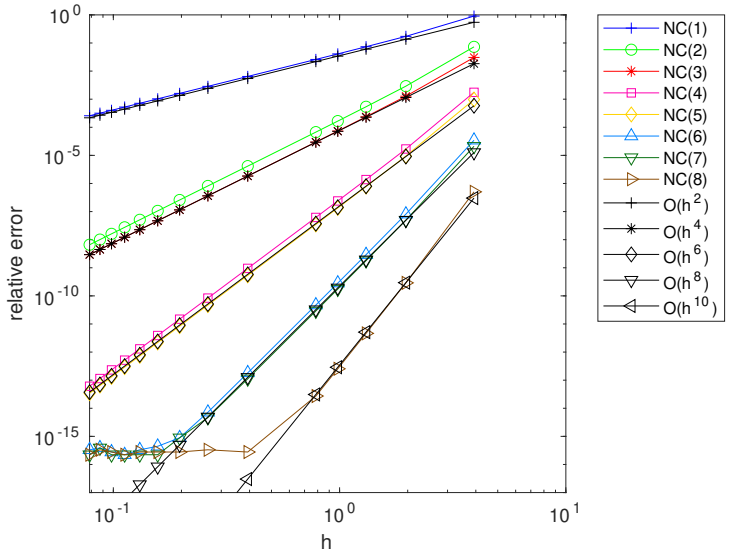


Figure 5.9: Erreur des méthodes de Newton-Cotes composées pour le calcul de $\int_0^{5\pi/2} cos(x)dx$, `NC(n)` correspondant à $\mathcal{Q}_{k,n}^{\text{comp}}$ et $h = \frac{5\pi}{2k}$.

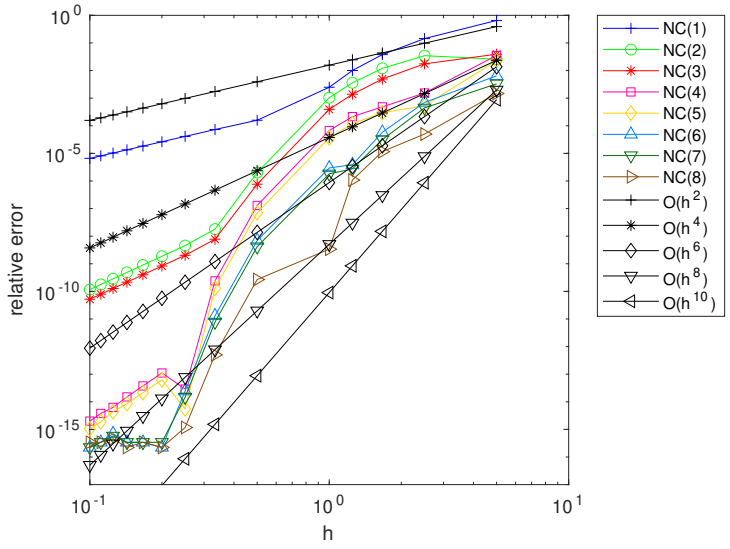


Figure 5.10: Erreur des méthodes de Newton-Cotes composées pour le calcul de $\int_{-5}^{5} \frac{1}{1+x^2}dx$, `NC(n)` correspondant à $\mathcal{Q}_{k,n}^{\text{comp}}$ et $h = \frac{10}{k}$.

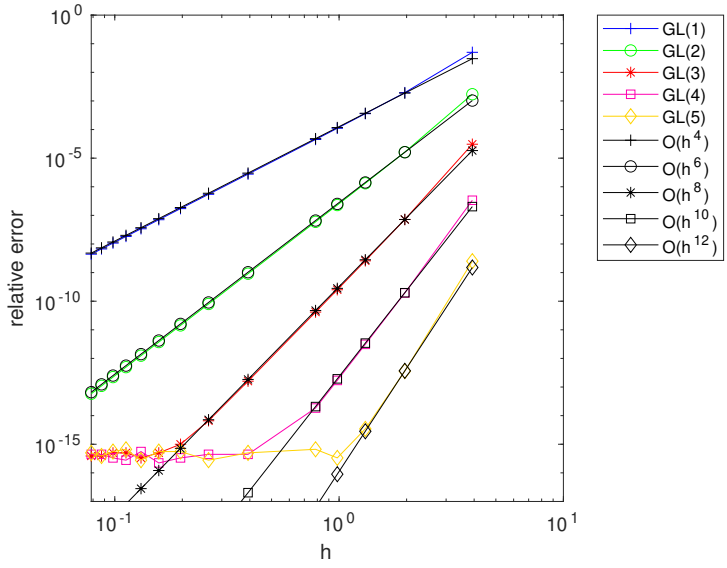


Figure 5.11: Erreur des méthodes de Gauss-Legendre composées pour le calcul de $\int_0^{5\pi/2} cos(x)dx$, `GL(n)` correspondant à $\mathcal{Q}_{k,n}^{\text{comp}}$ et $h = \frac{5\pi}{2k}$.

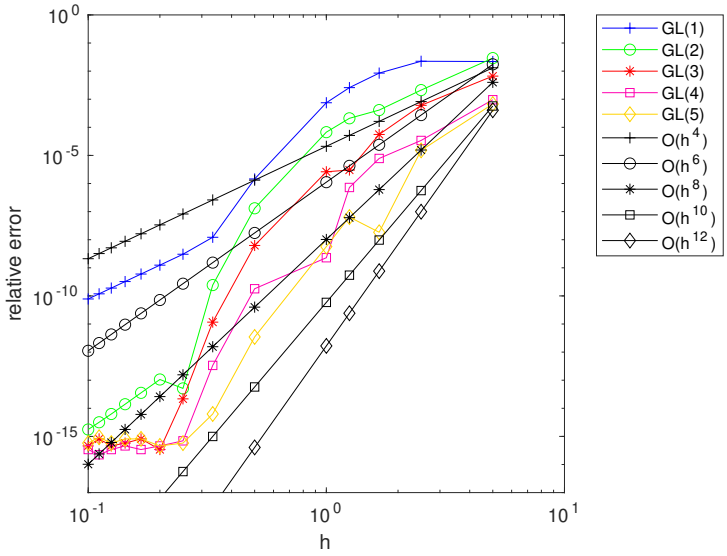


Figure 5.12: Erreur des méthodes de Gauss-Legendre composées pour le calcul de $\int_{-5}^{5} \frac{1}{1+x^2}dx$, `GL(n)` correspondant à $\mathcal{Q}_{k,n}^{\text{comp}}$ et $h = \frac{10}{k}$.

Comprendre les ordres de convergence de méthodes numériques

De manière générale, l'ordre de convergence de l'approximation d'une formule, d'un schéma, d'une méthode est donné par une majoration de la norme de la différence entre la solution exacte u_{ex} d'un problème et son approximation u_h où usuellement h est un paramètre correspondant à la *finesse* de l'approximation numérique (plus la méthode numérique est *précise*, plus h est proche de 0). On dit alors qu'une méthode numérique est convergente d'**ordre** p si

$$\|u_{\text{ex}} - u_h\| = \mathcal{O}(h^p).$$

On retouve cette notion d'ordre dans la résolution numérique d'E.D.O.[1], dans la résolution numérique d'E.D.P.[2] par différences finies, éléments finis, volumes finis. Ceci permettra de **vérifier/valider** les méthodes numériques implémentées.

Dans le cadre de l'intégration numérique, le Théorème 5.18 et plus particulièrement la formule (5.2.2), affirme que si une méthode de quadrature élémentaire $\mathcal{Q}_n$ est de degré d'exactitude $p \geqslant n$ et que $f \in \mathcal{C}^{p+1}([\alpha,\beta];\mathbb{R})$ alors

$$|\int_\alpha^\beta f(x)dx - \mathcal{Q}_{k,n}^{\text{comp}}(f,\alpha,\beta)| = \mathcal{O}(h^{p+1}).$$

et donc son ordre de convergence est $p+1$.

Il est possible de "retrouver" numériquement l'**ordre** de convergence de méthodes numériques en représentant en échelle logarithmique (en abscisses et ordonnées) la fonction erreur $h \to E(h)$ pour chacune des méthodes.
Par exemple, pour la méthode composite de Simpson, on a vu que pour un $h = (b-a)/N$ donné (i.e. un N donné)

$$E(h) = |\int_a^b f(x)dx - \frac{h}{6}\sum_{i=1}^{N}(f(x_{i-1}) + 4f(m_i) + f(x_i))| = \mathcal{O}(h^4)$$

Pour h suffisament petit, il existe alors $C > 0$ tel que

$$E(h) = Ch^4.$$

On en déduit que si l'on multiplie N par 10 alors l'erreur sera divisée par 10^4. En effet, on a

$$E(\frac{h}{10}) = C(\frac{h}{10})^4 = \frac{E(h)}{10^4}.$$

Pour illustrer ceci, on propose le Listing 5.1 où est calculé une approximation de $\int_0^{\pi/2} cos(x)dx$ par la méthode composée des trapèzes (degré d'exactitude 1) et par la méthode composée de Simpson (degré d'exactitude 3): on remarque que l'erreur pour la méthode des trapèzes entre $N = 10$ et $N = 100$ est bien divisée par 10^2 et, pour la méthode de Simpson, elle est bien divisée par 10^4.

Listing 5.1: : script Matlab pour illustrer l'ordre des méthodes des Trapèzes et de Simpson

```
f=@(x) cos(x);
F=@(x) sin(x);
a=0;b=pi/2;
Iex=F(b)-F(a);
I1=QuadTrapeze(f,a,b,10);
I2=QuadTrapeze(f,a,b,100);
fprintf('Erreurs␣Trapeze:␣(N=10)␣%.5e␣-␣(N=100)␣%.5e\n',abs(I1-Iex),abs(I2-Iex))
I1=QuadSimpson(f,a,b,10);
I2=QuadSimpson(f,a,b,100);
fprintf('Erreurs␣Simpson:␣(N=10)␣%.5e␣-␣(N=100)␣%.5e\n',abs(I1-Iex),abs(I2-Iex))
```

Output

```
Erreurs Trapeze: (N=10) 2.05701e-03 - (N=100) 2.05618e-05
Erreurs Simpson: (N=10) 2.11547e-07 - (N=100) 2.11393e-11
```

Nous allons maintenant représenter graphiquement ce phénomène. On note tout d'abord que

$$\log(E(h)) = \log(C) + 4\log(h),$$

et donc, en échelle logarithmique, on va représenter $\log(h) \to \log(E(h))$ qui est une droite de pente 4. En Figure 5.13, on représente en échelle logarithmique les différentes erreurs ainsi que les fonctions $h \to h^2$ et $h \to h^4$. Le code Matlab/Octave est donné en Listing 5.2.

```
f=@(x) cos(x);
F=@(x) sin(x);
a=0;b=pi/2;
Iex=F(b)-F(a);
LN=25:25:200;
k=1;
for N=LN
  H(k)=(b-a)/N;
```

[1] Equations Différentielles Ordinaires
[2] Equations aux Dérivées Partielles

```
  E1(k)=abs(QuadTrapeze(f,a,b,N)-Iex);
  E2(k)=abs(QuadSimpson(f,a,b,N)-Iex);
  k=k+1;
end
loglog(H,E1,'b',H,E2,'r',H,H.^2,'k-.s', ...
       H,0.01*H.^4,'k-.d')
xlabel('h');ylabel('E(h)')
legend('Trapeze','Simpson','O(h^2)','O(h^4)')
```

Listing 5.2: Ordre des méthodes composites des trapèzes et de Simpson

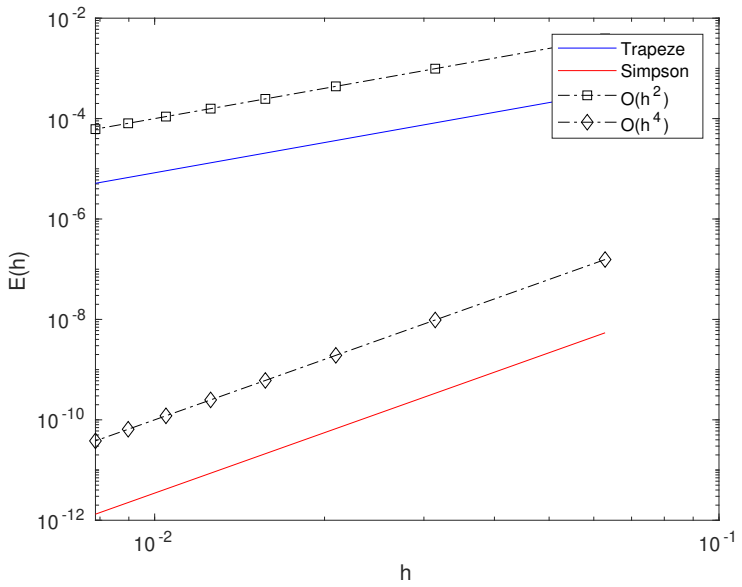


Figure 5.13: Ordre des méthodes composites des Trapèzes et de Simpson pour le calcul de $\int_0^{\pi/2} \cos(x)dx$

On donne en Listing 5.3, un exemple pour le calcul approché de $\int_0^1 x^{3/2}dx$ dont les résultats posent question.

Listing 5.3: : Exemple 1, script Matlab/Octave pour illustrer des changements de comportements des ordres

```
f=@(x) x.^(3/2);
F=@(x) x.^(3/2+1)/(3/2+1);
a=0;b=1;
Iex=F(b)-F(a);
I1=QuadTrapeze(f,a,b,10);
I2=QuadTrapeze(f,a,b,100);
fprintf('Erreurs␣Trapeze:␣(N=10)␣%.5e␣-␣(N=100)␣%.5e\n',abs(I1-Iex),abs(I2-Iex))
I1=QuadSimpson(f,a,b,10);
I2=QuadSimpson(f,a,b,100);
fprintf('Erreurs␣Simpson:␣(N=10)␣%.5e␣-␣(N=100)␣%.5e\n',abs(I1-Iex),abs(I2-Iex))
```

Output

```
Erreurs Trapeze: (N=10) 1.16946e-03 - (N=100) 1.22452e-05
Erreurs Simpson: (N=10) 7.85521e-06 - (N=100) 2.48802e-08
```

On ne retrouve le *bon* ordre pour la méthode de Simpson car la fonction ne vérifie par les hypothèses du Théorème 5.18: la dérivée seconde de la fonction $x \mapsto x^{3/2}$ n'est pas définie en 0.

On donne en Listing 5.4, un autre exemple pour le calcul approché de $\int_{-1}^1 |x|dx$ dont les résultats posent question.

Listing 5.4: : Exemple 2, script Matlab/Octave pour illustrer des changements de comportements des erreurs

```
f=@(x) abs(x);
a=-1;b=1;
Iex=1;
I1=QuadTrapeze(f,a,b,10);
I2=QuadTrapeze(f,a,b,11);
I3=QuadTrapeze(f,a,b,12);
I4=QuadTrapeze(f,a,b,13);
fprintf('Erreurs␣Trapeze:␣(N=10)␣%.5e␣-␣(N=11)␣%.5e␣-␣(N=12)␣%.5e␣-(N=13)␣...
    %.5e\n',abs(I1-Iex),abs(I2-Iex),abs(I3-Iex),abs(I4-Iex))
I1=QuadSimpson(f,a,b,10);
I2=QuadSimpson(f,a,b,11);
I3=QuadSimpson(f,a,b,12);
I4=QuadSimpson(f,a,b,13);
fprintf('Erreurs␣Simpson:␣(N=10)␣%.5e␣-␣(N=11)␣%.5e␣-␣(N=12)␣%.5e␣-(N=13)␣...
    %.5e\n',abs(I1-Iex),abs(I2-Iex),abs(I3-Iex),abs(I4-Iex))
```

Output

```
Erreurs Trapeze: (N=10) 0.00000e+00 - (N=11) 8.26446e-03 - (N=12) 0.00000e+00 -(N=13) 5.91716e-03
Erreurs Simpson: (N=10) 0.00000e+00 - (N=11) 2.75482e-03 - (N=12) 0.00000e+00 -(N=13) 1.97239e-03
```

Cette fois la fonction n'est pas dérivable en 0! Le Théorème 5.18 ne peut s'appliquer. Toutefois, les deux formules donnent un résultat exacte pour N pair. En effet, dans ce cas le point 0 est un point de discrétisation des méthodes composées et dans ce cas on est ramené à calculer numériquement les deux intégrales $\int_{-1}^{0}(-x)dx$ et $\int_{0}^{1}(x)dx$ ce qui donnera des résultats exactes puisque les 2 méthodes sont exactes pour les polynômes de degré 1.

5.3 Intégrales multiples

On veut approcher, en utilisant la formule de Simpson, l'intégrale

$$I = \int_a^b \int_c^d f(x,y)dydx$$

Par utilisation de la formule de quadrature de Simpson (5.13) en y on a

$$g(x) = \int_c^d f(x,y)dy \approx \tilde{g}(x) = \frac{d-c}{6}\left(f(x,c) + 4f(x,\frac{c+d}{2}) + f(x,d)\right).$$

Une nouvelle utilisation de Simpson en x donne

$$\begin{aligned} I = \int_a^b g(x)dx &\approx \frac{b-a}{6}\left(g(a) + 4g(\frac{a+b}{2}) + g(b)\right) \\ &\approx \frac{b-a}{6}\left(\tilde{g}(a) + 4\tilde{g}(\frac{a+b}{2}) + \tilde{g}(b)\right) \end{aligned}$$

En posant $\alpha = \frac{a+b}{2}$ et $\beta = \frac{c+d}{2}$, on obtient la formule de quadrature de Simpson "2D" :

$$I \approx \frac{b-a}{6}\frac{d-c}{6}\begin{pmatrix} f(a,c) + 4f(a,\beta) + f(a,d) \\ +4(f(\alpha,c) + 4f(\alpha,\beta) + f(\alpha,d)) \\ +f(b,c) + 4f(b,\beta) + f(b,d) \end{pmatrix} \tag{5.31}$$

La méthodologie pour obtenir la formule composite de Simpson "2D" est la suivante

1. Discrétisation régulière de $[a,b]$: $\forall k \in [\![0,n]\!]$, $x_k = a + kh_x$ avec $h_x = \frac{b-a}{n}$.
2. Discrétisation régulière de $[c,d]$: $\forall l \in [\![0,m]\!]$, $y_l = a + lh_y$ avec $h_y = \frac{d-c}{m}$.
3. Relation de Chasles :
$$\int_a^b \int_c^d f(x,y)dydx = \sum_{k=1}^{n}\sum_{l=1}^{m}\int_{x_{k-1}}^{x_k}\int_{y_{l-1}}^{y_l} f(x,y)dydx.$$
4. Formule composite de Simpson "2D" :
$$\begin{gathered} \int_a^b \int_c^d f(x,y)dydx \\ = \\ \frac{h_x h_y}{36}\sum_{k=1}^{n}\sum_{l=1}^{m}\begin{pmatrix} f(x_{k-1},y_{l-1}) + 4f(x_{k-1},\beta_l) + f(x_{k-1},y_l) \\ +4(f(\alpha_k,y_{l-1}) + 4f(\alpha_k,\beta_l) + f(\alpha_k,y_l)) \\ +f(x_k,y_{l-1}) + 4f(x_k,\beta_l) + f(x_k,y_l) \end{pmatrix} \end{gathered}$$

avec $\alpha_k = \frac{x_{k-1}+x_k}{2}$ et $\beta_l = \frac{y_{l-1}+y_l}{2}$.

Chapitre A

Langage algorithmique

A.1 Pseudo-langage algorithmique

Pour uniformiser l'écriture des algorithmes nous employons un pseudo-langage contenant l'indispensable :

- variables,
- opérateurs (arithmétiques, relationnels, logiques),
- expressions,
- instructions (simples et composées),
- fonctions.

Ce pseudo-langage sera de fait très proche du langage de programmation de Matlab.

A.1.1 Données et constantes

Une donnée est une valeur introduite par l'utilisateur (par ex. une température, une vitesse, ...). Une constante est un symbole ou un identificateur non modifiable (par ex. π, la constante de gravitation,...).

A.1.2 Variables

Definition A.1

Une variable est un objet dont la valeur est modifiable, qui possède un nom et un type (entier, charactère, réel, complexe, ...). Elle est rangée en mémoire à partir d'une certaine adresse.

A.1.3 Opérateurs

Opérateurs arithmétiques

Nom	Symbole	example
addition	$+$	$a+b$
soustraction	$-$	$a-b$
opposé	$-$	$-a$
produit	$*$	$a*b$
division	$/$	a/b
puissance a^b	^	$a\hat{}b$

Table A.1: Opérateurs arithmétiques

Opérateurs relationnels

Nom	Symbole	example	Commentaires
identique	$==$	$a == b$	vrai si a et b ont même valeur, faux sinon.
différent	$\sim=$	$a \sim= b$	faux si a et b ont même valeur, vrai sinon.
inférieur	$<$	$a < b$	vrai si a est plus petit que b, faux sinon.
supérieur	$>$	$a > b$	vrai si a est plus grand que b, faux sinon.
inférieur ou égal	$<=$	$a <= b$	vrai si a est plus petit ou égal à b, faux sinon.
supérieur ou égal	$>=$	$a >= b$	vrai si a est plus grand ou égal à b, faux sinon.

Table A.2: Opérateurs relationnels

Opérateurs logiques

Nom	Symbole	example	Commentaires
négation	$\sim$	$\sim a$	vrai si a est faux (ou nul), faux sinon.
ou	$\vert$	$a\vert b$	vrai si a **ou** b est vrai (non nul), faux sinon.
et	&	$a\&b$	vrai si a **et** b sont vrais (non nul), faux sinon.

Table A.3: Opérateurs logiques

Opérateur d'affectation

Nom	Symbole	example	Commentaires
affectation	$\leftarrow$	$a \leftarrow b$	On affecte à la variable a le contenu de b

Table A.4: Opérateurs d'affectation

A.1.4 Expressions

Definition A.2

Une expression est un groupe d'opérandes (i.e. nombres, constantes, variables, ...) liées par certains opérateurs pour former un terme algébrique qui représente une valeur (i.e. un élément de donnée simple)

Exemple A.3

- Voici un exemple classique d'expression numérique :

$$(b*b-4*a*c)/(2*a).$$

On appelle **opérandes** les identifiants a, b et c, et les nombres 4 et 2. Les symboles $*$, $-$ et $/$ sont les **opérateurs**.

- Voici un exemple classique d'expression booléenne (logique) :

$$(x < 3.14)$$

On teste $(x < 3.14)$ avec x une variable numérique réelle. si $x < 3.14$ alors cette expression renverra la valeur *vraie* (i.e. 1), *faux* sinon (i.e. 0).

A.1.5 Instructions

Definition A.4

Une **instruction** est un ordre ou un groupe d'ordres qui déclenche l'exécution de certaines actions par l'ordinateur. Il y a deux types d'instructions : simple et structuré.
Les **instructions simples** sont essentiellement des ordres seuls et inconditionnels réalisant l'une des tâches suivantes :

1. affectation d'une valeur a une variable.
2. appel d'une fonction (procedure, subroutine, ... suivant les langages).

Les **instructions structurées** sont essentiellement :

1. les instructions composées, groupe de plusieurs instructions simples,
2. les instructions répétitives, permettant l'exécution répétée d'instructions simples, (i.e. boucles «pour», «tant que»)
3. les instructions conditionnelles, lesquelles ne sont exécutées que si une certaine condition est respectée (i.e. «si»)

Les exemples qui suivent sont écrits dans un pseudo langage algorithmique mais sont facilement transposable dans la plupart des langages de programmation.

Instructions simples

Voici un exemple de l'*instruction simple d'affectation* :

1: $a \leftarrow 3.14 * R$

On évalue l'expression $3.14 * R$ et affecte le résultat à la variable a.
Un autre exemple est donné par l'*instruction simple d'affichage* :

```
affiche('bonjour')
```

Affiche la chaine de caractères `'bonjour'` à l'écran. Cette instruction fait appel à la fonction `affiche`.

Instructions composées

Instructions répétitives, boucle «pour»

Algorithme A.1 Exemple de boucle «pour»

Données : n : un entier.

1: $S \leftarrow 0$
2: **Pour** $i \leftarrow 1$ **à** n **faire**
3: $\quad S \leftarrow S + \cos(i^2)$
4: **Fin Pour**

Instruction répétitive, boucle «tant que»

Algorithme A.2 Exemple de boucle «tant que»

1: $i \leftarrow 0,\ x \leftarrow 1$
2: **Tantque** $i < 1000$ **faire**
3: $\quad x \leftarrow x + i * i$
4: $\quad i \leftarrow i + 1$
5: **Fin Tantque**

Instruction répétitive, boucle «répéter ...jusqu'à»

Algorithme A.3 Exemple de boucle «répéter ...jusqu'à»

1: $i \leftarrow 0,\ x \leftarrow 1$
2: **Répéter**
3: $\quad x \leftarrow x + i * i$
4: $\quad i \leftarrow i + 1$
5: **jusqu'à** i>=1000

Instructions conditionnelles «si»

Algorithme A.4 Exemple d'instructions conditionnelle «si»

Données : *age* : un réel.

1: **Si** *age* >= 18 **alors**
2: affiche('majeur')
3: **Sinon Si** *age* >= 0 **alors**
4: affiche('mineur')
5: **Sinon**
6: affiche('en devenir')
7: **Fin Si**

A.1.6 Fonctions

Les fonctions permettent

- d'automatiser certaines tâches répétitives au sein d'un même programme,
- d'ajouter à la clarté d'un programme,
- l'utilisation de portion de code dans un autre programme,
- ...

Fonctions prédéfinies

Pour faciliter leur usage, tous les langages de programmation possèdent des fonctions prédéfinies. On pourra donc supposer que dans notre langage algorithmique un grand nombre de fonctions soient prédéfinies : par exemple, les fonctions mathématiques sin, cos, exp, abs, $\cdots$ (pour ne citer qu'elles)

Syntaxe

On utilise la syntaxe suivante pour la définition d'une fonction

Fonction $[args_1, \ldots, args_n] \leftarrow$ NomFonction ($arge_1, \ldots, arge_m$)
instructions
Fin Fonction

La fonction se nomme NomFonction . Elle admet comme paramètres d'entrée (données) les m arguments $arge_1, \ldots, arge_m$ et comme paramètres de sortie (résultats) les n arguments $args_1, \ldots, args_n$. Ces derniers doivent être déterminés dans le corps de la fonction (partie instructions).
Dans le cas ou la fonction n'admet qu'un seul paramètre de sortie, l'écriture se simplifie :

Fonction $args \leftarrow$ NomFonction ($arge_1, \ldots, arge_m$)
instructions
Fin Fonction

Ecrire ses propres fonctions

Pour écrire une fonction «propre», il faut tout d'abord déterminer exactement ce que devra faire cette fonction.

Puis, il faut pouvoir répondre à quelques questions :

1. Quelles sont les données (avec leurs limitations)?
2. Que doit-on calculer ?

Et, ensuite il faut la **commenter** : expliquer son usage, type des paramètres, ...

Exemple : résolution d'une équation du premier degré

Nous voulons écrire une fonction calculant la solution de l'équation

$$ax + b = 0,$$

où nous supposons que $a \in \mathbb{R}^*$ et $b \in \mathbb{R}$. La solution de ce problème est donc

$$x = -\frac{b}{a}.$$

Les données de cette fonction sont $a \in \mathbb{R}^*$ et $b \in \mathbb{R}$. Elle doit retourner $x = -\frac{b}{a}$ solution de $ax + b = 0$.

Algorithme A.5 Exemple de fonction : Résolution de l'équation du premier degré $ax + b = 0$.

Données : a : nombre réel différent de 0
b : nombre réel.
Résultat : x : un réel.

1: **Fonction** $x \leftarrow$ REPD(a, b)
2: $x \leftarrow -b/a$
3: **Fin Fonction**

Remarque A.5 Cette fonction est très simple, toutefois pour ne pas «alourdir» le code nous n'avons pas vérifié la validité des données fournies.

Exercice A.1.1

Ecrire un algorithme permettant de valider cette fonction.

Exemple : résolution d'une équation du second degré

Nous cherchons les solutions réelles de l'équation

$$ax^2 + bx + c = 0, \tag{A.1}$$

où nous supposons que $a \in \mathbb{R}^*$, $b \in \mathbb{R}$ et $c \in \mathbb{R}$ sont donnés.

Mathématiquement, l'étude des solutions réelles de cette équation nous amène à envisager trois cas suivant les valeurs du discriminant $\Delta = b^2 - 4ac$

- si $\Delta < 0$ alors les deux solutions sont complexes,
- si $\Delta = 0$ alors la solution est $x = -\frac{b}{2a}$,
- si $\Delta > 0$ alors les deux solutions sont $x_1 = \frac{-b-\sqrt{\Delta}}{2*a}$ et $x_2 = \frac{-b-\sqrt{\Delta}}{2*a}$.

Exercice A.1.2

1. Ecrire la fonction `discriminant` permettant de calculer le discriminant de l'équation (A.1).
2. Ecrire la fonction `RESD` permettant de résoudre l'équation (A.1) en utilisant la fonction `discriminant`.
3. Ecrire un programme permettant de valider ces deux fonctions.

Exercice A.1.3

Même question que précédemment dans le cas complexe (solutions et coefficients).

A.2 Méthodologie d'élaboration d'un algorithme

A.2.1 Description du problème

- Spécification d'un ensemble de données
  Origine : énoncé, hypothèses, sources externes, ...
- Spécification d'un ensemble de buts à atteindre
  Origine : résultats, opérations à effectuer, ...
- Spécification des contraintes

A.2.2 Recherche d'une méthode de résolution

- Clarifier l'énoncé.
- Simplifier le problème.
- Ne pas chercher à le traiter directement dans sa globalité.
- S'assurer que le problème est soluble (sinon problème d'indécidabilité!)
- Recherche d'une stratégie de construction de l'algorithme
- Décomposer le problème en sous problèmes partiels plus simples : raffinement.
- Effectuer des raffinements successifs.
- Le niveau de raffinement le plus élémentaire est celui des instructions.

A.2.3 Réalisation d'un algorithme

Il doit être conçu indépendamment du langage de programmation et du système informatique (sauf cas très particulier)

- L'algorithme doit être exécuté en un nombre fini d'opérations.
- L'algorithme doit être spécifié clairement, sans la moindre ambiguïté.
- Le type de données doit être précisé.
- L'algorithme doit fournir au moins un résultat.
- L'algorithme doit être effectif : toutes les opérations doivent pouvoir être simulées par un homme en temps fini.

Pour écrire un algorithme détaillé, il faut tout d'abord savoir répondre a quelques questions :

- Que doit-il faire ? (i.e. Quel problème est-il censé résoudre?)
- Quelles sont les données necessaires à la résolution de ce problème?
- Comment résoudre ce problème «à la main» (sur papier)?

Si l'on ne sait pas répondre à l'une de ces questions, l'écriture de l'algorithme est fortement compromise.

A.2.4 Exercices

Exercice A.2.1: Algorithme pour une somme

Ecrire un algorithme permettant de calculer

$$S(x) = \sum_{k=1}^{n} k \sin(2 * k * x)$$

Correction Exercice A.2.1 L'énoncé de cet exercice est imprécis. On choisit alors $x \in \mathbb{R}$ et $n \in \mathbb{N}$ pour rendre possible le calcul. Le problème est donc de calculer

$$\sum_{k=1}^{n} k \sin(2kx).$$

Toutefois, on aurait pu choisir $x \in \mathbb{C}$ ou encore un tout autre problème :

$$\text{Trouver } x \in \mathbb{R} \text{ tel que } S(x) = \sum_{k=1}^{n} k \sin(2kx)$$

où $n \in \mathbb{N}$ et S, fonction de $\mathbb{R}$ à valeurs réelles, sont les données!

Algorithme A.6 Calcul de $S = \sum_{k=1}^{n} k \sin(2kx)$

Données : x : nombre réel,
n : nombre entier.
Résultat : S : un réel.

1: $S \leftarrow 0$
2: **Pour** $k \leftarrow 1$ **à** n **faire**
3: $S \leftarrow S + k * \sin(2 * k * x)$
4: **Fin Pour**

◇

Exercice A.2.2: Algorithme pour un produit

Ecrire un algorithme permettant de calculer

$$P(z) = \prod_{n=1}^{k} sin(2 * k * z/n)^k$$

Correction Exercice A.2.2 L'énoncé de cet exercice est imprécis. On choisit alors $z \in \mathbb{R}$ et $k \in \mathbb{N}$ pour rendre possible le calcul.

Algorithme A.7 Calcul de $P = \prod_{n=1}^{k} \sin(2kz/n)^k$

Données : z : nombre réel,
k : nombre entier.
Résultat : P : un réel.

```
1: P ← 1
2: Pour n ← 1 à k faire
3:    P ← P * sin(2 * k * z/n)^k
4: Fin Pour
```

◇

Exercice A.2.3: Série de Fourier

Soit la série de Fourier

$$x(t) = \frac{4A}{\pi}\left\{\cos\omega t - \frac{1}{3}\cos 3\omega t + \frac{1}{5}\cos 5\omega t - \frac{1}{7}\cos 7\omega t + \cdots\right\}.$$

Ecrire la fonction SFT permettant de calculer $x_n(t)$.

Correction Exercice A.2.3 Nous devons écrire la fonction permettant de calculer

$$x_n(t) = \frac{4A}{\pi}\sum_{k=1}^{n}(-1)^{k+1}\frac{1}{2k-1}\cos((2k-1)\omega t)$$

Les données de la fonction sont $A \in \mathbb{R}$, $\omega \in \mathbb{R}$, $n \in \mathbb{N}^*$ et $t \in \mathbb{R}$.

Grâce à ces renseignements nous pouvons déjà écrire l'entête de la fonction :

Algorithme A.8 En-tête de la fonction SFT retournant valeur de la série de Fourier en t tronquée au n premiers termes de l'exercice A.2.3.

Données : t : nombre réel,
n : nombre entier strictement positif
$A,\ \omega$: deux nombres réels.
Résultat : x : un réel.

```
1: Fonction x ← SFT( t, n, A, ω )
2:    ...
3: Fin Fonction
```

Maintenant nous pouvons écrire progressivement l'algorithme pour aboutir au final à une version ne contenant que des opérations élémentaires.

Finalement la fonction est

Algorithme A.9 Fonction SFT retournant la valeur de la série de Fourier en t tronquée au n premiers termes de l'exercice A.2.3.

Données : t : nombre réel,
n : nombre entier strictement positif
$A,\ \omega$: deux nombres réels.
Résultat : x : un réel.

1: **Fonction** $x \leftarrow$ SFT(t, n, A, ω)
2: $S \leftarrow 0$
3: **Pour** $k = 1$ **à** n **faire**
4: $S \leftarrow S + ((-1)\hat{}(k+1)) * \cos((2*k-1)*\omega*t)/(2*k-1)$
5: **Fin Pour**
6: $S \leftarrow 4 * A * S/\pi$
7: **Fin Fonction**

◇

Exercice A.2.4

Reprendre les trois exercices précédents en utilisant les boucles «tant que».

A.3 Principes de «bonne» programmation pour attaquer de «gros» problèmes

Tous les exemples vus sont assez courts. Cependant, il peut arriver que l'on ait des programmes plus longs à écrire (milliers de lignes, voir des dizaines de milliers de lignes). Dans l'industrie, il arrive que des équipes produisent des codes de millions de lignes, dont certains mettent en jeux des vies humaines (contrôler un avion de ligne, une centrale nucléaire, ...). Le problème est évidemment d'écrire des programmes sûrs. Or, *un programme à* 100% *sûr, cela n'existe pas!* Cependant, plus un programme est simple, moins le risque d'erreur est grand : c'est sur cette remarque de bon sens que se basent les «bonnes» méthodes. Ainsi :

Tout problème compliqué doit être découpé en sous-problèmes plus simples

Il s'agit, lorsqu'on a un problème P à résoudre, de l'analyser et de le décomposer en un ensemble de problèmes $P_1,\ P_2,\ P_3, \ldots$ plus simples. Puis, P_1, est lui-même analysé et décomposé en $P_{11}, P_{12}, \ldots$, et P_2 en P_{21}, P_{22}, etc. On poursuit cette analyse jusqu'à ce qu'on n'ait plus que des problèmes élémentaires à résoudre. Chacun de ces problèmes élémentaires est donc traité séparément dans un module, c'est à dire un morceau de programme relativement indépendant du reste. Chaque module sera *testé et validé* séparément dans la mesure du possible et naturellement *largement documenté*. Enfin ces modules élémentaires sont assemblés en modules de plus en plus complexes, jusqu'à remonter au problème initiale. A chaque niveau, il sera important de bien réaliser les phases de test, validation et documentation des modules.

Dans ce document, on s'évertue à écrire des algorithmes!
Ceux-ci ne seront pas optimisés[a]!

[a] améliorés pour minimiser le nombre d'opérations élémentaires, l'occupation mémoire, ..

Chapitre B

Annexes

B.1 Analyse : rappels

B.1.1 En vrac

Théorème B.1: Théorème de Bolzano ou des valeurs intermédiaires

Soit $f : [a, b] \subset \mathbb{R} \longrightarrow \mathbb{R}$ une application continue. Si $f(a)$ et $f(b)$ ne sont pas de même signe (i.e. $f(a)f(b) < 0$) alors il existe au moins $c \in]a, b[$ tel que $f(c) = 0$.

Théorème B.2: Théorème des accroissements finis

Soient a et b deux réels, $a < b$ et f une fonction continue sur l'intervalle fermé $[a, b]$, dérivable sur l'intervalle ouvert $]a, b[$. Alors il existe $\xi \in]a, b[$ tel que

$$f'(\xi) = \frac{f(b) - f(a)}{b - a}.$$

Proposition B.3: Formule de Taylor-Lagrange

Soit $n \in N^*$ et $f \in \mathcal{C}^n([a, b])$ dont la dérivée n-ième est dérivable. Alors pour tout x, y dans $[a, b]$, $x \neq y$, il existe $\xi \in] \min(x, y), \max(x, y)[$ tel que

$$f(x) = f(y) + \sum_{k=1}^{n} \frac{(x - y)^k}{k!} f^{(k)}(y) + \frac{(x - y)^{n+1}}{(n + 1)!} f^{(n+1)}(\xi) \tag{B.1}$$

Corollaire B.4: Théorème de la bijection

Si f est une fonction continue et strictement monotone sur un intervalle $[a, b]$ et à valeurs réelles,

alors elle constitue une bijection entre $[a, b]$ et l'intervalle fermé dont les bornes sont $f(a)$ et $f(b)$.

Preuve. Notons $J = f^{-1}([a, b])$ cet intervalle fermé, c'est-à-dire l'ensemble des réels compris entre $f(a)$ et $f(b)$.

- La monotonie de la fonction implique que l'image de l'intervalle $[a, b]$ est contenue dans J :
  - si f est croissante, pour tout $x \in [a, b]$ on a $f(a) \leqslant f(x) \leqslant f(b)$
  - si f est décroissante, pour tout $x \in [a, b]$ on a $f(b) \leqslant f(x) \leqslant f(a)$.
- Le fait que cette monotonie soit stricte assure que deux réels distincts ne peuvent avoir la même image, autrement dit la fonction est injective sur $[a, b]$.
- Enfin, le théorème des valeurs intermédiaires (qui s'appuie sur l'hypothèse de continuité) garantit que tout élément de J admet au moins un antécédent par f , c'est-à-dire que la fonction est surjective dans J.

□

Proposition B.5

Soit f est une fonction bijective continue d'un intervalle ouvert $I \subset \mathbb{R}$ sur un intervalle ouvert $J \subset \mathbb{R}$. Si f est dérivable en $\alpha \in I$ et que $f'(\alpha) \neq 0$ alors sa réciproque f^{-1} est dérivable en $\beta = f(\alpha) \in J$ et

$$(f^{-1})'(\beta) = \frac{1}{f'(\alpha)} \quad \text{ou encore} \quad (f^{-1})'(\beta) = \frac{1}{f'(f^{-1}(\beta))}$$

Preuve. On pose $g = f^{-1}$ et on écrit son taux d'accroissement :

$$\frac{g(y) - g(\beta)}{y - \beta} = \frac{x - \alpha}{f(x) - f(\alpha)}$$

avec $y = f(x)$. Cette fraction est l'inverse de $\frac{f(x)-f(\alpha)}{x-\alpha}$ qui tend vers $f'(\alpha) \neq 0$ quand x tend vers α. □

B.1.2 Espace métrique

Definition B.6: Distance sur un ensemble

On appelle **distance** sur un ensemble E, une application d de E^2 dans $\mathbb{R}^+$ telle que pour tout $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in E^3$ on a

- symétrie : $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$,
- séparation : $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$,
- inégalité triangulaire : $d(\boldsymbol{x}, \boldsymbol{z}) \leqslant d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$

Voici quelques exemples de distances:

- $\mathrm{d}(x, y) = |x - y|$ dans $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Z}$ ou $\mathbb{Q}$
- $\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$ dans $\mathbb{R}^n$, où $\|.\|$ est l'une quelconque des normes habituelles.

Definition B.7: Espace métrique

On appelle (E, d) un **espace métrique** si E est un ensemble et d une distance sur E.

Definition B.8: Suite convergente

Soient (E, d) un **espace métrique** et $(\boldsymbol{u}^{[k]})_{k\in\mathbb{N}}$ une suite d'éléments de E. On dit que la suite $(\boldsymbol{u}^{[k]})_{k\in\mathbb{N}}$ converge vers $\boldsymbol{\alpha} \in E$ si

$$\forall \epsilon > 0,\ \exists N \in \mathbb{N} \ \text{ tel que } \ \forall k > N, \quad \mathrm{d}(\boldsymbol{u}^{[k]}, \boldsymbol{\alpha}) < \epsilon. \tag{B.2}$$

Definition B.9: Ordre de convergence

Soient (E, d) un **espace métrique** et $(\boldsymbol{u}^{[k]})_{k\in\mathbb{N}}$ une suite d'éléments de E convergeant vers $\boldsymbol{\alpha} \in E$. On dit que cette suite **converge vers $\boldsymbol{\alpha}$ avec un ordre** $p \geqslant 1$ si

$$\exists k_0 \in \mathbb{N},\ \exists C > 0 \ \text{ tels que } \ \mathrm{d}(\boldsymbol{u}^{[k+1]}, \boldsymbol{\alpha}) \leqslant C\, \mathrm{d}(\boldsymbol{u}^{[k]}, \boldsymbol{\alpha})^p,\ \forall k \geqslant k_0. \tag{B.3}$$

où $C < 1$ si $p = 1$.

Definition B.10: Suite de Cauchy

Soit (E, d) un **espace métrique**. Une suite $(\boldsymbol{x}^{[k]})_{k\in\mathbb{N}}$ d'éléments de E est dites **de Cauchy** si

$$\forall \epsilon > 0,\ \exists M \in \mathbb{N},\ \text{tel que } \forall (p, q) \in \mathbb{N}^2,\ p, q \geqslant M, \quad \mathrm{d}(\boldsymbol{x}^{[p]}, \boldsymbol{x}^{[q]}) < \epsilon.$$

ce qui correspond à

$$\lim_{m\to+\infty} \sup_{p,q\geqslant m} \mathrm{d}(\boldsymbol{x}^{[p]}, \boldsymbol{x}^{[q]}) = 0.$$

Une autre manière de l'écrire est

$$\forall \epsilon > 0,\ \exists M \in \mathbb{N},\ \text{tel que } \forall k \in \mathbb{N},\ k \geqslant M,\ \forall l \in \mathbb{N}, \quad \mathrm{d}(\boldsymbol{x}^{[k+l]}, \boldsymbol{x}^{[k]}) < \epsilon.$$

ce qui correspond à

$$\lim_{m\to+\infty} \sup_{k\geqslant m, l\geqslant 0} \mathrm{d}(\boldsymbol{x}^{[k+l]}, \boldsymbol{x}^{[k]}) = 0.$$

Definition B.11: Espace métrique complet

Un espace metrique est dit **complet** si toute suite de Cauchy converge.

Proposition B.12

Si E est un espace vectoriel normé de norme $\|.\|$ alors E est un espace métrique pour la distance d issue de sa norme et définie par $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|$, $\forall (\boldsymbol{x}, \boldsymbol{y}) \in E^2$.

Definition B.13: Espace de Banach

On appelle **espace de Banach** un espace vectoriel normé complet pour la distance issue de sa norme.

B.2 Algèbre linéaire

Toute cette partie peut être joyeusement omise par tout Homo sapiens *algebra linearis* compatible. Toutefois une lecture rapide permet de se raffraichir la mémoire.

Soit V un **espace vectoriel** de dimension finie n, sur le corps $\mathbb{R}$ des nombres réels, ou sur le corps $\mathbb{C}$ des nombres complexes. Notons plus généralement $\mathbb{K}$ le corps $\mathbb{R}$ ou $\mathbb{C}$.

B.2.1 Vecteurs

Une **base** de V est un ensemble $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$ de n **vecteurs linéairement indépendants**. Le vecteur $\boldsymbol{v} = \sum_{i=1}^{n} v_i \boldsymbol{e}_i$ sera représenté par le **vecteur colonne**

$$\boldsymbol{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

et on désignera par $\boldsymbol{v}^{\mathsf{t}}$ et $\boldsymbol{v}^*$ les **vecteurs lignes** suivants

$$\boldsymbol{v}^{\mathsf{t}} = \begin{pmatrix} v_1 & v_2 & \cdots & v_n \end{pmatrix}, \quad \boldsymbol{v}^* = \begin{pmatrix} \overline{v_1} & \overline{v_2} & \cdots & \overline{v_n} \end{pmatrix}$$

où $\overline{\alpha}$ est le nombre **complexe conjugué** du nombre α.

Definition B.14

- Le vecteur ligne $\boldsymbol{v}^{\mathsf{t}}$ est le **vecteur transposé** du vecteur colonne $\boldsymbol{v}$.
- Le vecteur ligne $\boldsymbol{v}^*$ est le **vecteur adjoint** du vecteur colonne $\boldsymbol{v}$.

Definition B.15

L'application $\langle \bullet, \bullet \rangle : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}$ définie pour tout $(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{K}^n \times \mathbb{K}^n$ par

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^{\mathsf{t}}.\boldsymbol{v} = \boldsymbol{v}^{\mathsf{t}}.\boldsymbol{u} = \sum_{i=1}^{n} u_i v_i, \quad \text{si} \quad \mathbb{K} = \mathbb{R} \tag{B.4}$$

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^*.\boldsymbol{v} = \overline{\boldsymbol{v}^*.\boldsymbol{u}} = \overline{\langle \boldsymbol{v}, \boldsymbol{u} \rangle} = \sum_{i=1}^{n} \overline{u_i} v_i, \quad \text{si} \quad \mathbb{K} = \mathbb{C} \tag{B.5}$$

est appelée **produit scalaire** euclidien si $\mathbb{K} = \mathbb{R}$, hermitien[a] si $\mathbb{K} = \mathbb{C}$. Pour rappeler la dimension de l'espace, on écrit

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \langle \boldsymbol{u}, \boldsymbol{v} \rangle_n .$$

[a] La convention choisie pour le produit scalaire hermitien étant ici : linéarité à droite et semi-linéarité à gauche. Il est aussi possible de définir le produit scalaire hermitien par le complexe conjugué de (B.5) :

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{v}^*.\boldsymbol{u} = \sum_{i=1}^{n} u_i \overline{v_i}.$$

Dans ce cas le produit scalaire est une forme sesquilinéaire à droite.

Definition B.16

Soit V est un espace vectoriel muni d'un produit scalaire.

◇ Deux **vecteurs** $\boldsymbol{u}$ et $\boldsymbol{v}$ sont **orthogonaux** si $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$.

◇ Un **vecteur** $\boldsymbol{v}$ est **orthogonal à une partie** U de V si

$$\forall \boldsymbol{u} \in U, \ \langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0.$$

On note $\boldsymbol{v} \perp U$.

◇ Un ensemble de vecteurs $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\}$ de l'espace V est dit **orthonormal** si

$$\langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_{ij}, \ \forall (i,j) \in [\![1,k]\!]^2$$

où δ_{ij} est le **symbole de Kronecker** : $\delta_{ij} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{si } i \neq j. \end{cases}$

Definition B.17

Le vecteur nul de $\mathbb{K}^n$ est représenté par $\mathbf{0}_n$ ou $\mathbf{0}$ lorsqu'il n'y a pas d'ambiguité.

Definition B.18

Soit $\boldsymbol{u} \in \mathbb{K}^n$ non nul. On définit l'**opérateur de projection** sur $\boldsymbol{u}$ par

$$\text{proj}_{\boldsymbol{u}}(\boldsymbol{v}) = \frac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\langle \boldsymbol{u}, \boldsymbol{u} \rangle} \boldsymbol{u} = \frac{1}{\langle \boldsymbol{u}, \boldsymbol{u} \rangle} \boldsymbol{u}\boldsymbol{u}^* \boldsymbol{v}, \ \forall \boldsymbol{v} \in \mathbb{K}^n. \tag{B.6}$$

La matrice $\mathbb{P}_{\boldsymbol{u}} = \boldsymbol{u}\boldsymbol{u}^*$ s'appelle la matrice de la projection orthogonale suivant le vecteur $\boldsymbol{u}$.

Proposition B.19: Procédé de Gram-Schmidt

Soit $\{\boldsymbol{v}_i\}_{i \in [\![1,n]\!]}$ une base de $\mathbb{K}^n$. On construit successivement les vecteurs $\boldsymbol{u}_i$

$$\boldsymbol{u}_i = \boldsymbol{v}_i - \sum_{k=1}^{i-1} \text{proj}_{\boldsymbol{u}_k}(\boldsymbol{v}_i) = \boldsymbol{v}_i - \sum_{k=1}^{i-1} \frac{\langle \boldsymbol{u}_k, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{u}_k, \boldsymbol{u}_k \rangle} \boldsymbol{u}_k, \ \forall i \in [\![1,n]\!].$$

Ils forment une **base orthogonale** de $\mathbb{K}^n$ et $\text{Vect}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_i) = \text{Vect}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_i)$, $\forall i \in [\![1,n]\!]$ (voir Exercice B.3.5, page 215).
Pour construire une **base orthonormale** $\{\boldsymbol{z}_i\}_{i \in [\![1,n]\!]}$, il suffit de normaliser les vecteurs de la base orthogonale:

$$\boldsymbol{z}_i = \frac{\boldsymbol{u}_i}{\langle \boldsymbol{u}_i, \boldsymbol{u}_i \rangle^{1/2}}, \quad \forall i \in [\![1,n]\!].$$

B.2.2 Matrices

Généralités

Une matrice à m lignes et n colonnes est appelée *matrice de type* (m,n), et on note $\mathcal{M}_{m,n}(\mathbb{K})$, ou simplement $\mathcal{M}_{m,n}$, l'espace vectoriel sur le corps $\mathbb{K}$ formé par les matrices de type (m,n) à éléments dans $\mathbb{K}$.

Une matrice $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{K})$ d'éléments $\text{A}_{ij} \in \mathbb{K}$ est notée

$$\mathbb{A} = (\text{A}_{ij})_{1 \leqslant i \leqslant m, \ 1 \leqslant j \leqslant n},$$

le premier indice i correspond aux lignes et le second j aux colonnes. On désigne par $(\mathbb{A})_{ij}$ l'élément de la $i^{\text{ème}}$ ligne et de la $j^{\text{ème}}$ colonne. On peut aussi le noter $\text{A}_{i,j}$.

Definition B.20

La matrice nulle de $\mathcal{M}_{m,n}(\mathbb{K})$ est représentée par $\mathbb{O}_{m,n}$ ou $\mathbb{O}$ lorsqu'il n'y a pas d'ambiguité. Si $m = n$ on peut aussi noter $\mathbb{O}_n$ cette matrice.

Definition B.21

◇ Soit une matrice $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{C})$, on note $\mathbb{A}^* \in \mathcal{M}_{n,m}(\mathbb{C})$ la **matrice adjointe** de la matrice $\mathbb{A}$, définie de façon unique par

$$\langle \mathbb{A}\boldsymbol{u}, \boldsymbol{v} \rangle_m = \langle \boldsymbol{u}, \mathbb{A}^*\boldsymbol{v} \rangle_n , \ \forall \boldsymbol{u} \in \mathbb{C}^n, \ \forall \boldsymbol{v} \in \mathbb{C}^m$$

qui entraine $(\mathbb{A}^*)_{ij} = \overline{\mathrm{A}_{ji}}$.

◇ Soit une matrice $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{R})$, on note $\mathbb{A}^\mathsf{t} \in \mathcal{M}_{n,m}(\mathbb{R})$ la **matrice transposée** de la matrice $\mathbb{A}$, définie de façon unique par

$$\langle \mathbb{A}\boldsymbol{u}, \boldsymbol{v} \rangle_m = \langle \boldsymbol{u}, \mathbb{A}^\mathsf{t}\boldsymbol{v} \rangle_n , \ \forall \boldsymbol{u} \in \mathbb{R}^n, \ \forall \boldsymbol{v} \in \mathbb{R}^m$$

qui entraine $(\mathbb{A}^\mathsf{t})_{ij} = \mathrm{A}_{ji}$.

Definition B.22

Si $\mathbb{A} \in \mathcal{M}_{m,p}(\mathbb{K})$ et $\mathbb{B} \in \mathcal{M}_{p,n}(\mathbb{K})$, leur **produit** $\mathbb{AB} \in \mathcal{M}_{m,n}(\mathbb{K})$ est défini par

$$(\mathbb{AB})_{ij} = \sum_{k=1}^{p} \mathrm{A}_{ik}\mathrm{B}_{kj}, \forall i \in [\![1, m]\!], \ \forall j \in [\![1, n]\!]. \tag{B.7}$$

Exercice B.2.1: résultats à savoir

Soient $\mathbb{A} \in \mathcal{M}_{m,p}(\mathbb{K})$ et $\mathbb{B} \in \mathcal{M}_{p,n}(\mathbb{K})$, montrer que

$$(\mathbb{AB})^\mathsf{t} = \mathbb{B}^\mathsf{t}\mathbb{A}^\mathsf{t}, \quad \text{si } \mathbb{K} = \mathbb{R}, \tag{B.8}$$
$$(\mathbb{AB})^* = \mathbb{B}^*\mathbb{A}^*, \quad \text{si } \mathbb{K} = \mathbb{C} \tag{B.9}$$

Les matrices considérées jusqu'à la fin de ce paragraphe sont carrées.

Definition B.23

Si $\mathbb{A} \in \mathcal{M}_n$ alors les éléments $\mathrm{A}_{ii} = (\mathbb{A})_{ii}$ sont appelés **éléments diagonaux** et les éléments $\mathrm{A}_{ii} = (\mathbb{A})_{ij}, i \neq j$ sont appelés **éléments hors-diagonaux**.

Definition B.24

On appelle **matrice identitée** de $\mathcal{M}_n$ la matrice dont les éléments diagonaux sont tous égals à 1 et les éléments hors-diagonaux nulles. On la note $\mathbb{I}$ ou encore $\mathbb{I}_n$ et on a

$$(\mathbb{I})_{i,j} = \delta_{ij}, \ \forall (i, j) \in [\![1, n]\!]^2.$$

Definition B.25

Une matrice $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$ est **inversible** ou **régulière** s'il existe une matrice $\mathbb{B} \in \mathcal{M}_n(\mathbb{K})$ vérifiant

$$\mathbb{AB} = \mathbb{BA} = \mathbb{I} \tag{B.10}$$

Dans le cas contraire, on dit que la matrice $\mathbb{A}$ est **singulière** ou **non inversible**.

On peut noter que la matrice $\mathbb{B}$ est unique. En effet, soient $\mathbb{B}_1$ et $\mathbb{B}_2$ vérifiant (B.10). On a alors $\mathbb{AB}_2 = \mathbb{I}$ et donc $\mathbb{B}_1(\mathbb{AB}_2) = \mathbb{B}_1$. On a aussi $\mathbb{B}_1\mathbb{A} = \mathbb{I}$ et donc $(\mathbb{B}_1\mathbb{A})\mathbb{B}_2 = \mathbb{B}_2$. Le produit des matrices étant associatif on a $\mathbb{B}_1(\mathbb{AB}_2) = (\mathbb{B}_1\mathbb{A})\mathbb{B}_2$ et donc $\mathbb{B}_1 = \mathbb{B}_2$.

Definition B.26

Soit $\mathbb{A} \in \mathcal{M}_n$ une matrice inversible. On note $\mathbb{A}^{-1} \in \mathcal{M}_n$ l'unique matrice vérifiant

$$\mathbb{A}\mathbb{A}^{-1} = \mathbb{A}^{-1}\mathbb{A} = \mathbb{I}. \tag{B.11}$$

Cette matrice est appelée **matrice inverse** de $\mathbb{A}$.

Definition B.27

Soit $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{K})$

- On note $\ker(\mathbb{A}) = \{\boldsymbol{v} \in \mathbb{K}^n \ ; \ \mathbb{A}\boldsymbol{v} = 0\}$ le **noyau** de $\mathbb{A}$.
- On note $\mathrm{im}(\mathbb{A}) = \{\mathbb{A}\boldsymbol{v} \in \mathbb{K}^m \ ; \ \boldsymbol{v} \in \mathbb{K}^n\}$ l'**image** de $\mathbb{A}$.
- On note $\mathrm{rank}(\mathbb{A}) \overset{\text{def}}{=} \dim(\mathrm{im}(\mathbb{A}))$ le **rang** de $\mathbb{A}$.

Théorème B.28: (théorème du rang)

Soit $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{K})$. On a

$$\mathrm{rank}(\mathbb{A}) + \dim(\ker(\mathbb{A})) = n$$

Proposition B.29

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$. Les propriétés suivantes sont équivalentes

1. $\mathbb{A}$ est inversible,
2. $\mathrm{rank}(\mathbb{A}) = n$,
3. $\boldsymbol{x} \in \mathbb{K}^n$, $\mathbb{A}\boldsymbol{x} = 0 \ \Rightarrow \ \boldsymbol{x} = 0$, (i.e. $\ker \mathbb{A} = \{0\}$)
4. $\det(\mathbb{A}) \neq 0$,
5. toutes les valeurs propres de $\mathbb{A}$ sont non nulles,
6. il existe $\mathbb{B} \in \mathcal{M}_n(\mathbb{K})$ tel que $\mathbb{A}\mathbb{B} = \mathbb{I}$,
7. il existe $\mathbb{B} \in \mathcal{M}_n(\mathbb{K})$ tel que $\mathbb{B}\mathbb{A} = \mathbb{I}$.

Exercice B.2.2: ***résultats à savoir***

Soient $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$ et $\mathbb{B} \in \mathcal{M}_n(\mathbb{K})$ inversibles. Montrer que $\mathbb{A}\mathbb{B}$ inversible et

$$(\mathbb{A}^{\mathsf{t}})^{-1} = (\mathbb{A}^{-1})^{\mathsf{t}}, \quad \text{si } \mathbb{K} = \mathbb{R}, \tag{B.12}$$
$$(\mathbb{A}^{*})^{-1} = (\mathbb{A}^{-1})^{*}, \quad \text{si } \mathbb{K} = \mathbb{C}. \tag{B.13}$$
$$(\mathbb{A}\mathbb{B})^{-1} = \mathbb{B}^{-1}\mathbb{A}^{-1} \tag{B.14}$$
$$(\mathbb{A}^{-1})^{-1} = \mathbb{A} \tag{B.15}$$

Definition B.30

Une matrice **carrée** $\mathbb{A}$ est :

- **symétrique** si $\mathbb{A}$ est réelle et $\mathbb{A} = \mathbb{A}^{\mathsf{t}}$,
- **hermitienne** si $\mathbb{A} = \mathbb{A}^{*}$,

◇ **normale** si $\mathbb{A}\mathbb{A}^* = \mathbb{A}^*\mathbb{A}$,

◇ **orthogonale** si $\mathbb{A}$ est réelle et $\mathbb{A}\mathbb{A}^t = \mathbb{A}^t\mathbb{A} = \mathbb{I}$,

◇ **unitaire** si $\mathbb{A}\mathbb{A}^* = \mathbb{A}^*\mathbb{A} = \mathbb{I}$,

Proposition B.31

- une matrice symétrique ou hermitienne est nécessairement normale.
- une matrice orthogonale (resp. unitaire) est nécessairement normale et inversible d'inverse $\mathbb{A}^t$ (resp. $\mathbb{A}^*$).

Definition B.32

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice **hermitienne**.

◇ Elle est **définie positive** si

$$\langle \mathbb{A}\boldsymbol{u}, \boldsymbol{u} \rangle > 0, \ \forall \boldsymbol{u} \in \mathbb{C}^n \backslash \{0\} \tag{B.16}$$

◇ Elle est **semi définie positive** si

$$\langle \mathbb{A}\boldsymbol{u}, \boldsymbol{u} \rangle \geqslant 0, \ \forall \boldsymbol{u} \in \mathbb{C}^n \backslash \{0\} \tag{B.17}$$

Exercice B.2.3

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$.

Q. 1 *Que peut-on dire de la matrice $\mathbb{A}\mathbb{A}^*$? Et si la matrice $\mathbb{A}$ est inversible?*

Q. 2 *Proposer une technique permettant de générer une matrice hermitienne semi-définie positive à partir d'une matrice aléatoire quelconque.*

Q. 3 *Proposer une technique permettant de générer une matrice hermitienne définie positive à partir d'une matrice triangulaire inférieure inversible aléatoire.*

Definition B.33

Soit $\mathbb{A} \in \mathcal{M}_n$. La trace d'une matrice carrée $\mathbb{A} = (a_{ij})$ est définie par

$$\mathrm{tr}\,(\mathbb{A}) = \sum_{i=1}^{n} a_{ii}.$$

Definition B.34

Soit $\mathcal{T}_n$ le **groupe des permutations** de l'ensemble $\{1, 2, \ldots, n\}$. A tout élément $\sigma \in \mathcal{T}_n$, on associe la **matrice de permutation** de $\mathbb{P}_\sigma \in \mathcal{M}_n$ est définie par

$$(\mathbb{P}_\sigma)_{i,j} = \delta_{i\sigma(j)}.$$

Exercice B.2.4: *résultats à savoir*

Montrer qu'une matrice de permutation est orthogonale.

Definition B.35

Soient $\mathbb{A} = (\mathrm{A}_{i,j})_{i,j=1}^{n} \in \mathcal{M}_n$ et $\mathcal{T}_n$ le **groupe des permutations** de l'ensemble $\{1, 2, \ldots, n\}$.. Le **déterminant** d'une matrice $\mathbb{A}$ est défini par

$$\det(\mathbb{A}) = \sum_{\sigma \in \mathcal{T}_n} \varepsilon_\sigma \prod_{j=1}^{n} \mathrm{A}_{\sigma(j)j}$$

où ε_σ désigne la signature de la permutation σ.

Proposition B.36: Méthode de Laplace ou des coffacteurs

Soit $\mathbb{A} = (\mathrm{A}_{i,j})_{i,j=1}^{n} \in \mathcal{M}_n$. On note $\mathbb{A}^{[i,j]} \in \mathcal{M}_{n-1}$ la matrice obtenue en supprimant la ligne i et la colonne j de $\mathbb{A}$. On a alors le **développement par rapport à la ligne** $i \in [\![1, n]\!]$

$$\det(\mathbb{A}) = \sum_{j=1}^{n} (-1)^{i+j} \mathrm{A}_{i,j} \det\left(\mathbb{A}^{[i,j]}\right), \tag{B.18}$$

et le **développement par rapport à la colonne** $j \in [\![1, n]\!]$

$$\det(\mathbb{A}) = \sum_{i=1}^{n} (-1)^{i+j} \mathrm{A}_{i,j} \det\left(\mathbb{A}^{[i,j]}\right). \tag{B.19}$$

Le terme $(-1)^{i+j} \det\left(\mathbb{A}^{[i,j]}\right)$ est appellé le **cofacteur** du terme $\mathrm{A}_{i,j}$.

Definition B.37

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$. On dit que $\lambda \in \mathbb{C}$ est **valeur propre** de $\mathbb{A}$ s'il existe $\boldsymbol{u} \in \mathbb{C}^n$ **non nul** tel que

$$\mathbb{A}\boldsymbol{u} = \lambda \boldsymbol{u}. \tag{B.20}$$

Le vecteur $\boldsymbol{u}$ est appelé **vecteur propre** associé à la valeur propre λ.
Le couple $(\lambda, \boldsymbol{u})$ est appelé **élément propre** de $\mathbb{A}$.

Definition B.38

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$. Soit $\lambda \in \mathbb{C}$ une valeur propre de $\mathbb{A}$. Le sous-espace

$$E_\lambda = \{\boldsymbol{u} \in \mathbb{C}^n : \mathbb{A}\boldsymbol{u} = \lambda \boldsymbol{u}\} = \ker(\mathbb{A} - \lambda \mathbb{I}) \tag{B.21}$$

est appelé **sous-espace propre** associé à la valeur propre λ. La dimension de E_λ est appelée **multiplicité géométrique** de la valeur propre λ.

Definition B.39

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$. Le polynôme de degré n défini par

$$\mathcal{P}_{\mathbb{A}}(\lambda) = \det(\mathbb{A} - \lambda \mathbb{I}) \tag{B.22}$$

est appellé **polynôme caractéristique** de la matrice $\mathbb{A}$.

Proposition B.40

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$.

- ◇ Les racines complexes du polynôme caractéristique $\mathcal{P}_{\mathbb{A}}$ sont les valeurs propres de la matrice $\mathbb{A}$.
- ◇ Si la racine λ de $\mathcal{P}_{\mathbb{A}}$ est de multiplicité k, on dit que la valeur propre λ est de **multiplicité algébrique** k.
- ◇ La matrice $\mathbb{A}$ possède n valeurs propres distinctes ou non.

Definition B.41

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$. On note $\lambda_i(\mathbb{A})$, $i \in [\![1, n]\!]$, les n valeurs propres de $\mathbb{A}$. Le **spectre** de la matrice $\mathbb{A}$ est le sous-ensemble

$$\mathrm{Sp}(\mathbb{A}) = \bigcup_{i=1}^{n} \{\lambda_i(\mathbb{A})\} \tag{B.23}$$

du plan complexe.

Proposition B.42

Soient $\mathbb{A} \in \mathcal{M}_n$ et $\mathbb{B} \in \mathcal{M}_n$. On a les relations suivantes

$$\begin{aligned}
\mathrm{tr}(\mathbb{A}) &= \sum_{i=1}^{n} \lambda_i(\mathbb{A}), && \text{(B.24)}\\
\det(\mathbb{A}) &= \prod_{i=1}^{n} \lambda_i(\mathbb{A}), && \text{(B.25)}\\
\mathrm{tr}(\mathbb{AB}) &= \mathrm{tr}(\mathbb{BA}), && \text{(B.26)}\\
\mathrm{tr}(\mathbb{A}+\mathbb{B}) &= \mathrm{tr}\,A + \mathrm{tr}\,B, && \text{(B.27)}\\
\det(\mathbb{AB}) &= \det(\mathbb{A})\det(\mathbb{B}) = \det(\mathbb{BA}), && \text{(B.28)}\\
\det(\mathbb{A}^*) &= \overline{\det(\mathbb{A})}. && \text{(B.29)}
\end{aligned}$$

Definition B.43

Le **rayon spectral** d'une matrice $\mathbb{A} \in \mathcal{M}_n$ est le nombre $\geqslant 0$ défini par

$$\rho(\mathbb{A}) = \max\{|\lambda_i(\mathbb{A})|\,;\, i \in [\![1, n]\!]\}$$

Matrices particulières

Definition B.44

Une matrice carrée $\mathbb{A} \in \mathcal{M}_n$ est :

- ◇ **diagonale** si $a_{ij} = 0$ pour $i \neq j$,
- ◇ **triangulaire supérieure** si $a_{ij} = 0$ pour $i > j$,
- ◇ **triangulaire inférieure** si $a_{ij} = 0$ pour $i < j$,
- ◇ **triangulaire** si elle est triangulaire supérieure ou triangulaire inférieure

◇ **à diagonale dominante** si

$$|a_{ii}| \geqslant \sum_{j \neq i} |a_{ij}|\,, \ \forall i \in [\![1, n]\!], \tag{B.30}$$

◇ **à diagonale strictement dominante** si

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|\,, \ \forall i \in [\![1, n]\!]. \tag{B.31}$$

Proposition B.45

Soient $\mathbb{A}$ et $\mathbb{B}$ deux matrices de $\mathcal{M}_n(\mathbb{K})$ triangulaires inférieures (resp. triangulaires supérieures). Alors la matrice $\mathbb{AB}$ est aussi triangulaire inférieure (resp. triangulaire supérieure).
De plus on a

$$(\mathbb{AB})_{i,i} = A_{i,i}B_{i,i}, \ \forall i \in [\![1, n]\!].$$

Preuve. (voir Exercice B.3.2, page 214) □

Proposition B.46

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$ une matrice triangulaire inférieure (resp. triangulaire supérieure).

1. $\mathbb{A}$ est inversible si et seulement si ses éléments diagonaux sont tous non nuls (i.e. $A_{i,i} \neq 0$, $\forall i \in [\![1, n]\!]$).
2. Si $\mathbb{A}$ est inversible alors son inverse est triangulaire inférieure (resp. triangulaire supérieure) et

$$(\mathbb{A}^{-1})_{i,i} = \frac{1}{(\mathbb{A})_{i,i}}$$

Preuve. (voir Exercice B.3.10, page 220) □

Definition B.47

On appelle **matrice bande** une matrice $\mathbb{A}$ telle que $a_{ij} \neq 0$ pour $|j - i| \leqslant c$. c est la **demi largeur de bande**.

Lorsque $c = 1$, la matrice est dite **tridiagonale** . Lorsque $c = 2$, la matrice est dite **pentadiagonale** .

Definition B.48

On appelle **sous-matrice** d'une matrice donnée, la matrice obtenue en supprimant certaines lignes et certaines colonnes. En particulier, si on supprime les $(n - k)$ dernières lignes et colonnes d'une matrice carrée $\mathbb{A}$ d'ordre n, on obtient la **sous matrice principale** d'ordre k.

Definition B.49

On appelle **matrice bloc** une matrice $\mathbb{A} \in \mathcal{M}_{N,M}$ écrite sous la forme

$$\mathbb{A} = \left(\begin{array}{c:c:c} \mathbb{A}_{1,1} & \cdots & \mathbb{A}_{1,q} \\ \hdashline \vdots & & \vdots \\ \hdashline \mathbb{A}_{p,1} & \cdots & \mathbb{A}_{p,q} \end{array}\right)$$

où $\forall i \in [\![1,p]\!]$, $\forall j \in [\![1,q]\!]$, $\mathbb{A}_{i,j}$ est une matrice de $\mathcal{M}_{n_i,m_j}$. On a $N = \sum_{i=1}^{p} n_i$ et $M = \sum_{j=1}^{q} m_j$.

On dit que $\mathbb{A}$ et une matrice **bloc-carrée** si $p = q$ et si tous les blocs diagonaux sont des matrices carrées.

Propriété B.50: Multiplication de matrices blocs

Soient $\mathbb{A} \in \mathcal{M}_{N,M}$ et $\mathbb{B} \in \mathcal{M}_{M,S}$. Le produit $\mathbb{P} = \mathbb{A}\mathbb{B} \in \mathcal{M}_{N,S}$ peut s'écrire sous forme bloc si les matrices $\mathbb{A}$ et $\mathbb{B}$ sont *compatibles par blocs* : il faut que le nombre de blocs colonne de $\mathbb{A}$ soit égale au nombre de blocs ligne de $\mathbb{B}$ avec correspondance des dimensions.

$$\mathbb{A} = \left(\begin{array}{c|c|c} \mathbb{A}_{1,1} & \cdots & \mathbb{A}_{1,q} \\ \hline \vdots & & \vdots \\ \hline \mathbb{A}_{p,1} & \cdots & \mathbb{A}_{p,q} \end{array}\right) \text{ et } \mathbb{B} = \left(\begin{array}{c|c|c} \mathbb{B}_{1,1} & \cdots & \mathbb{A}_{1,r} \\ \hline \vdots & & \vdots \\ \hline \mathbb{B}_{q,1} & \cdots & \mathbb{A}_{q,r} \end{array}\right)$$

avec $\mathbb{A}_{i,k} \in \mathcal{M}_{n_i,m_k}$ et $\mathbb{B}_{k,j} \in \mathcal{M}_{m_k,s_j}$ pour tout $i \in [\![1,p]\!]$, $k \in [\![1,q]\!]$ et $j \in [\![1,r]\!]$. La matrice produit $\mathbb{P}$ s'écrit alors sous la forme bloc

$$\mathbb{P} = \left(\begin{array}{c|c|c} \mathbb{P}_{1,1} & \cdots & \mathbb{P}_{1,r} \\ \hline \vdots & & \vdots \\ \hline \mathbb{P}_{p,1} & \cdots & \mathbb{P}_{p,r} \end{array}\right)$$

avec $\forall i \in [\![1,p]\!]$, $\forall j \in [\![1,r]\!]$ $\mathbb{P}_{i,j} \in \mathcal{M}_{n_i,s_j}$ et

$$\mathbb{P}_{i,j} = \sum_{k=1}^{q} \mathbb{A}_{i,k}\mathbb{B}_{k,j}.$$

Definition B.51

On dit qu'une matrice bloc-carrée $\mathbb{A}$ est **triangulaire inférieure** (resp. **supérieure**) **par blocs** si elle peut s'écrire sous la forme d'une matrice bloc avec les sous matrices $\mathbb{A}_{i,j} = 0$ pour $i < j$ (resp. $i > j$). . Elle s'écrit donc sous la forme

$$\mathbb{A} = \left(\begin{array}{c|c|c|c} \mathbb{A}_{1,1} & \mathbb{0} & \cdots & \mathbb{0} \\ \hline \vdots & \ddots & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \mathbb{0} \\ \hline \mathbb{A}_{n,1} & \cdots & \cdots & \mathbb{A}_{n,n} \end{array}\right) \text{ (resp. } \mathbb{A} = \left(\begin{array}{c|c|c|c} \mathbb{A}_{1,1} & \cdots & \cdots & \mathbb{A}_{n,1} \\ \hline \mathbb{0} & \ddots & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \vdots \\ \hline \mathbb{0} & \cdots & \mathbb{0} & \mathbb{A}_{n,n} \end{array}\right)\text{).}$$

Definition B.52

On dit qu'une matrice bloc-carrée $\mathbb{A}$ est **diagonale par blocs** ou **bloc-diagonale** si elle peut s'écrire sous la forme d'une matrice bloc avec les sous matrices $\mathbb{A}_{i,j} = 0$ pour $i \neq j$. Elle s'écrit donc sous la forme

$$\mathbb{A} = \left(\begin{array}{c|c|c|c} \mathbb{A}_{1,1} & \mathbb{0} & \cdots & \mathbb{0} \\ \hline \mathbb{0} & \ddots & \ddots & \vdots \\ \hline \vdots & \ddots & \ddots & \mathbb{0} \\ \hline \mathbb{0} & \cdots & \mathbb{0} & \mathbb{A}_{n,n} \end{array}\right)$$

Proposition B.53

Soit $\mathbb{A}$ une matrice bloc-carré décomposée en $n \times n$ blocs. Si $\mathbb{A}$ est **bloc-diagonale** ou **triangulaire par blocs** alors son déterminant est le produit des déterminant des blocs diagonaux :

$$\det \mathbb{A} = \prod_{i=1}^{n} \det \mathbb{A}_{i,i} \tag{B.32}$$

Proposition B.54

Soit $\mathbb{A}$ une matrice bloc-carré **inversible** décomposée en $n \times n$ blocs.

- Si $\mathbb{A}$ est **bloc-diagonale** alors son inverse (décomposée en $n \times n$ blocs) est aussi **bloc-diagonale**.
- Si $\mathbb{A}$ est **triangulaire inférieure par blocs** (resp. supérieure) alors son inverse (décomposée en $n \times n$ blocs) est aussi **triangulaire inférieure par blocs** (resp. supérieure).

Dans ces deux cas les blocs diagonaux de la matrice inverse sont les inverses des blocs diagonaux de $\mathbb{A}$. On a donc

$$\mathbb{A} = \begin{pmatrix} \mathbb{A}_{1,1} & \mathbb{0} & \cdots & \mathbb{0} \\ \mathbb{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbb{0} \\ \mathbb{0} & \cdots & \circ & \mathbb{A}_{n,n} \end{pmatrix} \text{ et } \mathbb{A}^{-1} = \begin{pmatrix} \mathbb{A}_{1,1}^{-1} & \mathbb{0} & \cdots & \mathbb{0} \\ \mathbb{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbb{0} \\ \mathbb{0} & \cdots & \mathbb{0} & \mathbb{A}_{n,n}^{-1} \end{pmatrix}$$

$$\mathbb{A} = \begin{pmatrix} \mathbb{A}_{1,1} & \mathbb{0} & \cdots & \mathbb{0} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbb{0} \\ \mathbb{A}_{n,1} & \cdots & \cdots & \mathbb{A}_{n,n} \end{pmatrix} \text{ et } \mathbb{A}^{-1} = \begin{pmatrix} \mathbb{A}_{1,1}^{-1} & \mathbb{0} & \cdots & \mathbb{0} \\ \bullet & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbb{0} \\ \bullet & \cdots & \bullet & \mathbb{A}_{n,n}^{-1} \end{pmatrix}$$

$$\mathbb{A} = \begin{pmatrix} \mathbb{A}_{1,1} & \cdots & \cdots & \mathbb{A}_{n,1} \\ \mathbb{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbb{0} & \cdots & \mathbb{0} & \mathbb{A}_{n,n} \end{pmatrix} \text{ et } \mathbb{A}^{-1} = \begin{pmatrix} \mathbb{A}_{1,1}^{-1} & \bullet & \cdots & \bullet \\ \mathbb{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \bullet \\ \mathbb{0} & \cdots & \mathbb{0} & \mathbb{A}_{n,n}^{-1} \end{pmatrix}$$

B.2.3 Normes vectorielles et normes matricielles

Definition B.55

Une **norme** sur un espace vectoriel V est une application $\|\bullet\| : V \to \mathbb{R}^+$ qui vérifie les propriétés suivantes

- $\|\boldsymbol{v}\| = 0 \iff \boldsymbol{v} = 0$,
- $\|\alpha \boldsymbol{v}\| = |\alpha| \, \|\boldsymbol{v}\|, \ \forall \alpha \in \mathbb{K}, \ \forall \boldsymbol{v} \in V$,
- $\|\boldsymbol{u} + \boldsymbol{v}\| \leqslant \|\boldsymbol{u}\| + \|\boldsymbol{v}\|, \ \forall (\boldsymbol{u}, \boldsymbol{v}) \in V^2$ (inégalité triangulaire).

Une norme sur V est également appelée **norme vectorielle** . On appelle **espace vectoriel normé** un espace vectoriel muni d'une norme.

Les trois normes suivantes sont les plus couramment utilisées :

$$\|\boldsymbol{v}\|_1 = \sum_{i=1}^{n} |v_i|$$
$$\|\boldsymbol{v}\|_2 = \left(\sum_{i=1}^{n} |v_i|^2 \right)^{1/2}$$
$$\|\boldsymbol{v}\|_\infty = \max_i |v_i| .$$

Théorème B.56

Soit V un espace de dimension finie. Pour tout nombre réel $p \geqslant 1$, l'application $\|\bullet\|_p$ définie par

$$\|\boldsymbol{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p}$$

est une norme.

Proposition B.57

Pour $p > 1$ et $\frac{1}{p} + \frac{1}{q} = 1$, on a $\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{K}^n$

$$\sum_{i=1}^{n} |u_i v_i| \leqslant \left(\sum_{i=1}^{n} |u_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |v_i|^q \right)^{1/q} = \|\boldsymbol{u}\|_p \|\boldsymbol{v}\|_q . \tag{B.33}$$

Cette inégalité s'appelle l'**inégalité de Hölder**.

Definition B.58

Deux **normes** $\|\bullet\|$ et $\|\bullet\|'$, définies sur un même espace vectoriel V, sont **équivalentes** s'il exite deux constantes C et C' telles que

$$\|\boldsymbol{v}\|' \leqslant C \|\boldsymbol{v}\| \quad \text{et} \quad \|\boldsymbol{v}\| \leqslant C' \|\boldsymbol{v}\|' \quad \text{pour tout } \boldsymbol{v} \in V. \tag{B.34}$$

Proposition B.59

Sur un espace vectoriel de dimension finie toutes les normes sont équivalentes.

Definition B.60

Une **norme matricielle** sur $\mathcal{M}_n(\mathbb{K})$ est une application $\|\bullet\| : \mathcal{M}_n(\mathbb{K}) \to \mathbb{R}^+$ vérifiant

1. $\|\mathbb{A}\| = 0 \iff \mathbb{A} = 0$,
2. $\|\alpha\mathbb{A}\| = |\alpha| \|\mathbb{A}\|$, $\forall \alpha \in \mathbb{K}$, $\forall \mathbb{A} \in \mathcal{M}_n(\mathbb{K})$,
3. $\|\mathbb{A} + \mathbb{B}\| \leqslant \|\mathbb{A}\| + \|\mathbb{B}\|$, $\forall (\mathbb{A}, \mathbb{B}) \in \mathcal{M}_n(\mathbb{K})^2$ (inégalité triangulaire)
4. $\|\mathbb{A}\mathbb{B}\| \leqslant \|\mathbb{A}\| \|\mathbb{B}\|$, $\forall (\mathbb{A}, \mathbb{B}) \in \mathcal{M}_n(\mathbb{K})^2$

Proposition B.61

Etant donné une norme vectorielle $\|\bullet\|$ sur $\mathbb{K}^n$, l'application $\|\bullet\|_s : \mathcal{M}_n(\mathbb{K}) \to \mathbb{R}^+$ définie par

$$\|\mathbb{A}\|_s = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n \\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|}{\|\boldsymbol{v}\|} = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n \\ \|\boldsymbol{v}\|\leqslant 1}} \|\mathbb{A}\boldsymbol{v}\| = \sup_{\substack{\boldsymbol{v}\in\mathbb{K}^n \\ \|\boldsymbol{v}\|=1}} \|\mathbb{A}\boldsymbol{v}\|, \tag{B.35}$$

est une norme matricielle, appelée **norme matricielle subordonnée** (à la norme vectorielle donnée).
De plus

$$\|\mathbb{A}\boldsymbol{v}\| \leqslant \|\mathbb{A}\|_s \|\boldsymbol{v}\| \quad \forall \boldsymbol{v} \in \mathbb{K}^n \tag{B.36}$$

et la norme $\|\mathbb{A}\|$ peut se définir aussi par

$$\|\mathbb{A}\|_s = \inf\{\alpha \in \mathbb{R} : \|\mathbb{A}\boldsymbol{v}\| \leqslant \alpha \|\boldsymbol{v}\|, \ \forall \boldsymbol{v} \in \mathbb{K}^n\}. \tag{B.37}$$

Il existe au moins un vecteur $\boldsymbol{u} \in \mathbb{K}^n$ tel que

$$\boldsymbol{u} \neq 0 \ \text{ et } \ \|\mathbb{A}\boldsymbol{u}\| = \|\mathbb{A}\|_s \|\boldsymbol{u}\|. \tag{B.38}$$

Enfin une norme subordonnée vérifie toujours

$$\|\mathbb{I}\|_s = 1 \tag{B.39}$$

Théorème B.62

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$. On a

$$\|\mathbb{A}\|_1 \overset{\text{déf.}}{=} \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|_1}{\|\boldsymbol{v}\|_1} = \max_{j\in[\![1,n]\!]} \sum_{i=1}^{n} |a_{ij}| \tag{B.40}$$

$$\|\mathbb{A}\|_2 \overset{\text{déf.}}{=} \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2} = \sqrt{\rho\left(\mathbb{A}^*\mathbb{A}\right)} = \sqrt{\rho\left(\mathbb{A}\mathbb{A}^*\right)} = \|\mathbb{A}^*\|_2 \tag{B.41}$$

$$\|\mathbb{A}\|_\infty \overset{\text{déf.}}{=} \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq 0}} \frac{\|\mathbb{A}\boldsymbol{v}\|_\infty}{\|\boldsymbol{v}\|_\infty} = \max_{i\in[\![1,n]\!]} \sum_{j=1}^{n} |a_{ij}| \tag{B.42}$$

La norme $\|\bullet\|_2$ est invariante par transformation unitaire :

$$\mathbb{U}\mathbb{U}^* = \mathbb{I} \Longrightarrow \|\mathbb{A}\|_2 = \|\mathbb{A}\mathbb{U}\|_2 = \|\mathbb{U}\mathbb{A}\|_2 = \|\mathbb{U}^*\mathbb{A}\mathbb{U}\|_2. \tag{B.43}$$

Par ailleurs, si la matrice $\mathbb{A}$ est normale :

$$\mathbb{A}\mathbb{A}^* = \mathbb{A}^*\mathbb{A} \Longrightarrow \|\mathbb{A}\|_2 = \rho(\mathbb{A}). \tag{B.44}$$

Proposition B.63

1. Si une matrice $\mathbb{A}$ est hermitienne, ou symétrique (donc normale), on a $\|\mathbb{A}\|_2 = \rho(\mathbb{A})$.
2. Si une matrice $\mathbb{A}$ est unitaire, ou orthogonale (donc normale), on a $\|\mathbb{A}\|_2 = 1$.

Théorème B.64

1. Soit $\mathbb{A}$ une matrice carrée quelconque et $\|\bullet\|$ une norme matricielle subordonnée ou non, quel-

conque. Alors

$$\rho(\mathbb{A}) \leqslant \|\mathbb{A}\|. \tag{B.45}$$

2. Etant donné une matrice $\mathbb{A}$ et un nombre $\varepsilon > 0$, il existe au moins une norme matricielle subordonnée telle que

$$\|\mathbb{A}\| \leqslant \rho(\mathbb{A}) + \varepsilon. \tag{B.46}$$

Théorème B.65

L'application $\|\bullet\|_E : \mathcal{M}_n \to \mathbb{R}^+$ définie par

$$\|\mathbb{A}\|_E = \left(\sum_{(i,j)=\in[\![1,n]\!]^2} |a_{ij}|^2 \right)^{1/2} = \sqrt{\mathrm{tr}\,(\mathbb{A}^*\mathbb{A})}, \tag{B.47}$$

pour toute matrice $\mathbb{A} = (a_{ij})$ d'ordre n, est une norme matricielle non subordonnée (pour $n \geqslant 2$), invariante par transformation unitaire et qui vérifie

$$\|\mathbb{A}\|_2 \leqslant \|\mathbb{A}\|_E \leqslant \sqrt{n}\,\|\mathbb{A}\|_2\,, \ \forall \mathbb{A} \in \mathcal{M}_n. \tag{B.48}$$

De plus $\|\mathbb{I}\|_E = \sqrt{n}$.

Théorème B.66

1. Soit $\|\bullet\|$ une norme matricielle subordonnée, et $\mathbb{B}$ une matrice vérifiant

$$\|\mathbb{B}\| < 1.$$

Alors la matrice $(\mathbb{I} + \mathbb{B})$ est inversible, et

$$\left\|(\mathbb{I} + \mathbb{B})^{-1}\right\| \leqslant \frac{1}{1 - \|\mathbb{B}\|}.$$

2. Si une matrice de la forme $(\mathbb{I} + \mathbb{B})$ est singulière, alors nécessairement

$$\|\mathbb{B}\| \geqslant 1$$

pour toute norme matricielle, subordonnée ou non.

B.2.4 Réduction des matrices

Definition B.67

Soit $A : V \to V$ une application linéaire, représenté par une matrice carrée $\mathbb{A} \in \mathcal{M}_n$ relativement à une base $\{\boldsymbol{e}_i\}_{i\in[\![1,n]\!]}$. Relativement à une autre base $\{\boldsymbol{f}_i\}_{i\in[\![1,n]\!]}$, la même application est représentée par la matrice

$$\mathbb{B} = \mathbb{P}^{-1}\mathbb{A}\mathbb{P} \tag{B.49}$$

où $\mathbb{P}$ est la matrice inversible dont le j-ème vecteur colonne est formé des composantes du vecteur $\boldsymbol{f}_j$ dans la base $\{\boldsymbol{e}_i\}_{i\in[\![1,n]\!]}$:

$$\mathbb{P} = \begin{pmatrix} \langle \boldsymbol{e}_1, \boldsymbol{f}_1\rangle & \langle \boldsymbol{e}_1, \boldsymbol{f}_2\rangle & \cdots & \langle \boldsymbol{e}_1, \boldsymbol{f}_n\rangle \\ \langle \boldsymbol{e}_2, \boldsymbol{f}_1\rangle & \langle \boldsymbol{e}_1, \boldsymbol{f}_2\rangle & \ddots & \vdots \\ \vdots & \ddots & \ddots & \langle \boldsymbol{e}_{n-1}, \boldsymbol{f}_n\rangle \\ \langle \boldsymbol{e}_n, \boldsymbol{f}_1\rangle & \cdots & \langle \boldsymbol{e}_n, \boldsymbol{f}_{n-1}\rangle & \langle \boldsymbol{e}_n, \boldsymbol{f}_n\rangle \end{pmatrix} \tag{B.50}$$

La matrice $\mathbb{P}$ est appelée **matrice de passage de la base** $\{\boldsymbol{e}_i\}_{i\in[\![1,n]\!]}$ **dans le base** $\{\boldsymbol{f}_i\}_{i\in[\![1,n]\!]}$.

Definition B.68

On dit que la matrice carrée $\mathbb{A}$ est diagonalisable s'il existe une matrice inversible $\mathbb{P}$ telle que la matrice $\mathbb{P}^{-1}\mathbb{A}\mathbb{P}$ soit diagonale.

Remarque B.69 On notera que, dans le cas où $\mathbb{A}\in\mathcal{M}_n$ est diagonalisable, les éléments diagonaux de la matrice $\mathbb{P}^{-1}\mathbb{A}\mathbb{P}$ sont les valeurs propres $\lambda_1,\lambda_2,\ldots,\lambda_n$ de la matrice $\mathbb{A}$, et que le j-ème vecteur colonne $\boldsymbol{p}_j$ de la matrice $\mathbb{P}$ est formé des composantes, dans la même base que $\mathbb{A}$, d'un vecteur propre associé à la valeur propre λ_j. On a

$$\mathbb{P}^{-1}\mathbb{A}\mathbb{P}=\operatorname{diag}(\lambda_1,\ldots,\lambda_n)\Longleftrightarrow\mathbb{A}\boldsymbol{p}_j=\lambda_j\boldsymbol{p}_j,\ \forall j\in[\![1,n]\!]. \tag{B.51}$$

C'est à dire qu'une matrice est diagonalisable si, et seulement si, il existe une base de vecteurs propres.

Théorème B.70

1. Etant donnée une matrice **carrée** $\mathbb{A}$, il existe une matrice **unitaire** $\mathbb{U}$ telle que la matrice $\mathbb{U}^{-1}\mathbb{A}\mathbb{U}$ soit **triangulaire**.
2. Etant donnée une matrice **normale** $\mathbb{A}$, il existe une matrice **unitaire** $\mathbb{U}$ telle que la matrice $\mathbb{U}^{-1}\mathbb{A}\mathbb{U}$ soit **diagonale**.
3. Etant donnée une matrice **symétrique** $\mathbb{A}$, il existe une matrice **orthogonale** $\mathbb{O}$ telle que la matrice $\mathbb{O}^{-1}\mathbb{A}\mathbb{O}$ soit **diagonale**.

B.2.5 Suites de vecteurs et de matrices

Definition B.71

Soit V un espace vectoriel muni d'une norme $\|\bullet\|$, on dit qu'une suite $(\boldsymbol{v}_k)$ d'éléments de V **converge vers un élément** $\boldsymbol{v}\in V$, si

$$\lim_{k\to\infty}\|\boldsymbol{v}_k-\boldsymbol{v}\|=0$$

et on écrit

$$\boldsymbol{v}=\lim_{k\to\infty}\boldsymbol{v}_k.$$

Théorème B.72

Soit $\mathbb{B}$ une matrice carrée. Les conditions suivantes sont équivalentes :

1. $\lim_{k\to\infty}\mathbb{B}^k=0$,
2. $\lim_{k\to\infty}\mathbb{B}^k\boldsymbol{v}=0$ pour tout vecteur $\boldsymbol{v}$,
3. $\rho(\mathbb{B})<1$,
4. $\|\mathbb{B}\|<1$ pour au moins une norme matricielle subordonnée $\|\bullet\|$.

Théorème B.73

Soit $\mathbb{B}$ une matrice carrée, et $\|\bullet\|$ une norme matricielle quelconque. Alors

$$\lim_{k\to\infty} \left\|\mathbb{B}^k\right\|^{1/k} = \rho(\mathbb{B}).$$

B.3 Receuil d'exercices

B.3.1 Algèbre linéaire

Sur les matrices

Exercice B.3.1

Soit $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{R})$ et $\mathbb{B} \in \mathcal{M}_{n,m}(\mathbb{R})$ telles que

$$\langle \mathbb{A}\boldsymbol{u}, \boldsymbol{v}\rangle_m = \langle \boldsymbol{u}, \mathbb{B}\boldsymbol{v}\rangle_n\,,\ \forall \boldsymbol{u} \in \mathbb{R}^n,\ \forall \boldsymbol{v} \in \mathbb{R}^m.$$

Exprimer les éléments de la matrice $\mathbb{B}$ en fonction de ceux de la matrice $\mathbb{A}$.

Exercice B.3.2

Soient $\mathbb{A}$ et $\mathbb{B}$ deux matrices triangulaires supérieures de $\mathcal{M}_n$. Soient $\mathbb{E}$ et $\mathbb{F}$ deux matrices triangulaires inférieures de $\mathcal{M}_n$.

Q. 1 1. *Que peut-on dire des matrices* $\mathbb{A}^*$ *et* $(\mathbb{A}^*)^*$ *?*

2. *Montrer que* $\mathbb{C} = \mathbb{A}\mathbb{B}$ *est triangulaire supérieure et que* $C_{i,i} = A_{i,i}B_{i,i}$, $\forall i \in [\![1, n]\!]$.

3. *Montrer que* $\mathbb{G} = \mathbb{E}\mathbb{F}$ *est triangulaire inférieure et que* $G_{i,i} = E_{i,i}F_{i,i}$, $\forall i \in [\![1, n]\!]$.

4. *Que peut-on dire des matrices* $\mathbb{A}\mathbb{E}$ *et* $\mathbb{E}\mathbb{A}$*?*

Q. 2 1. *Calculer* $\det(\mathbb{A})$.

2. *Déterminer les valeurs propres de* $\mathbb{A}$.

3. *Que peut-on dire si les éléments diagonaux de* $\mathbb{A}$ *sont tous distincts?*

Q. 3 *Soit* $\mathbb{D}$ *la matrice définie par*

$$\mathbb{D}\begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}.$$

1. *La matrice* $\mathbb{D}$ *est-elle inversible? Si oui calculer son inverse.*

2. *Pour chacune des valeurs propres, déterminer l'espace propre associé.*

3. *La matrice* $\mathbb{D}$ *est-elle diagonalisable? Justifier.*

Exercice B.3.3

Q. 1 *Soit* $\mathbb{T} \in \mathcal{M}_{n,n}(\mathbb{C})$ *une matrice triangulaire supérieure. Montrer que si* $\mathbb{T}$ *est une matrice normale alors elle est diagonale.*

Q. 2 *Montrer que* $\mathbb{A} \in \mathcal{M}_{n,n}(\mathbb{C})$ *est une matrice normale si et seulement si il existe* $\mathbb{U} \in \mathcal{M}_{n,n}(\mathbb{C})$ *unitaire et* $\mathbb{D} \in \mathcal{M}_{n,n}(\mathbb{C})$ *diagonale telle que* $\mathbb{A} = \mathbb{U}\mathbb{D}\mathbb{U}^*$.

Q. 3 *En déduire qu'une matrice normale est diagonalisable et que ses vecteurs propres sont orthog-*

onaux.

Exercice B.3.4

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice hermitienne

Q. 1 *Montrer que*

$$\langle \mathbb{A}\boldsymbol{u}, \boldsymbol{u} \rangle \in \mathbb{R}, \ \forall \boldsymbol{u} \in \mathbb{C}^n. \tag{B.52}$$

On suppose de plus que la matrice $\mathbb{A}$ est définie positve.

Q. 2 *1. Montrer que les éléments diagonaux de $\mathbb{A}$ sont strictement positifs.*

2. Montrer que les sous matrices principales de $\mathbb{A}$ sont elles aussi hermitiennes et définies positves.

Exercice B.3.5: Procédé de Gram-Schmidt

Soit $\{\boldsymbol{v}_i\}_{i\in[\![1,n]\!]}$ une base de $\mathbb{K}^n$. On construit successivement les vecteurs $\boldsymbol{u}_i$

$$\boldsymbol{u}_i = \boldsymbol{v}_i - \sum_{k=1}^{i-1} \frac{\langle \boldsymbol{u}_k, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{u}_k, \boldsymbol{u}_k \rangle} \boldsymbol{u}_k, \ \forall i \in [\![1, n]\!].$$

Montrer qu'ils forment une **base orthogonale** de $\mathbb{K}^n$ et que $\text{Vect}\,(\boldsymbol{u}_1, \dots, \boldsymbol{u}_i) = \text{Vect}\,(\boldsymbol{v}_1, \dots, \boldsymbol{v}_i)$, $\forall i \in [\![1, n]\!]$.

Correction Exercice B.3.5 Montrons par récurrence sur i que

$$(\mathcal{H})_i : \text{Vect}\,(\boldsymbol{u}_1, \dots, \boldsymbol{u}_i) \text{ est une famille orthogonale et } \text{Vect}\,(\boldsymbol{u}_1, \dots, \boldsymbol{u}_i) = \text{Vect}\,(\boldsymbol{v}_1, \dots, \boldsymbol{v}_i)$$

Initialisation : Pour $i = 1$, on a $\boldsymbol{u}_1 = \boldsymbol{v}_1$ et $(\mathcal{H})_1$ est vérifiée.

Hérédité : Soit $i < n$. Supposons $(\mathcal{H})_i$ vérifiée. Montrons alors que $(\mathcal{H})_{i+1}$ est vraie.
On a

$$\boldsymbol{u}_{i+1} = \boldsymbol{v}_{i+1} - \sum_{k=1}^{i} \frac{\langle \boldsymbol{u}_k, \boldsymbol{v}_{i+1} \rangle}{\langle \boldsymbol{u}_k, \boldsymbol{u}_k \rangle} \boldsymbol{u}_k. \tag{B.53}$$

- En effectuant le produit scalaire de (B.53) par $\boldsymbol{u}_j$ avec $j \in [\![1, i]\!]$ on obtient

$$\langle \boldsymbol{u}_j, \boldsymbol{u}_{i+1} \rangle = \langle \boldsymbol{u}_j, \boldsymbol{v}_{i+1} \rangle - \sum_{k=1}^{i} \frac{\langle \boldsymbol{u}_k, \boldsymbol{v}_{i+1} \rangle}{\langle \boldsymbol{u}_k, \boldsymbol{u}_k \rangle} \langle \boldsymbol{u}_j, \boldsymbol{u}_k \rangle.$$

  Par hypothèse de récurrence, la famille $\text{Vect}\,(\boldsymbol{u}_1, \dots, \boldsymbol{u}_i)$ est orthogonale, c'est à dire $\forall (r, s) \in [\![1, i]\!]^2$, $\langle \boldsymbol{u}_r, \boldsymbol{u}_s \rangle = 0$ si $r \neq s$ et $\boldsymbol{u}_r \neq 0$. On obtient donc

$$\langle \boldsymbol{u}_j, \boldsymbol{u}_{i+1} \rangle = \langle \boldsymbol{u}_j, \boldsymbol{v}_{i+1} \rangle - \frac{\langle \boldsymbol{u}_j, \boldsymbol{v}_{i+1} \rangle}{\langle \boldsymbol{u}_j, \boldsymbol{u}_j \rangle} \langle \boldsymbol{u}_j, \boldsymbol{u}_j \rangle = 0, \ \forall j \in [\![1, i]\!].$$

- On montre maintenant par l'absurde que $\boldsymbol{u}_{i+1} \neq 0$.
  Supposons $\boldsymbol{u}_{i+1} = 0$. Alors de (B.53), on obtient

$$\boldsymbol{v}_{i+1} = \sum_{k=1}^{i} \frac{\langle \boldsymbol{u}_k, \boldsymbol{v}_{i+1} \rangle}{\langle \boldsymbol{u}_k, \boldsymbol{u}_k \rangle} \boldsymbol{u}_k$$

  et donc $\boldsymbol{v}_{i+1} \in \text{Vect}\,(\boldsymbol{u}_1, \dots, \boldsymbol{u}_i) \overset{(\mathcal{H})_i}{=} \text{Vect}\,(\boldsymbol{v}_1, \dots, \boldsymbol{v}_i)$. Ceci entre en contradiction avec $\text{Vect}\,(\boldsymbol{v}_1, \dots, \boldsymbol{v}_n)$ base de $\mathbb{K}^n$.
- On déduit de (B.53) que $\boldsymbol{u}_{i+1} \in \text{Vect}\,(\boldsymbol{u}_1, \dots, \boldsymbol{u}_i, \boldsymbol{v}_{i+1})$. Par hypothèse de récurrence, $\text{Vect}\,(\boldsymbol{u}_1, \dots, \boldsymbol{u}_i) = \text{Vect}\,(\boldsymbol{v}_1, \dots, \boldsymbol{v}_i)$, ce qui donne $\boldsymbol{u}_{i+1} \in \text{Vect}\,(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i+1})$ et donc

$$\text{Vect}\,(\boldsymbol{u}_1, \dots, \boldsymbol{u}_{i+1}) = \text{Vect}\,(\boldsymbol{v}_1, \dots, \boldsymbol{v}_{i+1}).$$

◇

Exercice B.3.6: factorisation $\mathbb{QR}$

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$ une matrice inversible. Pour tout $i \in [\![1, n]\!]$, on note $\boldsymbol{a}_i = \mathbb{A}_{:,i}$ ses n vecteurs colonnes. En utilisant le procédé de Gram-schmidt sur la base $\{\boldsymbol{a}_1, \dots, \boldsymbol{a}_n\}$ montrer qu'il existe une matrice $\mathbb{Q}$ unitaire et une matrice triangulaire supérieure $\mathbb{R}$ à coefficients diagonaux strictement positifs tel que $\mathbb{A} = \mathbb{QR}$.

Correction Exercice B.3.6 On utilise le procédé d'orthonormalisation de Gram-Schmidt (voir Proposition B.19, page 201) pour obtenir la base orthogonale $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_n\}$ en calculant successivement

$$\boldsymbol{u}_i = \boldsymbol{a}_i - \sum_{k=1}^{i-1} \frac{\langle \boldsymbol{u}_k, \boldsymbol{a}_i \rangle}{\langle \boldsymbol{u}_k, \boldsymbol{u}_k \rangle} \boldsymbol{u}_k, \ \forall i \in [\![1, n]\!]. \tag{B.54}$$

De plus on a $\text{Vect}\,(\boldsymbol{u}_1, \dots, \boldsymbol{u}_i) = \text{Vect}\,(\boldsymbol{a}_1, \dots, \boldsymbol{a}_i)\,, \forall i \in [\![1, n]\!]$ On normalise la base orthogonale $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_n\}$ pour obtenir la base orthonormée $\mathcal{Q} = \{\boldsymbol{q}_1, \dots, \boldsymbol{q}_n\}$:

$$\boldsymbol{q}_i = \frac{\boldsymbol{u}_i}{\|\boldsymbol{u}_i\|_2}, \ \forall i \in [\![1, n]\!]$$

et l'on a aussi $\text{Vect}\,(\boldsymbol{q}_1, \dots, \boldsymbol{q}_i) = \text{Vect}\,(\boldsymbol{a}_1, \dots, \boldsymbol{a}_i)\,, \forall i \in [\![1, n]\!]$.
On note $\mathbb{Q} \in \mathcal{M}_n(\mathbb{K})$ la matrice définie par

$$\mathbb{Q} = \left(\begin{array}{c:c:c} & & \\ \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_n \\ & & \end{array} \right)$$

Cette matrice est clairement unitaire puisque la base $\{\boldsymbol{q}_1, \dots, \boldsymbol{q}_n\}$ est orthonormée.
Montrons que $\mathbb{Q}^*\mathbb{A}$ est triangulaire supérieure. On a

$$\mathbb{Q}^*\mathbb{A} = \left(\begin{array}{c} \boldsymbol{q}_1^* \\ \hdashline \vdots \\ \hdashline \boldsymbol{q}_n^* \end{array} \right) \left(\begin{array}{c:c:c} & & \\ \boldsymbol{a}_1 & \cdots & \boldsymbol{a}_n \\ & & \end{array} \right) = \left(\begin{array}{c:c:c} \langle \boldsymbol{q}_1, \boldsymbol{a}_1 \rangle & \cdots & \langle \boldsymbol{q}_1, \boldsymbol{a}_n \rangle \\ \hdashline \vdots & \ddots & \vdots \\ \hdashline \langle \boldsymbol{q}_n, \boldsymbol{a}_1 \rangle & \cdots & \langle \boldsymbol{q}_n, \boldsymbol{a}_n \rangle \end{array} \right)$$

c'est à dire $(\mathbb{Q}^*\mathbb{A})_{i,j} = \langle \boldsymbol{q}_i, \boldsymbol{a}_j \rangle$, $\forall (i, j) \in [\![1, n]\!]$. Par définition cette matrice est triangulaire supérieure si $(\mathbb{Q}^*\mathbb{A})_{i,j} = 0$ pour tout $i > j$. Soit $i \in [\![1, n-1]\!]$. La base $\mathcal{Q}$ étant orthonormée, on a $\boldsymbol{q}_i \perp \text{Vect}\left(\boldsymbol{q}_1, \dots, \boldsymbol{q}_{i-1}\right)$. Comme $\text{Vect}\left(\boldsymbol{q}_1, \dots, \boldsymbol{q}_{i-1}\right) = \text{Vect}\,(\boldsymbol{a}_1, \dots, \boldsymbol{a}_{i-1})$, on en déduit que

$$\langle \boldsymbol{q}_i, \boldsymbol{a}_j \rangle = 0, \ \forall j \in [\![1, i-1]\!].$$

La matrice $\mathbb{Q}^*\mathbb{A}$ est donc triangulaire supérieure.
De plus, on a

$$(\mathbb{Q}^*\mathbb{A})_{i,i} = \langle \boldsymbol{q}_i, \boldsymbol{a}_i \rangle = \frac{\langle \boldsymbol{u}_i, \boldsymbol{a}_i \rangle}{\|\boldsymbol{u}_i\|_2}$$

En prenant le produit scalaire de (B.54) avec $\boldsymbol{u}_i$ on obtient

$$\begin{aligned} \langle \boldsymbol{u}_i, \boldsymbol{u}_i \rangle &= \langle \boldsymbol{u}_i, \boldsymbol{a}_i \rangle - \sum_{k=1}^{i-1} \frac{\langle \boldsymbol{u}_k, \boldsymbol{a}_i \rangle}{\langle \boldsymbol{u}_k, \boldsymbol{u}_k \rangle} \langle \boldsymbol{u}_i, \boldsymbol{u}_k \rangle \\ &= \langle \boldsymbol{u}_i, \boldsymbol{a}_i \rangle \quad \text{car } \langle \boldsymbol{u}_i, \boldsymbol{u}_k \rangle = 0, \ \forall k \neq i \text{ (base orthogonale)} \end{aligned}$$

Comme $\boldsymbol{u}_i \neq 0$, on obtient $(\mathbb{Q}^*\mathbb{A})_{i,i} > 0$.
On note $\mathbb{R} = \mathbb{Q}^*\mathbb{A}$ cette matrice triangulaire supérieure avec $R_{i,i} > 0$, $\forall i \in [\![1, n]\!]$. La matrice $\mathbb{Q}$ étant unitaire (i.e. $\mathbb{QQ}^* = \mathbb{I}$) alors $\mathbb{A} = \mathbb{QR}$. ◇

Exercice: 3.1.2

Soit $\mathbb{A} \in \mathcal{M}_{n,n}(\mathbb{C})$ une matrice et $(\lambda, \boldsymbol{u})$ un élément propre de $\mathbb{A}$ avec $\|\boldsymbol{u}\|_2 = 1$.

Q. 1 *En s'aidant de la base canonique* $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$, *construire une base orthonormée* $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ *telle que* $\boldsymbol{x}_1 = \boldsymbol{u}$.

Notons $\mathbb{P}$ la matrice de changement de base canonique $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$ dans la base $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$:

$$\mathbb{P} = \left(\begin{array}{c:c:c:c} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{array} \right)$$

Soit $\mathbb{B}$ la matrice définie par $\mathbb{B} = \mathbb{P}^*\mathbb{A}\mathbb{P}$.

Q. 2 1. *Exprimer les coefficients de la matrice* $\mathbb{B}$ *en fonction de la matrice* $\mathbb{A}$ *et des vecteurs* $\boldsymbol{x}_i$, $i \in [\![1, n]\!]$.

$$\mathbb{B} = \mathbb{P}^*\mathbb{A}\mathbb{P}.$$

2. *En déduire que la première colonne de* $\mathbb{B}$ *est* $(\lambda, 0, \dots, 0)^t$.

Q. 3 *Montrer par récurrence sur l'ordre de la matrice que la matrice* $\mathbb{A}$ *s'écrit*

$$\mathbb{A} = \mathbb{U}\mathbb{T}\mathbb{U}^*$$

où $\mathbb{U}$ *est une matrice unitaire et* $\mathbb{T}$ *une matrice triangulaire supérieure.*

Q. 4 *En supposant* $\mathbb{A}$ *inversible et la décomposition* $\mathbb{A} = \mathbb{U}\mathbb{T}\mathbb{U}^*$ *connue, expliquer comment résoudre "simplement" le système linéaire* $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$.

Correction Exercice 3.1.2

Q. 1 La première chose à faire est de construire une base contenant $\boldsymbol{u}$ à partir de la base canonique $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$. Comme le vecteur propre $\boldsymbol{u}$ est non nul, il existe $j \in [\![1, n]\!]$ tel que $\langle \boldsymbol{u}, \boldsymbol{e}_j \rangle \neq 0$. La famille $\{\boldsymbol{u}, \boldsymbol{e}_1, \dots, \boldsymbol{e}_{j-1}, \boldsymbol{e}_{j+1}, \dots, \boldsymbol{e}_n\}$ forme alors une base de $\mathbb{C}^n$ car $\boldsymbol{u}$ n'est pas combinaison linéaire des $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_{j-1}, \boldsymbol{e}_{j+1}, \dots, \boldsymbol{e}_n\}$.
On note $\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\}$ la base dont le premier élément est $\boldsymbol{z}_1 = \boldsymbol{u}$:

$$\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\} = \{\boldsymbol{u}, \boldsymbol{e}_1, \dots, \boldsymbol{e}_{j-1}, \boldsymbol{e}_{j+1}, \dots, \boldsymbol{e}_n\}.$$

On peut ensuite utiliser le **procédé de Gram-Schmidt**, rappelé en Proposition B.19, pour construire une base orthonormée à partir de cette base.
On calcule successivement les vecteurs $\boldsymbol{x}_i$ à partir de la base $\{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\}$ en construisant un vecteur $\boldsymbol{w}_i$ orthogonal aux vecteurs $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{i-1}$.

$$\boldsymbol{w}_i = \boldsymbol{z}_i - \sum_{k=1}^{i-1} \langle \boldsymbol{x}_k, \boldsymbol{z}_i \rangle \boldsymbol{x}_k$$

puis on obtient le vecteur $\boldsymbol{x}_i$ en normalisant

$$\boldsymbol{x}_i = \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|}$$

Q. 2 1. En conservant l'écriture colonne de la matrice $\mathbb{P}$ on obtient

$$\mathbb{B} = \left(\begin{array}{c} \boldsymbol{x}_1^* \\ \hdashline \boldsymbol{x}_2^* \\ \hdashline \vdots \\ \hdashline \boldsymbol{x}_n^* \end{array} \right) \mathbb{A} \left(\begin{array}{c:c:c:c} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n \end{array} \right) = \left(\begin{array}{c} \boldsymbol{x}_1^* \\ \hdashline \boldsymbol{x}_2^* \\ \hdashline \vdots \\ \hdashline \boldsymbol{x}_n^* \end{array} \right) \left(\begin{array}{c:c:c:c} \mathbb{A}\boldsymbol{x}_1 & \mathbb{A}\boldsymbol{x}_2 & \dots & \mathbb{A}\boldsymbol{x}_n \end{array} \right)$$

Ce qui donne

$$\mathbb{B} = \begin{pmatrix} \boldsymbol{x}_1^* \mathbb{A} \boldsymbol{x_1} & \boldsymbol{x_1}^* \mathbb{A} \boldsymbol{x}_2 & \dots & \boldsymbol{x_1}^* \mathbb{A} \boldsymbol{x}_n \\ \boldsymbol{x}_2^* \mathbb{A} \boldsymbol{x_1} & \boldsymbol{x}_2^* \mathbb{A} \boldsymbol{x}_2 & \dots & \boldsymbol{x}_2^* \mathbb{A} \boldsymbol{x}_n \\ \vdots & \vdots & & \vdots \\ \boldsymbol{x}_n^* \mathbb{A} \boldsymbol{x_1} & \boldsymbol{x}_n^* \mathbb{A} \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n^* \mathbb{A} \boldsymbol{x}_n \end{pmatrix}$$

On a donc

$$B_{i,j} = \boldsymbol{x}_i^* \mathbb{A} \boldsymbol{x_j}, \quad \forall (i,j) \in [\![1,n]\!]^2$$

.

2. On a $\mathbb{A}\boldsymbol{u} = \lambda \boldsymbol{u}$, $\|\boldsymbol{u}\| = 1$, la base $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ est orthonormée et $\boldsymbol{x}_1 = \boldsymbol{u}$. on obtient alors

$$\mathbb{B} = \left(\begin{array}{c|ccc} \lambda \boldsymbol{u}^* \boldsymbol{u} & \boldsymbol{u}^* \mathbb{A} \boldsymbol{x}_2 & \dots & \boldsymbol{u}^* \mathbb{A} \boldsymbol{x}_n \\ \hline \lambda \boldsymbol{x}_2^* \boldsymbol{u} & \boldsymbol{x}_2^* \mathbb{A} \boldsymbol{x}_2 & \dots & \boldsymbol{x}_2^* \mathbb{A} \boldsymbol{x}_n \\ \vdots & \vdots & & \vdots \\ \lambda \boldsymbol{x}_n^* \boldsymbol{u} & \boldsymbol{x}_n^* \mathbb{A} \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n^* \mathbb{A} \boldsymbol{x}_n \end{array}\right) = \left(\begin{array}{c|ccc} \lambda & \boldsymbol{u}^* \mathbb{A} \boldsymbol{x}_2 & \dots & \boldsymbol{u}^* \mathbb{A} \boldsymbol{x}_n \\ \hline 0 & \boldsymbol{x}_2^* \mathbb{A} \boldsymbol{x}_2 & \dots & \boldsymbol{x}_2^* \mathbb{A} \boldsymbol{x}_n \\ \vdots & \vdots & & \vdots \\ 0 & \boldsymbol{x}_n^* \mathbb{A} \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n^* \mathbb{A} \boldsymbol{x}_n \end{array}\right)$$

Q. 3 On veut démontrer, par récurrence faible, la proposition suivante pour $n \geqslant 2$

$(\mathcal{P}_n)$ $\forall \mathbb{A} \in \mathcal{M}_n(\mathbb{C})$, $\exists \mathbb{U} \in \mathcal{M}_n(\mathbb{C})$ unitaire, $\exists \mathbb{T} \in \mathcal{M}_n(\mathbb{C})$ triangulaire supérieure, telles que $\mathbb{A} = \mathbb{U}\mathbb{T}\mathbb{U}^*$.

Initialisation : Montrons que $(\mathcal{P}_2)$ est vérifié.
Soit $\mathbb{A}_2 \in \mathcal{M}_2(\mathbb{C})$. Elle admet au moins un élément propre $(\lambda, \boldsymbol{u})$ (voir Proposition B.40 par ex.) avec $\|\boldsymbol{u}\| = 1$. On peut donc appliquer le résultat de la question précédente : il existe une matrice unitaire $\mathbb{P}_2 \in \mathcal{M}_2(\mathbb{C})$ telle que la matrice $\mathbb{B}_2 = \mathbb{P}_2 \mathbb{A}_2 \mathbb{P}_2^*$ ait comme premier vecteur colonne $(\lambda, 0)^t$. La matrice $\mathbb{B}_2$ est donc triangulaire supérieure et comme $\mathbb{P}_2$ est unitaire on en déduit

$$\mathbb{A}_2 = \mathbb{P}_2^* \mathbb{B}_2 \mathbb{P}_2.$$

On pose $\mathbb{U}_2 = \mathbb{P}_2^*$ matrice unitaire et $\mathbb{T}_2 = \mathbb{B}_2$ matrice triangulaire supérieure pour conclure que la propostion $(\mathcal{P}_2)$ est vraie.

Hérédité : Supposons que $(\mathcal{P}_{n-1})$ soit vérifiée. Montrons que $(\mathcal{P}_n)$ est vraie.
Soit $\mathbb{A}_n \in \mathcal{M}_n(\mathbb{C})$. Elle admet au moins un élément propre $(\lambda, \boldsymbol{u})$ (voir Proposition B.40 par ex.) avec $\|\boldsymbol{u}\| = 1$. On peut donc appliquer le résultat de la question précédente : il existe une matrice unitaire $\mathbb{P}_n \in \mathcal{M}_n(\mathbb{C})$ telle que la matrice $\mathbb{B}_n = \mathbb{P}_n \mathbb{A}_n \mathbb{P}_n^*$ s'écrivent

$$\mathbb{B}_n = \left(\begin{array}{c|c} \lambda & \boldsymbol{c}_{n-1}^* \\ \hline 0 & \\ \vdots & \mathbb{A}_{n-1} \\ 0 & \end{array}\right)$$

où $\boldsymbol{c}_{n-1} \in \mathcal{M}_{n-1,1}(\mathbb{C})$ et $\mathbb{A}_{n-1} \in \mathcal{M}_{n-1}(\mathbb{C})$. Par hypothèse de récurrence, $\exists \mathbb{U}_{n-1} \in \mathcal{M}_{n-1}(\mathbb{C})$ unitaire et $\mathbb{T}_{n-1} \in \mathcal{M}_{n-1}(\mathbb{C})$ triangulaire supérieure telles que

$$\mathbb{A}_{n-1} = \mathbb{U}_{n-1} \mathbb{T}_{n-1} \mathbb{U}_{n-1}^*$$

ou encore

$$\mathbb{T}_{n-1} = \mathbb{U}_{n-1}^* \mathbb{A}_{n-1} \mathbb{U}_{n-1}.$$

Soit $\mathbb{Q}_n \in \mathcal{M}_n(\mathbb{C})$ la matrice définie par

$$\mathbb{Q}_n = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1} & \\ 0 & & & \end{array}\right).$$

La matrice $\mathbb{Q}_n$ est unitaire. En effet on a

$$\mathbb{Q}_n\mathbb{Q}_n^* = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1} & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1}^* & \\ 0 & & & \end{array}\right) = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \underbrace{\mathbb{U}_{n-1}\mathbb{U}_{n-1}^*}_{=\mathbb{I}_{n-1}} & \\ 0 & & & \end{array}\right) = \mathbb{I}_n.$$

On note $\mathbb{T}_n$ la matrice définie par $\mathbb{T}_n = \mathbb{Q}_n^*\mathbb{B}_n\mathbb{Q}_n$. On a alors

$$\mathbb{T}_n = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1}^* & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} \lambda & & \boldsymbol{c}_{n-1}^* & \\ \hline 0 & & & \\ \vdots & & \mathbb{A}_{n-1} & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1} & \\ 0 & & & \end{array}\right)$$

$$= \left(\begin{array}{c|ccc} \lambda & & \boldsymbol{c}_{n-1}^* & \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1}^*\mathbb{A}_{n-1} & \\ 0 & & & \end{array}\right)\left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & \mathbb{U}_{n-1} & \\ 0 & & & \end{array}\right) = \left(\begin{array}{c|ccc} \lambda & & \boldsymbol{c}_{n-1}^*\mathbb{U}_{n-1}^* & \\ \hline 0 & & & \\ \vdots & & \underbrace{\mathbb{U}_{n-1}^*\mathbb{A}_{n-1}\mathbb{U}_{n-1}}_{=\mathbb{T}_{n-1}} = & \\ 0 & & & \end{array}\right)$$

La matrice $\mathbb{T}_n$ est donc triangulaire supérieure et on a par définition de $\mathbb{B}_n$

$$\mathbb{T}_n = \mathbb{Q}_n^*\mathbb{P}_n\mathbb{A}_n\mathbb{P}_n^*\mathbb{Q}_n.$$

On note $\mathbb{U}_n = \mathbb{P}_n^*\mathbb{Q}_n$. Cette matrice est unitaire car les matrices $\mathbb{Q}_n$ et $\mathbb{P}_n$ le sont. En effet, on a

$$\mathbb{U}_n\mathbb{U}_n^* = \mathbb{P}_n^*\mathbb{Q}_n(\mathbb{P}_n^*\mathbb{Q}_n)^* = \mathbb{P}_n^*\underbrace{\mathbb{Q}_n\mathbb{Q}_n^*}_{=\mathbb{I}_n}\mathbb{P}_n = \mathbb{P}_n^*\mathbb{P}_n = \mathbb{I}_n.$$

On a $\mathbb{T}_n = \mathbb{U}_n^*\mathbb{A}_n\mathbb{U}_n$ et en multipliant cette équation à gauche par $\mathbb{U}_n$ et à droite par $\mathbb{U}_n^*$ on obtient l'équation équivalente $\mathbb{A}_n = \mathbb{U}_n\mathbb{T}_n\mathbb{U}_n^*$. La propriété $(\mathcal{P}_n)$ est donc vérifiée. Ce qui achève la démonstration.

Q. 4 Résoudre $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ est équivalent à résoudre

$$\mathbb{U}\mathbb{T}\mathbb{U}^*\boldsymbol{x} = \boldsymbol{b}. \tag{B.55}$$

Comme $\mathbb{U}$ est unitaire, on a $\mathbb{U}\mathbb{U}^* = \mathbb{I}$ et $\mathbb{U}^*$ inversible. Donc en multipliant (B.55) par $\mathbb{U}^*$ on obtient le système équivalent

$$\underbrace{\mathbb{U}^*\mathbb{U}}_{=\mathbb{I}}\mathbb{T}\mathbb{U}^*\boldsymbol{x} = \mathbb{U}^*\boldsymbol{b} \iff \mathbb{T}\mathbb{U}^*\boldsymbol{x} = \mathbb{U}^*\boldsymbol{b}.$$

On pose $\boldsymbol{y} = \mathbb{U}^*\boldsymbol{x}$. Le système précédant se résoud en deux étapes

1. on cherche $\boldsymbol{y}$ solution de $\mathbb{T}\boldsymbol{y} = \mathbb{U}^*\boldsymbol{b}$. Comme $\mathbb{U}$ est unitaire on a $\det(\mathbb{U})\det(\mathbb{U}^*) = \det(\mathbb{I}) = 1$ et donc

$$\begin{aligned}\det(\mathbb{A}) &= \det(\mathbb{U}\mathbb{T}\mathbb{U}^*) = \det(\mathbb{U})\det(\mathbb{T})\det(\mathbb{U}^*) \\ &= \det(\mathbb{T})\end{aligned}$$

   Or $\mathbb{A}$ inversible équivalent à $\det(\mathbb{A}) \neq 0$ et donc la matrice $\mathbb{T}$ est inversible. La matrice $\mathbb{T}$ étant triangulaire supérieure on peut résoudre facilement le système par la *méthode de remontée*.

2. une fois $\boldsymbol{y}$ déterminé, on résoud $\mathbb{U}^*\boldsymbol{x} = \boldsymbol{y}$. Comme $\mathbb{U}$ est unitaire, on obtient directement $\boldsymbol{x} = \mathbb{U}\boldsymbol{y}$.

◇

Inverse d'une matrice

Exercice B.3.7

Soit $\mathbb{A} \in \mathcal{M}_3(\mathbb{R})$ définie par

$$\mathbb{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

Q. 1 *Calculer le déterminant de la matrice $\mathbb{A}$. Que peut-on en conclure?*

Q. 2 *Calculer si possible l'inverse de la matrice $\mathbb{A}$ en utilisant la technique de la matrice augmentée.*

Exercice B.3.8

Soient $\mathbb{A}$ et $\mathbb{B}$, deux matrices de $\mathcal{M}_n(\mathbb{K})$.

Q. 1 *Montrer que*

$$\mathbb{AB} = \mathbb{I} \Rightarrow \mathbb{BA} = \mathbb{I} \tag{B.56}$$

Conclure.

Exercice B.3.9

Q. 1 *Soit $\mathbb{A}$ une matrice inversible et symétrique, montrer que $\mathbb{A}^{-1}$ est symétrique.*

Q. 2 *Soit $\mathbb{A}$ une matrice carrée telle que $\mathbb{I} - \mathbb{A}$ est inversible. Montrer que*

$$\mathbb{A}\left(\mathbb{I} - \mathbb{A}\right)^{-1} = \left(\mathbb{I} - \mathbb{A}\right)^{-1}\mathbb{A}.$$

Q. 3 *Soient $\mathbb{A}$, $\mathbb{B}$ des matrices carrées inversibles de même dimension telle que $\mathbb{A}+\mathbb{B}$ soit inversible. Montrer que*

$$\mathbb{A}\left(\mathbb{A}+\mathbb{B}\right)^{-1}\mathbb{B} = \mathbb{B}\left(\mathbb{A}+\mathbb{B}\right)^{-1}\mathbb{A} = \left(\mathbb{A}^{-1}+\mathbb{B}^{-1}\right)^{-1}$$

Exercice B.3.10

Soit $\mathbb{L} \in \mathcal{M}_n(\mathbb{C})$ une matrice triangulaire inférieure.

Q. 1 *A quelle(s) condition(s) la matrice $\mathbb{L}$ est-elle inversible?*

On suppose $\mathbb{L}$ inversible et on note $\mathbb{X} = \mathbb{L}^{-1}$.

Q. 2 *Montrer que $\mathbb{X}$ est une matrice triangulaire inférieure avec*

$$X_{i,i} = \frac{1}{L_{i,i}}, \ \forall i \in [\![1, n]\!].$$

Correction Exercice B.3.10

Q. 1 La matrice $\mathbb{L}$ est inversible si et seulement si son déterminant est non nul. Or le déterminant d'une matrice triangulaire est égal au produit de ses éléments diagonaux. Pour avoir $\mathbb{L}$, matrice triangulaire, inversible, il est nécessaire et suffisant d'avoir

$$L_{ii} \neq 0, \ \forall i \in [\![1, n]\!].$$

Q. 2 La matrice $\mathbb{X}$ étant la matrice inverse de $\mathbb{L}$, on a

$$\mathbb{LX} = \mathbb{I} \tag{B.57}$$

On note $\boldsymbol{X}^{[j]} = \mathbb{X}_{:,j}$ le j-ème vecteur colonne de la matrice $\mathbb{X}$ et $\boldsymbol{e}^{[j]}$ le j-ème vecteur de la base canonique de $\mathbb{C}^n$ ($e_i^{[j]} = \delta_{i,j}$).
L'équation (B.57) peut donc se réécrire

$$\mathbb{L}\left(\begin{array}{c:c:c} & & \\ \boldsymbol{X}^{[1]} & \cdots & \boldsymbol{X}^{[n]} \\ & & \end{array}\right) = \left(\begin{array}{c:c:c} & & \\ \boldsymbol{e}^{[1]} & \cdots & \boldsymbol{e}^{[n]} \\ & & \end{array}\right)$$

ou encore, déterminer la matrice $\mathbb{X}$ inverse de $\mathbb{L}$ reviens à résoudre les n systèmes linéaires suivants:

$$\mathbb{L}\boldsymbol{X}^{[j]} = \boldsymbol{e}^{[j]}, \quad \forall j \in [\![1, n]\!]. \tag{B.58}$$

- Pour montrer que $\mathbb{X}$ est triangulaire inférieure il suffit de vérifier que pour tout $j \in [\![2, n]\!]$

$$X_i^{[j]} = 0, \quad \forall i \in [\![1, j-1]\!].$$

Soit $j \in [\![2, n]\!]$. On décompose la matrice $\mathbb{L}$ en la matrice bloc carré 2 par 2 ou le premier bloc diagonal, noté $\mathbb{L}_{j-1}$, est une matrice triangulaire inférieure inversible de dimension $j-1$. Le système (B.58) s'écrit alors

On en déduit donc que nécessairement

$$\begin{pmatrix} L_{1,1} & 0 & \cdots & 0 \\ \bullet & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \bullet & \cdots & \bullet & L_{j-1,j-1} \end{pmatrix} \begin{pmatrix} X_1^{[j]} \\ \vdots \\ X_{j-1}^{[j]} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

Comme la matrice de ce système est inversible, on a bien $X_i^{[j]} = 0$, $\forall i \in [\![1, j-1]\!]$ et donc la matrice $\mathbb{X}$ est triangulaire inférieure.

- Par définition du produit matricielle, de l'équation (B.57) on tire pour tout $j \in [\![1, n]\!]$

$$(\mathbb{L}\mathbb{X})_{j,j} = (\mathbb{I})_{j,j} \iff \sum_{k=1}^{n} L_{j,k}X_{k,j} = 1 \iff \sum_{k=1}^{j-1} L_{j,k}X_{k,j} + L_{j,j}X_{j,j} + \sum_{k=j+1}^{n} L_{j,k}X_{k,j} = 1$$

Or les matrices $\mathbb{L}$ et $\mathbb{X}$ sont triangulaires inférieures et donc $X_{k,j} = 0$, $\forall k \in [\![1, j-1]\!]$, et $L_{j,k} = 0$, $\forall k \in [\![j+1, n]\!]$. On obtient alors $L_{j,j}X_{j,j} = 1$ et comme $L_{j,j} \neq 0$ on a bien $X_{j,j} = 1/L_{j,j}$.

◇

Exercice B.3.11

Soit $\mathbb{A} \in \mathcal{M}_{n,n}(\mathbb{K})$ et $\mathbb{U}$, $\mathbb{B}$, $\mathbb{V}$ trois matrices rectangulaires.

Q. 1 *Sous quelles hypothèses peut-on définir la matrice $\mathbb{G}$ suivante*

$$\mathbb{G} = \mathbb{A}^{-1} - \mathbb{A}^{-1}\mathbb{U}\left(\mathbb{I} + \mathbb{B}\mathbb{V}\mathbb{A}^{-1}\mathbb{U}\right)^{-1}\mathbb{B}\mathbb{V}\mathbb{A}^{-1}? \tag{B.59}$$

Q. 2 *Montrer que $(\mathbb{A} + \mathbb{U}\mathbb{B}\mathbb{V})\,\mathbb{G} = \mathbb{I}$. Conclure.*

Q. 3 *Soit $\beta \in \mathbb{R}$ et $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^n$. Calculer $(\mathbb{A} + \beta\boldsymbol{u}\boldsymbol{v}^t)^{-1}$ en fonction de l'inverse de $\mathbb{A}$.*

Exercice B.3.12

Etant donnée une matrice $\mathbb{D} \in \mathcal{M}_{n,n}(\mathbb{C})$, on pose

$$\mathbb{D} = \mathbb{A} + \imath\mathbb{B} \text{ avec } \mathbb{A}, \mathbb{B} \in \mathcal{M}_{n,n}(\mathbb{R})$$

Sous certaines hypothèses à préciser, établir la relation

$$\mathbb{D}^{-1} = \left(\mathbb{A} + \mathbb{B}\mathbb{A}^{-1}\mathbb{B}\right)^{-1} - \imath\mathbb{A}^{-1}\mathbb{B}\left(\mathbb{A} + \mathbb{B}\mathbb{A}^{-1}\mathbb{B}\right)^{-1}.$$

Exercice B.3.13

Soient $(z_i)_{i=0}^n$ $n+1$ points distincts 2 à 2 de $\mathbb{C}$. Soit $\mathbb{V} \in \mathcal{M}_{n+1}(\mathbb{C})$ la matrice définie par

$$\mathbb{V}_{i,j} = z_{i-1}^{j-1}, \ \forall (i,j) \in [\![1, n+1]\!].$$

Q. 1 *Ecrire la matrice* $\mathbb{V}$.

Soient $\boldsymbol{w} = (w_i)_{i=1}^{n+1}$ un vecteur de $\mathbb{C}^{n+1}$. On note $\mathrm{P}_{\boldsymbol{w}} \in \mathbb{C}_n[X]$, le polynôme défini par

$$\mathrm{P}_{\boldsymbol{w}}(z) = \sum_{i=0}^{n} w_{i+1} z^i$$

Q. 2 *Exprimer* $\boldsymbol{v} = \mathbb{V}\boldsymbol{w}$ *en fonction de* $\mathrm{P}_{\boldsymbol{w}}$.

Q. 3 *En déduire que* $\mathbb{V}$ *est inversible.*

Correction Exercice B.3.13

Q. 1 On a

$$\mathbb{V} = \begin{pmatrix} 1 & z_0 & \cdots & z_0^n \\ 1 & z_1 & \cdots & z_1^n \\ \vdots & \vdots & & \vdots \\ 1 & z_n & \cdots & z_n^n \end{pmatrix}$$

Q. 2 On a $\boldsymbol{v} \in \mathbb{C}^{n+1}$ et, pour tout $i \in [\![1, n+1]\!]$,

$$\begin{aligned} \boldsymbol{v}_i &= \sum_{j=1}^{n+1} \mathbb{V}_{i,j} \boldsymbol{w}_j \\ &= \sum_{j=1}^{n+1} z_{i-1}^{j-1} w_j \\ &= \sum_{j=0}^{n} w_{j+1} z_{i-1}^{j} = \mathrm{P}_{\boldsymbol{w}}(z_{i-1}). \end{aligned}$$

c'est à dire

$$\boldsymbol{v} = \begin{pmatrix} \mathrm{P}_{\boldsymbol{w}}(z_0) \\ \vdots \\ \mathrm{P}_{\boldsymbol{w}}(z_n) \end{pmatrix}.$$

Q. 3 La matrice $\mathbb{V}$ est inversible si et seulement si son noyau est réduit à l'élément nul, c'est à dire

$$\ker(\mathbb{V}) = \{\mathbf{0}\}.$$

Soit $\boldsymbol{u} = (u_1, \ldots, u_{n+1})^* \in \mathbb{C}^{n+1}$, tel que $\mathbb{V}\boldsymbol{u} = \mathbf{0}$, montrons qu'alors $\boldsymbol{u} = \mathbf{0}$.
On a

$$\mathbb{V}\boldsymbol{u} = \mathbb{V} \begin{pmatrix} u_1 \\ \vdots \\ u_{n+1} \end{pmatrix} = \begin{pmatrix} \mathrm{P}_{\boldsymbol{u}}(z_0) \\ \vdots \\ \mathrm{P}_{\boldsymbol{u}}(z_n) \end{pmatrix} = \mathbf{0}.$$

Les $n+1$ points $(z_i)_{i=0}^n$ sont distincts 2 à 2, donc le polynôme $\boldsymbol{u}$ admet $n+1$ racines distinctes hors $\mathrm{P}_{\boldsymbol{u}} \in \mathbb{C}_n[X]$, c'est donc le polynôme nul, c'est à dire $u_i = 0$, $\forall i \in [\![1, n+1]\!]$. On a donc $\boldsymbol{u} = \mathbf{0}$.
La matrice $\mathbb{V}$ est donc inversible.

◇

Matrices blocs

Exercice B.3.14

On considère les matrices blocs suivantes

$$\mathbb{A} = \left(\begin{array}{cc:cc} 1 & 2 & 1 & 0 \\ 3 & 4 & 0 & 1 \\ \hdashline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{array}\right) = \left(\begin{array}{c:c} \mathbb{C} & \mathbb{I} \\ \hdashline \mathbb{I} & 0 \end{array}\right) \quad \text{et} \quad \mathbb{B} = \left(\begin{array}{cc:cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hdashline 1 & 2 & 1 & 2 \\ 3 & 4 & 3 & 4 \end{array}\right) = \left(\begin{array}{c:c} \mathbb{I} & 0 \\ \hdashline \mathbb{C} & \mathbb{C} \end{array}\right)$$

avec par identification

$$\mathbb{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{et} \quad \mathbb{C} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

Q. 1 *Calculer les matrices $\mathbb{AB}$ et $\mathbb{BA}$ en utilisant l'écriture bloc.*

Q. 2 *Exprimer les matrices $\mathbb{A}(\mathbb{A}+\mathbb{B})$ et $(2\mathbb{B}-\mathbb{A})(\mathbb{B}+\mathbb{A})$ en fonction des matrices $\mathbb{C}$ et $\mathbb{I}$.*

Exercice B.3.15

Soient $\mathbb{A} \in \mathcal{M}_{n,k}(\mathbb{K})$ et $\mathbb{B} \in \mathcal{M}_{k,n}(\mathbb{K})$. On note $\mathbb{L}$ la matrice

$$\mathbb{L} = \left(\begin{array}{c:c} \mathbb{I} - \mathbb{BA} & \mathbb{B} \\ \hdashline 2\mathbb{A} - \mathbb{ABA} & \mathbb{AB} - \mathbb{I} \end{array}\right).$$

Q. 1 *Montrer que la matrice $\mathbb{L}$ est bien définie et spécifier les dimensions des blocs.*

Q. 2 *Calculer $\mathbb{L}^2$. Que peut-on en conclure?*

Exercice B.3.16: ***résultats à savoir*** ★★★★★

Soient $\mathbb{A} \in \mathcal{M}_m(\mathbb{C})$, $\mathbb{B} \in \mathcal{M}_n(\mathbb{C})$ et $\mathbb{D} \in \mathcal{M}_{m,n}(\mathbb{C})$.

Q. 1 *Calculer, en fonction des déterminant de $\mathbb{A}$ et $\mathbb{B}$, le déterminant des matrices*

$$\mathbb{E} = \left(\begin{array}{c:c} \mathbb{A} & \mathbb{O} \\ \hdashline \mathbb{O} & \mathbb{I}_n \end{array}\right), \quad \mathbb{F} = \left(\begin{array}{c:c} \mathbb{I}_m & \mathbb{O} \\ \hdashline \mathbb{O} & \mathbb{B} \end{array}\right), \quad \textit{et } \mathbb{G} = \left(\begin{array}{c:c} \mathbb{A} & 0 \\ \hdashline 0 & \mathbb{B} \end{array}\right).$$

Q. 2 *Soit $\mathbb{H} = \left(\begin{array}{c:c} \mathbb{A} & \mathbb{D} \\ \hdashline 0 & \mathbb{B} \end{array}\right)$. En utilisant les factorisations $\mathbb{QR}$ des matrices $\mathbb{A}$ et $\mathbb{B}$, montrer que*

$$\det(\mathbb{H}) = \det(\mathbb{A}) \det(\mathbb{B}). \tag{B.60}$$

Q. 3 *En déduire qu'**une matrice triangulaire supérieure par blocs est inversible si et seulement si ses matrices blocs diagonales sont inversibles.***

Q. 4 *En déduire qu'**une matrice triangulaire inférieure par blocs est inversible si et seulement si ses matrices blocs diagonales sont inversibles.***

B.3.2 Normes

Normes vectorielles

Exercice B.3.17

Soient $\boldsymbol{x}$ et $\boldsymbol{y}$ deux vecteurs de $\mathbb{C}^n$.

Q. 1 *Trouver $\alpha \in \mathbb{C}$ tel que $\langle \alpha\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x}\rangle = 0$.*

Q. 2 *En calculant $\|\alpha\boldsymbol{x} - \boldsymbol{y}\|_2^2$, montrer que*

$$|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leqslant \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 . \tag{B.61}$$

Q. 3 *Soit $\boldsymbol{x} \neq 0$. Montrer alors que l'inégalité (B.61) est une égalité si et seulement si $\boldsymbol{y} = \alpha\boldsymbol{x}$.*

Correction Exercice B.3.17

Q. 1

- Si $\boldsymbol{x} = 0$, alors α quelconque.
- Si $\boldsymbol{x} \neq 0$, alors

$$\langle \alpha\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x}\rangle = 0 \iff \overline{\alpha}\langle \boldsymbol{x}, \boldsymbol{x}\rangle - \langle \boldsymbol{y}, \boldsymbol{x}\rangle = 0$$

Or $\boldsymbol{x} \neq 0$, ce qui donne

$$\overline{\alpha} = \frac{\langle \boldsymbol{y}, \boldsymbol{x}\rangle}{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}.$$

et , comme $\langle \boldsymbol{x}, \boldsymbol{x}\rangle \in \mathbb{R}$ et $\overline{\langle \boldsymbol{y}, \boldsymbol{x}\rangle} = \langle \boldsymbol{x}, \boldsymbol{y}\rangle$, on obtient

$$\alpha = \frac{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}. \tag{B.62}$$

Q. 2 On a

$$\begin{aligned} \|\alpha\boldsymbol{x} - \boldsymbol{y}\|_2^2 &= \langle \alpha\boldsymbol{x} - \boldsymbol{y}, \alpha\boldsymbol{x} - \boldsymbol{y}\rangle \\ &= \alpha\langle \alpha\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x}\rangle - \langle \alpha\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{y}\rangle \\ &= -\langle \alpha\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{y}\rangle , \quad \text{car } \langle \alpha\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x}\rangle = 0 \\ &= -\overline{\alpha}\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \langle \boldsymbol{y}, \boldsymbol{y}\rangle \end{aligned}$$

En utilisant (B.62), on obtient alors

$$\begin{aligned} \|\alpha\boldsymbol{x} - \boldsymbol{y}\|_2^2 &= -\frac{\langle \boldsymbol{y}, \boldsymbol{x}\rangle}{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \langle \boldsymbol{y}, \boldsymbol{y}\rangle \\ &= \frac{-\langle \boldsymbol{y}, \boldsymbol{x}\rangle\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \langle \boldsymbol{y}, \boldsymbol{y}\rangle\langle \boldsymbol{x}, \boldsymbol{x}\rangle}{\langle \boldsymbol{x}, \boldsymbol{x}\rangle} \end{aligned}$$

Comme $\langle \boldsymbol{y}, \boldsymbol{x}\rangle = \overline{\langle \boldsymbol{x}, \boldsymbol{y}\rangle}$, on a $\langle \boldsymbol{y}, \boldsymbol{x}\rangle\langle \boldsymbol{x}, \boldsymbol{y}\rangle = |\langle \boldsymbol{x}, \boldsymbol{y}\rangle|^2$ et donc

$$\begin{aligned} \|\alpha\boldsymbol{x} - \boldsymbol{y}\|_2^2 &= \frac{1}{\langle \boldsymbol{x}, \boldsymbol{x}\rangle}\left(-|\langle \boldsymbol{x}, \boldsymbol{y}\rangle|^2 + \|\boldsymbol{x}\|_2^2 \|\boldsymbol{y}\|_2^2\right) \\ &\geqslant 0. \end{aligned} \tag{B.63}$$

On a alors

$$|\langle \boldsymbol{x}, \boldsymbol{y}\rangle|^2 \leqslant \|\boldsymbol{x}\|_2^2 \|\boldsymbol{y}\|_2^2$$

La fonction $x \mapsto \sqrt{x}$ étant croissante sur $[0; +\infty[$, on obtient (B.61).

Q. 3 Soit $\boldsymbol{x} \neq 0$. On veut montrer que

$$|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| = \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \iff \boldsymbol{y} = \alpha\boldsymbol{x}$$

$\boxed{\Leftarrow}$ On suppose $\boldsymbol{y} = \alpha\boldsymbol{x}$. On a alors

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \alpha \langle \boldsymbol{x}, \boldsymbol{x} \rangle = \alpha \|\boldsymbol{x}\|_2^2 \implies |\langle \boldsymbol{x}, \boldsymbol{y} \rangle| = |\alpha| \|\boldsymbol{x}\|_2^2 .$$

Comme $\|\boldsymbol{y}\|_2 = |\alpha| \|\boldsymbol{x}\|_2$, on a aussi

$$\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 = |\alpha| \|\boldsymbol{x}\|_2^2 .$$

On en déduit alors

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| = \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 .$$

$\boxed{\Rightarrow}$ On suppose $|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| = \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2$. Avec cette hypothèse, l'équation (B.63) devient

$$\|\alpha\boldsymbol{x} - \boldsymbol{y}\|_2^2 = 0$$

et donc $\alpha\boldsymbol{x} - \boldsymbol{y} = 0$, c'est à dire $\boldsymbol{y} = \alpha\boldsymbol{x}$.

◇

Exercice B.3.18

Soient $\boldsymbol{x}$ et $\boldsymbol{y}$ deux vecteurs de $\mathbb{C}^n$.

Q. 1 *Démontrer l'**inégalité triangulaire***

$$\|\boldsymbol{x} + \boldsymbol{y}\|_2 \leqslant \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2 . \tag{B.64}$$

Q. 2 *Si $\boldsymbol{x}$ et $\boldsymbol{y}$ sont non nuls, prouver que l'inégalité (B.64) est une égalité si et seulement si $\boldsymbol{y} = \alpha\boldsymbol{x}$ avec α un réel strictement positif.*

Q. 3 *Déduire de (B.64) l'inégalité suivante :*

$$|\,\|\boldsymbol{x}\|_2 - \|\boldsymbol{y}\|_2\,| \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|_2 . \tag{B.65}$$

Q. 4 *Soient $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$, p vecteurs de $\mathbb{C}^n$. Montrer que*

$$\left\| \sum_{i=1}^{p} \boldsymbol{x}_i \right\|_2 \leqslant \sum_{i=1}^{p} \|\boldsymbol{x}_i\|_2 . \tag{B.66}$$

Correction Exercice B.3.18

Q. 1 On rappele que $\|\boldsymbol{x}\|_2^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle$. On obtient, en utilisant les propriétés du produit scalaire,

$$\begin{aligned} \|\boldsymbol{x} + \boldsymbol{y}\|_2^2 &= \langle \boldsymbol{x} + \boldsymbol{y}, \boldsymbol{x} + \boldsymbol{y} \rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle + \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \langle \boldsymbol{y}, \boldsymbol{x} \rangle + \langle \boldsymbol{x}, \boldsymbol{x} \rangle \\ &= \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 + \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \langle \boldsymbol{y}, \boldsymbol{x} \rangle . \end{aligned}$$

Pour tout nombre complexe z, on a $z + \overline{z} = 2\,\mathrm{Re}(z)$, et $|z| \geqslant \mathrm{Re}(z)$. [1] Comme $\langle \boldsymbol{y}, \boldsymbol{x} \rangle = \overline{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}$, on a

$$\begin{aligned} \|\boldsymbol{x} + \boldsymbol{y}\|_2^2 &= \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 + 2\,\mathrm{Re}\left(\langle \boldsymbol{x}, \boldsymbol{y} \rangle\right) \\ &\leqslant \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 + 2|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| . \end{aligned} \tag{B.67}$$

L'inégalité de Cauchy-Schwarz donne

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leqslant \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2$$

et donc

$$\begin{aligned} \|\boldsymbol{x} + \boldsymbol{y}\|_2^2 &\leqslant \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 + 2\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \\ &= \left(\|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2\right)^2 . \end{aligned}$$

La fonction $x \mapsto \sqrt{x}$ étant croissante sur $[0; +\infty[$, on obtient

$$\|\boldsymbol{x} + \boldsymbol{y}\|_2 \leqslant \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2 .$$

[1] En effet, $z = a + \imath b$ et $\overline{z} = a - \imath b$ d'où $z + \overline{z} = 2a$. De plus, $|z|^2 = a^2 + b^2 \geqslant a^2$ ce qui donne $|z| \geqslant |a| \geqslant a$.

Q. 2 Soient $\boldsymbol{x}$ et $\boldsymbol{y}$ deux vecteurs de $\mathbb{C}^n$ non nuls. On veut démontrer que

$$\|\boldsymbol{x}+\boldsymbol{y}\|_2 = \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2 \Longleftrightarrow \boldsymbol{y} = \alpha\boldsymbol{x},\ \alpha > 0.$$

$\boxed{\Leftarrow}$ On suppose $\boldsymbol{y} = \alpha\boldsymbol{x}$ avec $\alpha > 0$.
On a alors

$$\begin{aligned}\|\boldsymbol{x}+\boldsymbol{y}\|_2 &= \|\boldsymbol{x}+\alpha\boldsymbol{x}\|_2\\ &= \|(1+\alpha)\boldsymbol{x}\|_2\\ &= |1+\alpha|\,\|\boldsymbol{x}\|_2\,.\end{aligned}$$

Comme $\alpha > 0$, on a $|1+\alpha| = 1+\alpha$ et donc

$$\|\boldsymbol{x}+\boldsymbol{y}\|_2 = \|\boldsymbol{x}\|_2 + \alpha\,\|\boldsymbol{x}\|_2\,. \tag{B.68}$$

De plus, on a

$$\begin{aligned}\|\boldsymbol{y}\|_2 &= \|\alpha\boldsymbol{x}\|_2\\ &= |\alpha|\,\|\boldsymbol{x}\|_2\\ &= \alpha\,\|\boldsymbol{x}\|_2\,,\quad \text{car}\quad \alpha > 0.\end{aligned}$$

D'après (B.68), on en déduit

$$\|\boldsymbol{x}+\boldsymbol{y}\|_2 = \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2\,.$$

$\boxed{\Rightarrow}$ On suppose que

$$\|\boldsymbol{x}+\boldsymbol{y}\|_2 = \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2\,. \tag{B.69}$$

Ce qui donne

$$\|\boldsymbol{x}+\boldsymbol{y}\|_2^2 = \|\boldsymbol{x}\|_2^2 + \|\boldsymbol{y}\|_2^2 + 2\,\|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2 \tag{B.70}$$

On déduit alors de l'égalité (B.67) que

$$\mathrm{Re}\,(\langle\boldsymbol{x},\boldsymbol{y}\rangle) = \|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2\,. \tag{B.71}$$

Or, on a, $\forall z \in \mathbb{C},\ \mathrm{Re}(z) \leqslant |z|$ et donc, en utilisant l'inégalité de Cauchy-Schwarz

$$\|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2 = \mathrm{Re}\,(\langle\boldsymbol{x},\boldsymbol{y}\rangle) \leqslant |\langle\boldsymbol{x},\boldsymbol{y}\rangle| \leqslant \|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2$$

ce qui impose

$$|\langle\boldsymbol{x},\boldsymbol{y}\rangle| = \|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2\,.$$

D'après l'exercice précédant, cette égalité est vérifiée si et seulement si $\boldsymbol{y} = \alpha\boldsymbol{x}$ avec $\alpha = \frac{\langle\boldsymbol{y},\boldsymbol{x}\rangle}{\langle\boldsymbol{x},\boldsymbol{x}\rangle}$.
Il nous reste à vérifier que $\alpha > 0$.
L'hypothèse (B.69) avec $\boldsymbol{y} = \alpha\boldsymbol{x}$ devient

$$\|\boldsymbol{x}+\boldsymbol{y}\|_2 = |1+\alpha|\,\|\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2 + \|\boldsymbol{y}\|_2 = 1 + |\alpha|\,\|\boldsymbol{x}\|_2\,. \tag{B.72}$$

On a donc

$$\begin{aligned}|1+\alpha| = 1+|\alpha| &\Longleftrightarrow |1+\alpha|^2 = (1+|\alpha|)^2\\ &\Longleftrightarrow (1+\alpha)\overline{(1+\alpha)} = 1+2|\alpha|+|\alpha|^2\\ &\Longleftrightarrow (1+\alpha)(1+\overline{\alpha}) = 1+2|\alpha|+|\alpha|^2\\ &\Longleftrightarrow 1+\alpha\overline{\alpha}+\alpha+\overline{\alpha} = 1+2|\alpha|+|\alpha|^2\\ &\Longleftrightarrow \mathrm{Re}(\alpha) = |\alpha|\end{aligned}$$

Donc α est un réel et $\alpha = \mathrm{Re}(\alpha) = |\alpha| \geqslant 0$.
Comme $\boldsymbol{y} = \alpha\boldsymbol{x}$ et $\boldsymbol{y} \neq 0$, on a $\alpha \neq 0$ et donc $\alpha > 0$.

Q. 3 On a $\boldsymbol{x} = (\boldsymbol{x}-\boldsymbol{y})+\boldsymbol{y}$, et par application de l'inégalité triangulaire (B.64) on obtient

$$\|\boldsymbol{x}\|_2 = \|(\boldsymbol{x}-\boldsymbol{y})+\boldsymbol{y}\|_2 \leqslant \|\boldsymbol{x}-\boldsymbol{y}\|_2 + \|\boldsymbol{y}\|_2 \implies \|\boldsymbol{x}\|_2 - \|\boldsymbol{y}\|_2 \leqslant \|\boldsymbol{x}-\boldsymbol{y}\|_2 .$$

De même, avec $\boldsymbol{y} = (\boldsymbol{y}-\boldsymbol{x})+\boldsymbol{x}$, et par application de l'inégalité triangulaire (B.64) on a

$$\|\boldsymbol{y}\|_2 = \|(\boldsymbol{y}-\boldsymbol{x})+\boldsymbol{x}\|_2 \leqslant \|\boldsymbol{y}-\boldsymbol{x}\|_2 + \|\boldsymbol{x}\|_2 \implies \|\boldsymbol{y}\|_2 - \|\boldsymbol{x}\|_2 \leqslant \|\boldsymbol{x}-\boldsymbol{y}\|_2 .$$

En combinant ces deux inégalités, on obtient alors

$$|\,\|\boldsymbol{x}\|_2 - \|\boldsymbol{y}\|_2\,| \leqslant \|\boldsymbol{x}-\boldsymbol{y}\|_2 .$$

Q. 4 On effectue une démonstration par récurrence.
Soit $n \in \mathbb{N}$, $n > 1$. On définit la propriété $\mathcal{P}(n)$ par

$$\mathcal{P}(n) \ : \quad \left\| \sum_{i=1}^{n} \boldsymbol{x}_i \right\|_2 \leqslant \sum_{i=1}^{n} \|\boldsymbol{x}_i\|_2 . \tag{B.73}$$

- On a démontré, en Q.1, que la propriété $\mathcal{P}(2)$ est vraie.
- Soit $n > 2$, on suppose que $\mathcal{P}(n)$ est vérifiée (hypothèse de récurrence). On veut alors montrer que $\mathcal{P}(n+1)$ est vraie.
  On a

$$\begin{aligned}
\left\| \sum_{i=1}^{n+1} \boldsymbol{x}_i \right\|_2 &= \left\| \sum_{i=1}^{n} \boldsymbol{x}_i + \boldsymbol{x}_{n+1} \right\|_2 \\
&\leqslant \left\| \sum_{i=1}^{n} \boldsymbol{x}_i \right\|_2 + \|\boldsymbol{x}_{n+1}\|_2 , \quad \text{d'après (B.64)} \\
&\leqslant \sum_{i=1}^{n} \|\boldsymbol{x}_i\|_2 + \|\boldsymbol{x}_{n+1}\|_2 , \quad \text{car } \mathcal{P}(n) \text{ est vérifiée} \\
&= \sum_{i=1}^{n+1} \|\boldsymbol{x}_i\|_2 .
\end{aligned}$$

  La propriété $\mathcal{P}(n+1)$ est donc vérifiée.

On a donc démontré par récurrence que la propriété $\mathcal{P}(n)$ est vraie pour tout $n \geqslant 2$.

◇

Exercice B.3.19

Q. 1 *Soit la fonction $f(t) = (1-\lambda)+\lambda t - t^{\lambda}$ avec $0 < \lambda < 1$. Montrer que pour tous $\alpha \geqslant 0$ et $\beta \geqslant 0$ on a*

$$\alpha^{\lambda}\beta^{1-\lambda} \leqslant \lambda\alpha + (1-\lambda)\beta. \tag{B.74}$$

Soient $\boldsymbol{x}$ et $\boldsymbol{y}$ deux vecteurs non nuls de $\mathbb{C}^n$. Soient $p > 1$ et $q > 1$ vérifiant $\frac{1}{p} + \frac{1}{q} = 1$.

Q. 2 *On pose $\boldsymbol{u} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_p}$ et $\boldsymbol{v} = \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|_q}$. En utilisant l'inégalite (B.74), montrer que l'on a l'inégalité*

$$\sum_{i=1}^{n} |u_i v_i| \leqslant \frac{1}{p}\sum_{i=1}^{n} |u_i|^p + \frac{1}{q}\sum_{i=1}^{n} |v_i|^q = 1. \tag{B.75}$$

Q. 3 *En déduire l'inégalité de Holder suivante*

$$|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leqslant \sum_{i=1}^{n} |x_i y_i| \leqslant \|\boldsymbol{x}\|_p \, \|\boldsymbol{y}\|_q . \tag{B.76}$$

Quel est le lien entre l'inégalité de Hölder et l'inégalité de Cauchy-Schwarz?

Correction Exercice B.3.19

Q. 1 L'inégalité (B.74) est vérifiée si $\alpha = 0$ ou $\beta = 0$. Il nous reste donc à la vérifier pour $\alpha > 0$ et $\beta > 0$. Dans ce cas (B.74) s'écrit

$$\left(\frac{\alpha}{\beta}\right)^{\lambda} \leqslant \lambda\frac{\alpha}{\beta} + (1-\lambda)$$

c'est à dire

$$f(\frac{\alpha}{\beta}) \geqslant 0.$$

Montrons que $f(t) \geqslant 0$, $\forall t \in]0, +\infty[$.
On a $f'(t) = \lambda(1 - t^{\lambda-1})$ et

$$f'(t) = 0 \quad \Leftrightarrow \quad 1 - t^{\lambda-1} = 0, \text{car } \lambda \neq 0$$

De plus, on a $t^{\lambda-1} = e^{(\lambda-1)\ln(t)}$ et comme $\lambda - 1 \neq 0$, on obtient

$$f'(t) = 0 \Leftrightarrow t = 1.$$

- Etudions la fonction sur $]0, 1[$. On a pour $t \in]0, 1[$, $\ln(t) < 0$ et donc $(\lambda - 1)\ln(t) > 0$. Comme la fonction exp est croissante, on en déduit $\exp((\lambda - 1)\ln(t)) > 1$ et alors $f'(t) < 0$.

- Etudions la fonction sur $]1, +\infty[$. On a pour $t \in]1, +\infty[$, $\ln(t) > 0$ et donc $(\lambda - 1)\ln(t) < 0$. Comme la fonction exp est croissante, on en déduit $0 < \exp((\lambda - 1)\ln(t)) < 1$ et alors $f'(t) > 0$.

Le minimum de f est donc atteint en $t = 1$ et on a

$$\forall t \in]0, +\infty[, \ f(x) \geqslant f(1) = 0.$$

L'inégalité (B.74) est donc vérifiée $\forall \alpha \geqslant 0$, $\forall \beta \geqslant 0$ et $\forall \lambda \in]0, 1[$.

Q. 2 On pose $\lambda = \frac{1}{p} \in]0, 1[$. on a alors $1 - \lambda = \frac{1}{q}$. On pose

$$\alpha = |u_i|^p \geqslant 0, \quad \beta = |v_i|^q \geqslant 0.$$

En utilisant (B.74), on obtient directement

$$|u_i||v_i| \leqslant \frac{1}{p}|u_i|^p + \frac{1}{q}|v_i|^q, \ \forall i \in [\![1, n]\!].$$

En sommant sur i on obtient:

$$\sum_{i=1}^{n} |u_i v_i| \leqslant \frac{1}{p}\sum_{i=1}^{n} |u_i|^p + \frac{1}{q}\sum_{i=1}^{n} |v_i|^q = \frac{1}{p}\|\boldsymbol{u}\|_p^p + \frac{1}{q}\|\boldsymbol{v}\|_q^q$$

Comme par construction $\|\boldsymbol{u}\|_p = \|\boldsymbol{v}\|_q = 1$, on obtient

$$\sum_{i=1}^{n} |u_i v_i| \leqslant \frac{1}{p} + \frac{1}{q} = 1.$$

Q. 3 Par construction, on a

$$\sum_{i=1}^{n} |u_i v_i| = \frac{1}{\|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q} \sum_{i=1}^{n} |x_i y_i|$$

et donc en utilisant l'inégalité (B.76) on obtient

$$\sum_{i=1}^{n} |x_i y_i| \leqslant \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q .$$

De plus

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| = |\sum_{i=1}^{n} \overline{x}_i y_i| \leqslant \sum_{i=1}^{n} |\overline{x}_i y_i| = \sum_{i=1}^{n} |x_i y_i| \leqslant \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q .$$

Pour $p = q = 2$, l'inégalité de Hölder entraine l'inégalité de Cauchy-Schwarz.

◇

Exercice B.3.20

Soit $p > 1$ et q le nombre tel que $\frac{1}{q} = 1 - \frac{1}{p}$.

Q. 1 *Vérifier que* $\forall(\alpha, \beta) \in \mathbb{C}^2$ *on a*

$$|\alpha + \beta|^p \leqslant |\alpha||\alpha + \beta|^{p/q} + |\beta||\alpha + \beta|^{p/q}. \tag{B.77}$$

Q. 2 *En utilisant l'inégalité de Hölder et (B.77), démontrer l'**inégalité de Minkowski** :*

$$\|\boldsymbol{x} + \boldsymbol{y}\|_p \leqslant \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p\,,\ \forall \boldsymbol{x} \in \mathbb{C}^n,\ \forall \boldsymbol{y} \in \mathbb{C}^n,\ p \geqslant 1. \tag{B.78}$$

Normes matricielles

Exercice B.3.21: Norme de Frobenious

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$. On définit l'application $\|\bullet\|_F$ par

$$\|\mathbb{A}\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2. \tag{B.79}$$

Q. 1 *On note, respectivement,* $\boldsymbol{A}_{i,:}$, $\forall i \in [\![1, n]\!]$ *et* $\boldsymbol{A}_{:,j}$, $\forall j \in [\![1, n]\!]$ *les vecteurs lignes et colonnes de* $\mathbb{A}$. *Montrer que*

$$\|\mathbb{A}\|_F^2 = \sum_{i=1}^{n} \|\boldsymbol{A}_{i,:}\|_2^2 = \sum_{j=1}^{n} \|\boldsymbol{A}_{:,j}\|_2^2 = \operatorname{tr} \mathbb{A}^*\mathbb{A}. \tag{B.80}$$

Q. 2 *Montrer que*

$$\|\mathbb{A}\boldsymbol{x}\|_2 \leqslant \|\mathbb{A}\|_F \|\boldsymbol{x}\|_2\,,\ \forall \boldsymbol{x} \in \mathbb{C}^n. \tag{B.81}$$

Q. 3 *Montrer que cette application est une norme matricielle (nommée norme de Frobenius).*

Q. 4 *Calculer* $\|\mathbb{A}^*\|_F$ *et* $\|\mathbb{I}_n\|_F$ *où* $\mathbb{I}_n$ *est la matrice identitée de* $\mathcal{M}_n(\mathbb{C})$.

Exercice B.3.22

Soit $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{C})$. Montrer les propriétés suivantes

1. $\|\mathbb{A}\|_2 = \max\limits_{\substack{\|\boldsymbol{x}\|_2=1 \\ \|\boldsymbol{y}\|_2=1}} |\langle \mathbb{A}\boldsymbol{x}, \boldsymbol{y}\rangle|$.
2. $\|\mathbb{A}\|_2 = \|\mathbb{A}^*\|_2$.
3. $\|\mathbb{A}^*\mathbb{A}\|_2 = \|\mathbb{A}\|_2^2$.
4. $\|\mathbb{U}^*\mathbb{A}\mathbb{V}\|_2 = \|\mathbb{A}\|_2$ quand $\mathbb{U}\mathbb{U}^* = \mathbb{I}$ et $\mathbb{V}\mathbb{V}^* = \mathbb{I}$

Exercice B.3.23

Soient $\mathbb{A} \in \mathcal{M}_{m,n}(\mathbb{C})$, $\mathbb{B} \in \mathcal{M}_{n,l}(\mathbb{C})$, $\boldsymbol{x} \in \mathbb{C}^n$ et $p \geqslant 1$. Montrer que

$$\begin{aligned} \|\mathbb{A}\boldsymbol{x}\|_p &\leqslant \|\mathbb{A}\|_p \|\boldsymbol{x}\|_p && \text{(B.82)} \\ \|\mathbb{A}\|_p &= \max_{\|\boldsymbol{x}\|_p=1} \|\mathbb{A}\boldsymbol{x}\|_p && \text{(B.83)} \\ \|\mathbb{A}\|_p &= \max_{\|\boldsymbol{x}\|_p\leqslant 1} \|\mathbb{A}\boldsymbol{x}\|_p && \text{(B.84)} \\ \|\mathbb{A}\mathbb{B}\|_p &\leqslant \|\mathbb{A}\|_p \|\mathbb{B}\|_p && \text{(B.85)} \end{aligned}$$

Exercice B.3.24

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{C})$. Montrer que l'on a

$$\begin{aligned} \|\mathbb{A}\|_1 &= \max_{j\in[\![1,n]\!]} \sum_{i=1}^{n} |a_{ij}|, && \text{(B.86)} \\ \|\mathbb{A}\|_2 &= \rho(\mathbb{A}\mathbb{A}^*)^{1/2}, && \text{(B.87)} \\ \|\mathbb{A}\|_\infty &= \max_{i\in[\![1,n]\!]} \sum_{j=1}^{n} |a_{ij}|. && \text{(B.88)} \end{aligned}$$

Correction Exercice B.3.24

- On démontre tout d'abord l'égalité (B.86) :

$$\|\mathbb{A}\|_1 = \max_{j\in[\![1,n]\!]} \sum_{i=1}^{n} |a_{ij}|$$

Soit $\boldsymbol{x} \in \mathbb{C}^n$ tel que $\|\boldsymbol{x}\|_1 = 1$. On a

$$\begin{aligned} \|\mathbb{A}\boldsymbol{x}\|_1 &= \sum_{i=1}^{n} |(\mathbb{A}\boldsymbol{x})_i| = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \\ &\leqslant \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}x_j| = \sum_{j=1}^{n} \left(|x_j| \sum_{i=1}^{n} |a_{ij}| \right) \\ &\leqslant \left(\max_{j\in[\![1,n]\!]} \sum_{i=1}^{n} |a_{ij}| \right) \sum_{j=1}^{n} |x_j| = \max_{j\in[\![1,n]\!]} \sum_{i=1}^{n} |a_{ij}| \text{ car } \sum_{j=1}^{n} |x_j| = 1. \end{aligned}$$

On obtient donc

$$\max_{\|\boldsymbol{x}\|_1=1} \|\mathbb{A}\boldsymbol{x}\|_1 \leqslant \max_{j\in[\![1,n]\!]} \sum_{i=1}^{n} |a_{i,j}|.$$

Pour démontrer que l'on a l'égalité

$$\max_{\|\boldsymbol{x}\|_1=1} \|\mathbb{A}\boldsymbol{x}\|_1 = \max_{j\in[\![1,n]\!]} \sum_{i=1}^{n} |a_{i,j}|.$$

il suffit alors de construire un vecteur $\boldsymbol{y} \in \mathbb{C}^n$, $\|\boldsymbol{y}\|_1 = 1$ particulier la vérifiant. Pour celà on note $k \in [\![1,n]\!]$ l'indice tel que

$$\sum_{i=1}^{n} |a_{i,k}| = \max_{j\in[\![1,n]\!]} \sum_{i=1}^{n} |a_{i,j}|.$$

On prend alors $\boldsymbol{y} = \boldsymbol{e}_k$ le $k^{\text{ème}}$ vecteur de la base canonique. Dans ce cas on a $\|\boldsymbol{y}\|_1 = 1$ et

$$\begin{aligned} \|\mathbb{A}\boldsymbol{y}\|_1 &= \|\mathbb{A}\boldsymbol{e}_k\|_1 = \|\mathbb{A}_{:,k}\|_1 \\ &= \sum_{i=1}^{n} |a_{i,k}| = \max_{j\in[\![1,n]\!]} \sum_{i=1}^{n} |a_{i,j}|. \end{aligned}$$

Ce qui démontre l'égalité (B.86).

- On démontre maintenant l'égalité (B.87) :

$$\|\mathbb{A}\|_2 = \rho(\mathbb{A}\mathbb{A}^*)^{1/2}.$$

On pose $\mathbb{B} = \mathbb{A}\mathbb{A}^*$. $\mathbb{B}$ est une matrice hermitienne : ses valeurs propres sont réelles. Soit $(\lambda, \boldsymbol{u})$ un élément propre de $\mathbb{B}$. On obtient alors

$$\begin{aligned} \langle \mathbb{B}\boldsymbol{u}, \boldsymbol{u} \rangle &= \langle \lambda \boldsymbol{u}, \boldsymbol{u} \rangle \\ &= \lambda \langle \boldsymbol{u}, \boldsymbol{u} \rangle \text{ car } \lambda \in \mathbb{R} \\ &= \lambda \|\boldsymbol{u}\|_2^2 \end{aligned}$$

De plus, on a

$$\langle \mathbb{B}\boldsymbol{u}, \boldsymbol{u} \rangle = \langle \mathbb{A}\mathbb{A}^*\boldsymbol{u}, \boldsymbol{u} \rangle = \langle \mathbb{A}^*\boldsymbol{u}, \mathbb{A}^*\boldsymbol{u} \rangle = \|\mathbb{A}^*\boldsymbol{u}\|_2^2 \geqslant 0$$

et comme $\|\boldsymbol{u}\|_2 > 0$ (c'est un vecteur propre, il est donc non nul) on en déduit

$$\lambda = \frac{\|\mathbb{A}^*\boldsymbol{u}\|_2^2}{\|\boldsymbol{u}\|_2^2} \geqslant 0.$$

La matrice $\mathbb{B}$ étant hermitienne (elle est donc normale), d'après le Théorème 3.2 page 63, il existe alors une matrice $\mathbb{U}$ unitaire et une matrice $\mathbb{D}$ diagonale telle que

$$\mathbb{B} = \mathbb{U}\mathbb{D}\mathbb{U}^*.$$

On note $(\lambda_i, \boldsymbol{e}_i)_{i \in [\![1,n]\!]}$ les éléments propres de $\mathbb{D}$. Les vecteurs $\boldsymbol{e}_i$ sont les vecteurs de la base canonique de $\mathbb{C}^n$ et $\lambda_i = d_{ii}$. Comme $\mathbb{D} = \mathbb{U}^*\mathbb{B}\mathbb{U}$, on obtient

$$\begin{aligned} \mathbb{D}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i &\iff \mathbb{U}^*\mathbb{B}\mathbb{U}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i \\ &\iff \mathbb{B}\mathbb{U}\boldsymbol{e}_i = \lambda_i \mathbb{U}\boldsymbol{e}_i. \end{aligned}$$

C'est à dire en posant $\boldsymbol{v}_i = \mathbb{U}\boldsymbol{e}_i$ (i-ème vecteur colonne de $\mathbb{U}$), les éléments propres de $\mathbb{B}$ sont les $(\lambda_i, \boldsymbol{v}_i)_{i \in [\![1,n]\!]}$. De plus comme $\mathbb{U}$ est unitaire, $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ forment une base orthonormée de $\mathbb{C}^n$.

Soit $\boldsymbol{x} \in \mathbb{C}^n$ tel que $\|\boldsymbol{x}\|_2 = 1$. le vecteur $\boldsymbol{x}$ peut s'écrire comme une combinaison linéaire de la base orthonormée $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$: $\exists (\alpha_1, \ldots, \alpha_n) \in \mathbb{C}^n$ tel que

$$\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i.$$

On peut voir que

$$\begin{aligned} \langle \boldsymbol{x}, \boldsymbol{x} \rangle = 1 &= \left\langle \sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i, \sum_{j=1}^{n} \alpha_j \boldsymbol{v}_j \right\rangle \\ &= \sum_{i=1}^{n} \overline{\alpha_i} \alpha_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle \\ &= \sum_{i=1}^{n} \overline{\alpha_i} \alpha_i \text{ car } \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle = \delta_{ij} \\ &= \sum_{i=1}^{n} |\alpha_i|^2. \end{aligned}$$

De plus on a

$$\begin{aligned}
\|\mathbb{A}\boldsymbol{x}\|_2^2 &= \langle \mathbb{A}\boldsymbol{x}, \mathbb{A}\boldsymbol{x}\rangle = \langle \mathbb{A}^*\mathbb{A}\boldsymbol{x}, \boldsymbol{x}\rangle = \langle \mathbb{B}\boldsymbol{x}, \boldsymbol{x}\rangle \\
&= \left\langle \sum_{i=1}^n \alpha_i \mathbb{B}\boldsymbol{v}_i, \sum_{j=1}^n \alpha_j \boldsymbol{v}_j \right\rangle \\
&= \left\langle \sum_{i=1}^n \alpha_i \lambda_i \boldsymbol{v}_i, \sum_{j=1}^n \alpha_j \boldsymbol{v}_j \right\rangle \\
&= \sum_{i=1}^n \overline{\alpha_i}\overline{\lambda_i} \sum_{j=1}^n \alpha_j \langle \boldsymbol{v}_i, \boldsymbol{v}_j\rangle \\
&= \sum_{i=1}^n \lambda_i |\alpha_i|^2 \quad \text{car } \lambda_i \in \mathbb{R} \text{ et } \langle \boldsymbol{v}_i, \boldsymbol{v}_j\rangle = \delta_{ij} \\
&\leqslant \max_{i\in[\![1,n]\!]} \lambda_i \sum_{i=1}^n |\alpha_i|^2 = \rho(\mathbb{A}^*\mathbb{A}) \quad \text{car } \lambda_i \geqslant 0 \text{ et } \sum_{i=1}^n |\alpha_i|^2 = 1.
\end{aligned}$$

Pour démontrer que l'on a en fait égalité il suffit de trouver un vecteur la vérifiant. Pour celà on note $k \in [\![1,n]\!]$ l'indice tel que $\lambda_k = \max_{i\in[\![1,n]\!]} \lambda_i$. En choisissant $\boldsymbol{x} = \boldsymbol{v}_k$ (qui est de norme 1) on obtient alors

$$\|\mathbb{A}\boldsymbol{v}_k\|_2^2 = \langle \mathbb{A}\boldsymbol{v}_k, \mathbb{A}\boldsymbol{v}_k\rangle = \langle \mathbb{A}^*\mathbb{A}\boldsymbol{v}_k, \boldsymbol{v}_k\rangle = \langle \lambda_k \boldsymbol{v}_k, \boldsymbol{v}_k\rangle = \lambda_k = \rho(\mathbb{A}^*\mathbb{A}).$$

Ce qui achève la démonstration de (B.87)

- Pour finir, on démontre l'égalité (B.88) :

$$\|\mathbb{A}\|_\infty = \max_{i\in[\![1,n]\!]} \sum_{j=1}^n |a_{ij}|.$$

Soit $\boldsymbol{x} \in \mathbb{C}^n$ tel que $\|\boldsymbol{x}\|_\infty = 1$. On a

$$\begin{aligned}
\|\mathbb{A}\boldsymbol{x}\|_\infty &= \max_{i\in[\![1,n]\!]} |(\mathbb{A}\boldsymbol{x})_i| = \max_{i\in[\![1,n]\!]} \left| \sum_{j=1}^n a_{i,j} x_j \right| \\
&\leqslant \max_{i\in[\![1,n]\!]} \sum_{j=1}^n |a_{i,j}||x_j| \\
&\leqslant \max_{i\in[\![1,n]\!]} \sum_{j=1}^n |a_{i,j}| \quad \text{car } |x_j| \leqslant \max_{i\in[\![1,n]\!]} |x_i| = \|\boldsymbol{x}\|_\infty = 1.
\end{aligned}$$

Pour démontrer l'égalité, il suffit de construire un vecteur $\boldsymbol{y}$ de norme 1 la vérifiant. Pour celà on note $k \in [\![1,n]\!]$ l'indice tel que

$$\max_{i\in[\![1,n]\!]} \sum_{j=1}^n |a_{ij}| = \sum_{j=1}^n |a_{k,j}|.$$

On veut alors construire un vecteur $\boldsymbol{y}$ tel que $\|\boldsymbol{y}\|_\infty = 1$ et vérifiant

$$\max_{i\in[\![1,n]\!]} \left| \sum_{j=1}^n a_{i,j} y_j \right| = \sum_{j=1}^n |a_{k,j}|.$$

Pour celà on pose, pour tout $j \in [\![1,n]\!]$

$$y_j = \begin{cases} \frac{|a_{k,j}|}{a_{k,j}} & \text{si } a_{kj} \neq 0 \\ 1 & \text{si } a_{k,j} = 0 \end{cases}.$$

On a alors $\|\boldsymbol{y}\|_\infty = 1$,

$$\sum_{j=1}^n a_{k,j} y_j = \sum_{j=1}^n |a_{k,j}|$$

On en déduit

$$\|\mathbb{A}\boldsymbol{y}\|_\infty = \max_{i\in[\![1,n]\!]} \left|\sum_{j=1}^{n} a_{i,j}y_j\right| = \sum_{j=1}^{n} |a_{k,j}| = \max_{i\in[\![1,n]\!]} \sum_{j=1}^{n} |a_{i,j}|$$

ce qui achève la démonstration.

◇

B.4 Listings

B.4.1 Codes sur la méthode de dichotomie/bissection

Transcription de l'Algorithme 2.1(page 18)

```
function x=dichotomie1(f,a,b,epsilon)
  kmin=floor( ...
       log((b-a)/epsilon)/log(2) );
  X=zeros(kmin+1,1);
  A=zeros(kmin+1,1);B=zeros(kmin+1,1);
  A(1)=a;B(1)=b;X(1)=(a+b)/2;
  for k=1:kmin
    if f(X(k))==0
      A(k+1)=X(k);B(k+1)=X(k);
    elseif f(B(k))*f(X(k))<0
      A(k+1)=X(k);B(k+1)=B(k);
    else
      A(k+1)=A(k);B(k+1)=X(k);
    end
    X(k+1)=(A(k+1)+B(k+1))/2;
  end
  x=X(kmin+1);
end
```

Listing B.1: fonction `dichotomie1` (Matlab)

```
double dichotomie1(double (*f)(double),
       double a, double b,double eps){
  double *A,*B,*X,x;
  int kmin,k;
  size_t Size;
  assert(f(a)*f(b)<0);
  kmin=(int)floor(log((b-a)/eps)/log(2.));
  Size=(kmin+1)*sizeof(double);
  assert( X=(double*)malloc(Size) );
  assert( A=(double*)malloc(Size) );
  assert( B=(double*)malloc(Size) );
  // ou assert( A && B && X );
  A[0]=a;B[0]=b;X[0]=(a+b)/2.;
  for (k=0;k<kmin;k++){
    if (f(X[k])==0){
      A[k+1]=X[k];B[k+1]=X[k];
    }else if ( f(B[k])*f(X[k])<0 ){
      A[k+1]=X[k];B[k+1]=B[k];
    }else{
      A[k+1]=A[k];B[k+1]=X[k];
    }
    X[k+1]=(A[k+1]+B[k+1])/2;
  }
  x=X[kmin];
  free(X);free(A);free(B);
  return x;
}
```

Listing B.2: fonction `dichotomie1` (C)

```
clear all
close all
f=@(x) (x+2)*(x+2)*(x-pi);
x=dichotomie1(f,-1,2*pi,1e-8);
fprintf('x=%.16f\n',x)
x=dichotomie1(@cos,2,pi,1e-8);
fprintf('x=%.16f\n',x)
```

Listing B.3: script `dichotomie1` (Matlab)

```
#include <stdio.h>
#include <stdlib.h>
#include <math.h>
#include <assert.h>

double dichotomie1(
  double (*f)(double),
  double a, double b, double eps
);
double g1(double x){
  return (x+2)*(x+2)*(x-M_PI);
}

int main(){
  double x;
  x=dichotomie1(g1,-1,2*M_PI,1e-8);
  printf("x=%.16lf, error=%.6e\n",
         x,fabs(x-M_PI));
  x=dichotomie1(cos,-1,M_PI,1e-8);
  printf("x=%.16lf, error=%.6e\n",
         x,fabs(x-M_PI_2));
  return 1;
}
// Definition de dichotomie1 ensuite ...
```

Listing B.4: main `dichotomie1` (C)

Transcription de l'Algorithme 2.5 (page 20)

```
function x=dichotomie5(f,a,b)
  assert(f(a)*f(b)<0, ...
         'test f(a)*f(b)<0 failed');
  A=a;B=b;x=(A+B)/2;xp=A;
  while x~=xp
    if f(B)*f(x)<0
      A=x;
    else
      B=x;
    end
    xp=x;
    x=(A+B)/2;
  end
end
```

Listing B.5: fonction `dichotomie5` (Matlab)

```
double dichotomie5(double (*f)(double),
                   double a, double b){
  double A,B,x,xp;
  assert(f(a)*f(b)<0);
  A=a;B=b;x=(A+B)/2.;xp=A;
  while (x!=xp){
    if (f(B)*f(x)<0)
      A=x;
    else
      B=x;
    xp=x;
    x=(A+B)/2;
  }
  return x;
}
```

Listing B.6: fonction `dichotomie5` (C)

```
clear all
close all
f=@(x) (x+2)*(x+2)*(x-pi);
x=dichotomie5(f,-1,2*pi);
fprintf('x=%.16f\n',x)
x=dichotomie5(@cos,2,pi);
fprintf('x=%.16f\n',x)
```

Listing B.7: script `dichotomie5` (Matlab)

```
#include <stdio.h>
#include <math.h>
#include <assert.h>

double dichotomie5(double (*f)(double),
                   double a, double b);
double g1(double x){
  return (x+2)*(x+2)*(x-M_PI);
}

int main(){
  double x;
  x=dichotomie5(g1,-1,2*M_PI);
  printf("x=%.16lf\n",x);
  x=dichotomie5(cos,-1,M_PI);
  printf("x=%.16lf\n",x);
```

Listing B.8: main `dichotomie5` (C)

Liste des algorithmes

Index

Bibliography

[1] P.G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. DUNOD, 2006.

[2] M. Crouzeix and A.L. Mignot. *Analyse numérique des équations différentielles*. Mathématiques appliquées pour la maîtrise. Masson, 1992.

[3] J.-P. Demailly. *Analyse numérique et équations différentielles*. Grenoble Sciences. EDP Sciences, 2006.

[4] J.P. Demailly. *Analyse Numérique et Equations Différentielles*. PUG, 1994.

[5] W. Gander, M.J. Gander, and F. Kwok. *Scientific computing : an introduction using Maple and MATLAB*. Springer, Cham, 2014.

[6] T. Huckle. Collection of software bugs http://wwwzenger.informatik.tu-muenchen.de/persons/huckle/bugse.html.

[7] P. Lascaux and R. Théodor. *Analyse numérique matricielle appliquée à l'art de l'ingénieur*. Number vol. 1 et 2 in Analyse numérique matricielle appliquée à l'art de l'ingénieur. Dunod, 2004.