Dauphine | PSL
UNIVERSITÉ PARIS

Notes de cours de méthodes numériques

Guillaume Legendre

(version du 19 janvier 2023)

Table des matières

Chapitre 1

Résolution numérique des équations non linéaires dans $\mathbb{R}$

Nous nous intéressons dans ce chapitre à l'approximation de zéros (ou de racines, dans le cas d'un polynôme [1]) d'une fonction réelle d'une variable réelle, c'est-à-dire, étant donné un intervalle $I \subseteq \mathbb{R}$ et une application f définie sur I et à valeurs réelles, la résolution approchée du problème : *trouver un réel ξ tel que*

$$f(\xi) = 0.$$

Ce problème intervient notamment dans l'étude générale d'une fonction d'une variable réelle, qu'elle soit motivée ou non par des applications, pour laquelle des solutions exactes de ce type d'équation ne sont pas explicitement connues [2].

Toutes les méthodes que nous allons présenter sont itératives et consistent donc en la construction d'une suite de réels $(x^{(k)})_{k\in\mathbb{N}}$ qui, on espère, sera telle que

$$\lim_{k\to+\infty} x^{(k)} = \xi.$$

En effet, à la différence du cas des systèmes linéaires étudié dans le chapitre 5, la convergence de ces méthodes itératives dépend en général du choix de l'approximation initiale $x^{(0)}$. On verra ainsi qu'on ne sait souvent qu'établir des résultats de *convergence locale*, valables lorsque $x^{(0)}$ appartient à un certain voisinage du zéro ξ.

Après avoir caractérisé la convergence de suites engendrées par les méthodes itératives présentées dans ce chapitre, en introduisant notamment la notion d'ordre de convergence, nous introduisons plusieurs méthodes parmi les plus connues et les plus utilisées : tout d'abord les méthodes de dichotomie et de la fausse position qui sont toutes deux des méthodes dites *d'encadrement* (*bracketing methods* en anglais), puis la méthode de Newton–Raphson, qui fait partie des *méthodes de point fixe* (*fixed-point methods* en anglais), et enfin la méthode de la sécante. Dans chaque cas, un ou plusieurs résultats de convergence *ad hoc* sont énoncés.

1.1 Ordre de convergence d'une méthode itérative

Afin de pouvoir évaluer à quelle « vitesse » la suite construite par une méthode itérative converge vers sa limite (ce sera souvent l'un des critères discriminants pour le choix d'une méthode), il nous faut introduire quelques notions.

Définition 1 (ordre de convergence d'une suite) *Soit une suite $(x^{(k)})_{k\in\mathbb{N}}$ de réels convergeant vers une limite ξ. On dit que cette suite est **convergente d'ordre** $r \geq 1$, s'il existe deux constantes strictement positives C_1 et C_2, avec $C_1 \leq C_2$, et un entier naturel k_0 tels que*

$$\forall k \in \mathbb{N},\ k \geq k_0 \Rightarrow C_1 \leq \frac{\left|x^{(k+1)} - \xi\right|}{\left|x^{(k)} - \xi\right|^r} \leq C_2. \tag{1.1}$$

1. On commettra parfois dans la suite un abus de langage en appelant *« polynôme »* toute fonction polynomiale, c'est-à-dire toute application associée à un polynôme à coefficients dans un anneau commutatif (le corps $\mathbb{R}$ dans notre cas). Dans ce cas particulier, tout zéro de la fonction est une *racine* du polynôme qui lui est sous-jacent.

2. Même dans le cas d'une équation algébrique, on rappelle qu'il n'existe pas de méthode de résolution générale à partir du degré cinq.

Par extension, une méthode itérative produisant une suite convergente vérifiant les relations (1.1) sera également dite *d'ordre* r. On notera que, dans plusieurs ouvrages, on trouve l'ordre d'une suite simplement défini par le fait qu'il existe une constante $C \geq 0$ et un entier k_0 tels que, pour tout entier k tel que $k \geq k_0 \geq 0$, $\left|x^{(k+1)} - \xi\right| \leq C \left|x^{(k)} - \xi\right|^r$. Il faut cependant observer [3] que cette définition n'assure pas l'unicité de r, l'ordre de convergence pouvant éventuellement être plus grand que r. On préférera donc dire dans ce cas que la suite est d'ordre r *au moins*. On remarquera aussi que, si r est égal à 1, on a nécessairement $C_2 < 1$ dans (1.1), faute de quoi la suite ne pourrait converger.

La définition 1 est très générale et n'exige pas que la suite $\left(\frac{\left|x^{(k+1)}-\xi\right|}{\left|x^{(k)}-\xi\right|^r}\right)_{k\in\mathbb{N}}$ admette une limite. Lorsque c'est le cas, on a coutume de se servir de la définition suivante.

Définition 2 *Soit une suite $(x^{(k)})_{k\in\mathbb{N}}$ de réels convergeant vers une limite ξ. On dit que cette suite est* ***convergente d'ordre*** *r, avec $r > 1$, vers ξ s'il existe un réel $\mu > 0$, appelé* ***constante asymptotique d'erreur****, tel que*

$$\lim_{k\to+\infty} \frac{\left|x^{(k+1)} - \xi\right|}{\left|x^{(k)} - \xi\right|^r} = \mu. \tag{1.2}$$

Dans le cas particulier où $r = 1$, on dit que la suite ***converge linéairement*** *si*

$$\lim_{k\to+\infty} \frac{\left|x^{(k+1)} - \xi\right|}{\left|x^{(k)} - \xi\right|} = \mu, \text{ avec } \mu \in]0,1[,$$

et ***super-linéairement*** *(resp.* ***sous-linéairement****) si l'égalité ci-dessus est vérifiée avec $\mu = 0$ (resp. $\mu = 1$).*

Ajoutons que la convergence d'ordre deux est dite *quadratique*, celle d'ordre trois *cubique*. On parle parfois de convergence *q-linéaire*, *q-quadratique*, etc. (*q-linear convergence*, *q-quadratic convergence*, etc. en anglais), la lettre q du préfixe signifiant « quotient ».

Si la dernière caractérisation est particulièrement adaptée à l'étude pratique de la plupart des méthodes itératives que nous allons présenter dans ce chapitre, elle a comme inconvénient de ne pouvoir fournir l'ordre d'une suite dont la « vitesse de convergence » est variable, ce qui se traduit par le fait que la limite (1.2) n'existe pas. On a alors recours à une définition « étendue ».

Définition 3 *On dit qu'une suite $(x^{(k)})_{k\in\mathbb{N}}$ de réels* ***converge avec un ordre au moins égal à*** *r, avec $r \geq 1$, vers une limite ξ s'il existe une suite positive $(\varepsilon^{(k)})_{k\in\mathbb{N}}$ tendant vers 0 vérifiant*

$$\forall k \in \mathbb{N},\ \left|x^{(k)} - \xi\right| \leq \varepsilon^{(k)}, \tag{1.3}$$

et un réel $\nu > 0$ ($0 < \nu < 1$ si $r = 1$) tel que

$$\lim_{k\to+\infty} \frac{\varepsilon^{(k+1)}}{{\varepsilon^{(k)}}^r} = \nu.$$

On remarquera l'ajout du qualificatif « au moins » dans la définition 3, qui provient du fait que l'on a dû procéder à une majoration par une suite convergeant vers zéro avec un ordre r au sens de la définition 2. Bien évidemment, on retrouve la définition 2 si l'on a égalité dans (1.3), mais ceci est souvent impossible à établir en pratique. En anglais, on qualifie parfois cette notion de convergence en utilisant le préfixe *r* (*r-linear convergence*, *r-quadratic convergence*, etc.) représentant le mot *"root"*.

Indiquons que les notions d'ordre et de constante asymptotique d'erreur ne sont pas purement théoriques, mais en relation avec le nombre de chiffres exacts obtenus dans l'approximation de ξ. Posons en effet $\delta^{(k)} = -\log_{10}(\left|x^{(k)} - \xi\right|)$; $\delta^{(k)}$ est alors le nombre de chiffres significatifs décimaux exacts de $x^{(k)}$. Pour k suffisamment grand, on a

$$\delta^{(k+1)} \approx r\,\delta^{(k)} - \log_{10}(\mu).$$

On voit donc que si r est égal à un, on ajoute environ $-\log_{10}(\mu)$ chiffres significatifs à chaque itération. Par exemple, si $\mu = 0,999$ alors $-\log_{10}(\mu) \approx 4,34\,10^{-4}$ et il faudra près de 2500 itérations pour gagner une seule décimale. Par contre, si r est strictement plus grand que un, on multiplie environ par r le nombre de chiffres significatifs à chaque itération. Ceci montre clairement l'intérêt des méthodes d'ordre plus grand que un.

3. On pourra considérer l'exemple de la suite positive définie par, $\forall k \in \mathbb{N}$, $x^{(k)} = \alpha^{\beta^k}$ avec $0 < \alpha < 1$ et $\beta > 1$. Cette suite est d'ordre β d'après la définition 1, alors que, $\forall k \in \mathbb{N}$, $x^{(k+1)} = {x^{(k)}}^{\beta} \leq x^{(k)^\gamma}$ pour $1 < \gamma < \beta$.

1.2 Méthodes d'encadrement

Cette première classe de méthodes repose sur la propriété fondamentale suivante, relative à l'existence d'un zéro pour une application d'une variable réelle à valeurs réelles.

Théorème 4 (existence d'un zéro d'une fonction à valeurs réelles continue) *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et f une application continue de $[a,b]$ dans $\mathbb{R}$ vérifiant $f(a)f(b)<0$. Alors il existe un réel ξ appartenant à l'intervalle $]a,b[$ tel que $f(\xi)=0$.*

DÉMONSTRATION. Si $f(a)<0$, on a $0\in]f(a),f(b)[$, sinon $f(a)>0$ et alors $0\in]f(b),f(a)[$. Dans ces deux cas, le résultat est une conséquence du théorème des valeurs intermédiaires. □

1.2.1 Méthode de dichotomie

La *méthode de dichotomie*, ou *méthode de la bissection* (*bisection method* en anglais), suppose que la fonction f est continue sur un intervalle $[a,b]$, vérifie $f(a)f(b)<0$ et admet donc (au moins) un zéro ξ dans $]a,b[$.

Son principe est le suivant. On pose $a^{(0)}=a$, $b^{(0)}=b$, on note $x^{(0)}=\frac{1}{2}(a^{(0)}+b^{(0)})$ le milieu de l'intervalle de départ et on évalue la fonction f en ce point. Si $f(x^{(0)})=0$, le point $x^{(0)}$ est le zéro de f et le problème est résolu. Sinon, si $f(a^{(0)})f(x^{(0)})<0$, alors le zéro ξ est contenu dans l'intervalle $]a^{(0)},x^{(0)}[$, alors qu'il appartient à $]x^{(0)},b^{(0)}[$ si $f(x^{(0)})f(b^{(0)})<0$. On réitère ensuite ce processus sur l'intervalle $[a^{(1)},b^{(1)}]$, avec $a^{(1)}=a^{(0)}$ et $b^{(1)}=x^{(0)}$ dans le premier cas, ou $a^{(1)}=x^{(0)}$ et $b^{(1)}=b^{(0)}$ dans le second, et ainsi de suite...

De cette manière, on construit de manière récurrente trois suites $(a^{(k)})_{k\in\mathbb{N}}$, $(b^{(k)})_{k\in\mathbb{N}}$ et $(x^{(k)})_{k\in\mathbb{N}}$ telles que $a^{(0)}=a$, $b^{(0)}=b$ et vérifiant, pour tout entier naturel k,

- $x^{(k)}=\dfrac{a^{(k)}+b^{(k)}}{2}$,
- $a^{(k+1)}=a^{(k)}$ et $b^{(k+1)}=x^{(k)}$ si $f(a^{(k)})f(x^{(k)})<0$,
- $a^{(k+1)}=x^{(k)}$ et $b^{(k+1)}=b^{(k)}$ si $f(x^{(k)})f(b^{(k)})<0$.

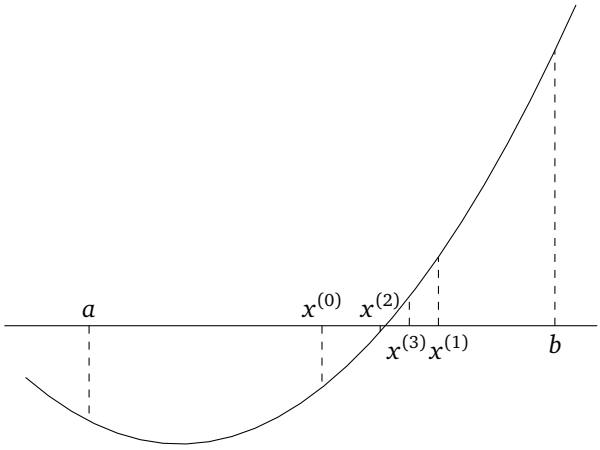

FIGURE 1.1: Construction des premiers itérés de la méthode de dichotomie.

La figure 1.1 illustre la construction des approximations du zéro produites par cette méthode.

Concernant la convergence de la méthode de dichotomie, on a le résultat suivant, dont la preuve est laissée en exercice.

Proposition 5 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et f une fonction réelle continue sur $[a,b]$, vérifiant $f(a)f(b)<0$ et possédant un unique zéro ξ dans $]a,b[$. Alors, la suite $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ construite par la méthode de dichotomie converge vers ξ et on a l'estimation*

$$\forall k\in\mathbb{N},\ \left|x^{(k)}-\xi\right|\leq\frac{b-a}{2^{k+1}}. \tag{1.4}$$

Il ressort de cette proposition que la méthode de dichotomie converge de manière certaine : c'est une méthode *globalement convergente*. L'estimation d'erreur (1.4) fournit par ailleurs directement un critère d'arrêt pour la méthode, puisque, à précision ε donnée, cette dernière permet d'approcher ξ en un nombre prévisible d'itérations. On voit en effet que, pour avoir $|x^{(k)} - \xi| \leq \varepsilon$, il faut que

$$\frac{b-a}{2^{k+1}} \leq \varepsilon \Leftrightarrow k \geq \frac{\ln\left(\frac{b-a}{\varepsilon}\right)}{\ln(2)} - 1. \tag{1.5}$$

Ainsi, pour améliorer la précision de l'approximation du zéro d'un ordre de grandeur, c'est-à-dire trouver $k > j$ tel que $\left|x^{(k)} - \xi\right| = \frac{1}{10}\left|x^{(j)} - \xi\right|$, on doit effectuer $k - j = \frac{\ln(10)}{\ln(2)} \simeq 3,32$ itérations.

On peut par ailleurs remarquer que, sous la seule hypothèse que $f(a)f(b) < 0$, le théorème 4 garantit l'existence d'*au moins* un zéro de la fonction f dans l'intervalle $]a, b[$ et il se peut donc que ce dernier en contienne plusieurs. Dans ce cas, la méthode converge vers l'un d'entre eux.

Comme on le constate sur la figure 1.2, la méthode de dichotomie ne garantit pas une réduction monotone de l'erreur absolue d'une itération à l'autre. Ce n'est donc pas une méthode d'ordre un au sens de la définition 1, mais sa convergence est néanmoins linéaire au sens de la définition 3.

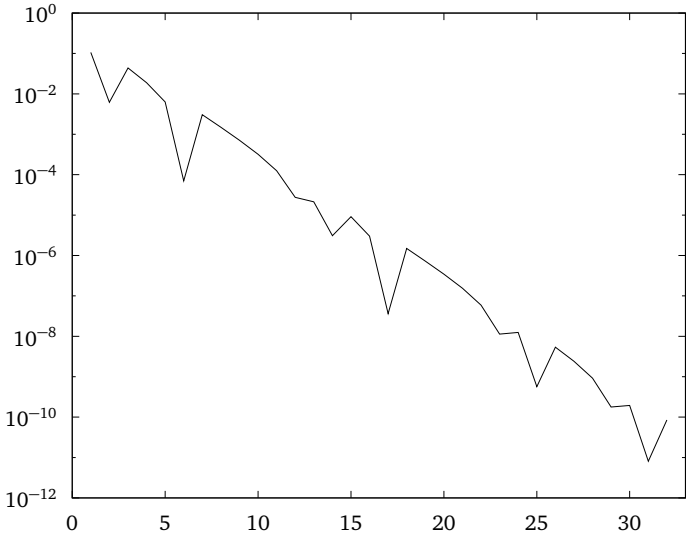

FIGURE 1.2: Historique de la convergence, c'est-à-dire le tracé de l'erreur absolue $\left|x^{(k)} - \xi\right|$ en fonction k, de la méthode de dichotomie pour l'approximation de la racine $\xi = 0,9061798459...$ du polynôme de Legendre de degré 5, $P_5(x) = \frac{1}{8}\,x\,(63x^4 - 70\,x^2 + 15)$, dont les racines se situent dans l'intervalle $]-1, 1[$. On a choisi les bornes $a = 0,6$ et $b = 1$ pour l'intervalle d'encadrement initial et une précision de 10^{-10} pour le test d'arrêt, qui est atteinte après 31 itérations (à comparer à la valeur $30,89735...$ fournie par l'estimation (1.5)). On observe que l'erreur a un comportement oscillant, mais diminue en moyenne de manière linéaire.

On gardera donc à l'esprit que la méthode de dichotomie est une méthode robuste. Si sa convergence est lente, on peut l'utiliser pour obtenir une approximation grossière (mais raisonnable) du zéro recherché servant d'initialisation à une méthode d'ordre plus élevé dont la convergence n'est que *locale*, comme la méthode de Newton–Raphson (voir la section 1.3.3). On peut voir cette approche comme une stratégie de « globalisation » de méthodes localement convergentes.

1.2.2 Méthode de la fausse position

La *méthode de la fausse position* (*false-position method* en anglais), encore appelée *méthode regula falsi*, est une méthode d'encadrement combinant les possibilités de la méthode de dichotomie avec celles de la méthode de la sécante, qui sera introduite en fin de chapitre. L'idée est d'utiliser l'information fournie par les valeurs de la fonction f aux extrémités de l'intervalle d'encadrement pour améliorer la vitesse de convergence de la méthode de dichotomie (cette dernière ne tenant compte que du signe de la fonction).

Comme précédemment, cette méthode suppose connus deux points a et b vérifiant $f(a)f(b) < 0$ et servant d'initialisation à la suite d'intervalles $[a^{(k)}, b^{(k)}]$, $k \geq 0$, contenant un zéro de la fonction f. Le procédé de construction des intervalles emboîtés est alors le même pour la méthode de dichotomie, à l'exception du choix de $x^{(k)}$, qui est à présent donné par l'abscisse du point d'intersection de la droite passant par les points $(a^{(k)}, f(a^{(k)}))$ et $(b^{(k)}, f(b^{(k)}))$ avec l'axe des abscisses, c'est-à-dire

$$x^{(k)} = a^{(k)} - \frac{a^{(k)} - b^{(k)}}{f(a^{(k)}) - f(b^{(k)})} f(a^{(k)}) = b^{(k)} - \frac{b^{(k)} - a^{(k)}}{f(b^{(k)}) - f(a^{(k)})} f(b^{(k)}) = \frac{f(a^{(k)})b^{(k)} - f(b^{(k)})a^{(k)}}{f(a^{(k)}) - f(b^{(k)})}. \tag{1.6}$$

De cette façon, le zéro est obtenu après une seule itération si f est une fonction affine, contre *a priori* une infinité pour la méthode de dichotomie.

On a représenté sur la figure 1.3 la construction des premières approximations $x^{(k)}$ ainsi trouvées.

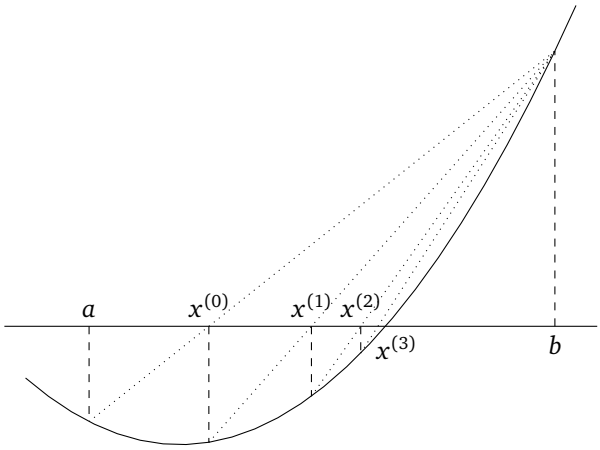

FIGURE 1.3: Construction des premiers itérés de la méthode de la fausse position.

Indiquons que si la mesure de l'intervalle d'encadrement $[a^{(k)}, b^{(k)}]$ ainsi obtenu décroît bien lorsque k tend vers l'infini, elle ne tend pas nécessairement vers zéro [4], comme c'est le cas pour la méthode de dichotomie. En effet, pour une fonction convexe ou concave dans un voisinage du zéro recherché, il apparaît que la méthode conduit inévitablement, à partir d'un certain rang, à l'une des configurations présentées sur la figure 1.4, pour chacune desquelles l'une des bornes de l'intervalle d'encadrement n'est plus jamais modifiée tandis que l'autre converge de manière monotone vers le zéro. On a alors affaire à une *méthode de point fixe* (voir la section 1.3, en comparant en particulier les relations de récurrence (1.7) et (1.8)).

L'analyse de la méthode de la fausse position est bien moins triviale que celle de la méthode de dichotomie. On peut cependant établir le résultat de convergence *linéaire* suivant moyennant quelques hypothèses sur la fonction f.

Théorème 6 *Soit $[a, b]$ un intervalle non vide de $\mathbb{R}$ et f une fonction réelle continue sur $[a, b]$, vérifiant $f(a)f(b) < 0$ et possédant un unique zéro ξ dans $]a, b[$. Supposons de plus que f est continûment dérivable sur $]a, b[$ et convexe ou concave dans un voisinage de ξ. Alors, la suite $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ construite par la méthode de la fausse position converge au moins linéairement vers ξ.*

DÉMONSTRATION. Compte tenu des hypothèses, l'une des configurations illustrées à la figure 1.4 est obligatoirement atteinte par la méthode de la fausse position à partir d'un certain rang et l'on peut se ramener sans perte de généralité au cas où l'une des bornes de l'intervalle de départ reste fixe tout au long du processus itératif.

On peut ainsi considérer le cas d'une fonction f convexe sur l'intervalle $[a, b]$ et telle que $f(a) < 0$ et $f(b) > 0$ (ce qui correspond à la première configuration décrite sur la figure 1.4). Dans ces conditions, on montre, en utilisant (1.6) et la convexité de f, que, $\forall k \in \mathbb{N}$, $f(x^{(k)}) \leq 0$. Par définition de la méthode, on a, $\forall k \in \mathbb{N}$, $a^{(k+1)} = x^{(k)}$ et $b^{(k+1)} = b$ si $f(x^{(k)}) < 0$, le point $x^{(k+1)}$ étant alors donné par la formule

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \frac{b - x^{(k)}}{f(b) - f(x^{(k)})} f(x^{(k)}), \tag{1.7}$$

4. Pour cette raison, le critère d'arrêt des itérations de la méthode doit être basé soit sur la longueur à l'étape k du plus petit des intervalles $[a^{(k)}, x^{(k)}]$ et $[x^{(k)}, b^{(k)}]$, $k \geq 0$, soit sur la valeur du résidu $f(x^{(k)})$.

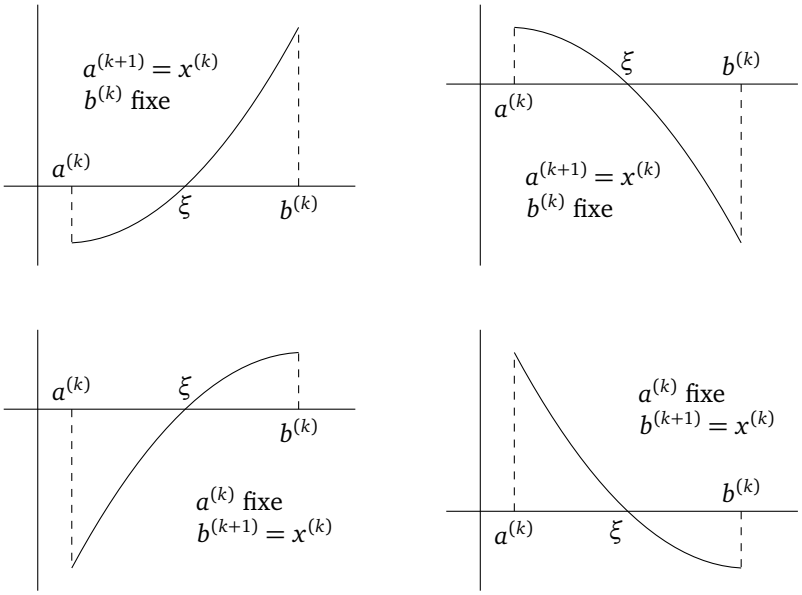

FIGURE 1.4: Différentes configurations atteintes par la méthode de la fausse position à partir d'un certain rang pour une fonction f supposée convexe ou concave dans un voisinage du zéro ξ.

ou bien $x^{(k+1)} = x^{(k)} = \xi$ si $f(x^{(k)}) = 0$, ce dernier cas mettant fin aux itérations.

Supposons à présent que, $\forall k \in \mathbb{N}$, $f(x^{(k)}) \neq 0$. Il découle de la relation (1.7) que la suite $(x^{(k)})_{k\in\mathbb{N}}$ est croissante et elle est par ailleurs majorée par b ; elle converge donc vers une limite ℓ, qui vérifie

$$(b-\ell)f(\ell) = 0.$$

Puisque, $\forall k \in \mathbb{N}$, $x^{(k)} < \xi$ on a $\ell \leq \xi < b$ et, par voie de conséquence, $f(\ell) = 0$, d'où $\ell = \xi$, par unicité du zéro ξ.

Il reste à prouver que la convergence de la méthode est au moins linéaire. En se servant une nouvelle fois de (1.7) et en faisant tendre k vers l'infini, on trouve que

$$\lim_{k\to+\infty} \frac{x^{(k+1)}-\xi}{x^{(k)}-\xi} = 1 - \frac{b-\xi}{f(b)-f(\xi)} f'(\xi).$$

La fonction f étant supposée convexe sur $[a,b]$, on a, $\forall x \in [a,b]$, $f(x) \geq f(\xi) + (x-\xi)f'(\xi)$; en choisissant $x = a$ et $x = b$ dans cette dernière inégalité, on obtient respectivement que $f'(\xi) > 0$ et $f'(\xi) \leq \frac{f(b)-f(\xi)}{b-\xi}$, d'où la conclusion.

La même technique de démonstration s'adapte pour traiter les trois cas possibles restants, ce qui achève la preuve. □

Dans de nombreuses situations, comme pour la résolution approchée de l'équation de Kepler dans le cas d'une orbite elliptique présentée sur la figure 1.5, la méthode de la fausse position converge plus rapidement que la méthode de dichotomie. Ceci n'est cependant pas une règle générale et l'on peut construire des exemples pour lesquels il en va tout autrement (voir la figure 1.6).

1.3 Méthodes de point fixe

Les méthodes d'approximation de zéros introduites dans la suite de ce chapitre sont plus générales au sens où elles se passent de l'hypothèse de changement de signe de f en ξ et ne consistent pas en la construction d'une suite d'intervalles contenant le zéro de la fonction ; bien qu'étant aussi des méthodes itératives, ce ne sont pas des méthodes d'encadrement. Rien ne garantit d'ailleurs qu'une suite $(x^{(k)})_{k\in\mathbb{N}}$ produite par l'un des algorithmes présentés prendra ses valeurs dans un intervalle fixé *a priori*.

Au sein de cette catégorie de méthodes itératives, les méthodes de point fixe sont basées sur le fait que tout problème de recherche de zéros d'une fonction peut se ramener à un problème de recherche de points fixes d'une autre fonction.

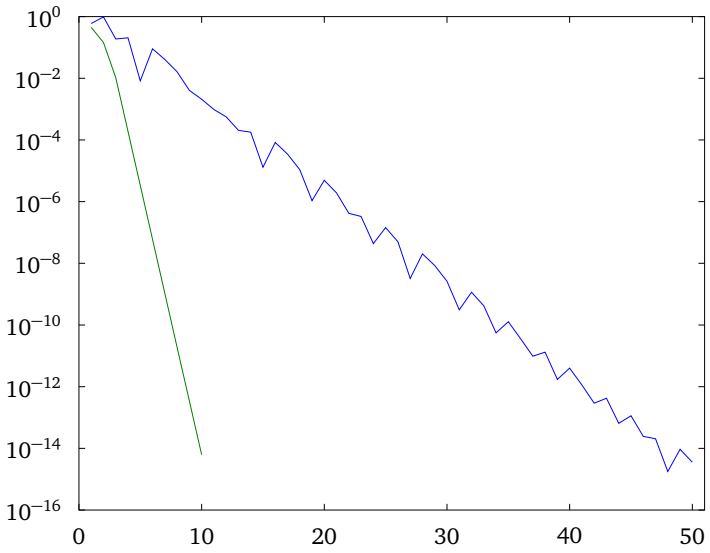

FIGURE 1.5: Tracés, en fonction du nombre d'itérations, des erreurs absolues de la méthode de dichotomie (en bleu) et de la méthode de la fausse position (en vert) utilisées pour résoudre de manière approchée l'équation de Kepler $x - e\sin(x) = M$, avec $e = 0,8$ et $M = \frac{4\pi}{3}$, de solution $x = 3,7388733587...$, à partir de l'intervalle d'encadrement initial $[0, 2\pi]$.

1.3.1 Principe

La famille de méthodes que nous allons maintenant étudier utilise le fait que le problème $f(x) = 0$ peut toujours ramener au problème équivalent $g(x) - x = 0$, pour lequel on a le résultat suivant.

Théorème 7 (« théorème du point fixe de Brouwer ») *Soit $[a, b]$ un intervalle non vide de $\mathbb{R}$ et g une application continue de $[a, b]$ dans lui-même. Alors, il existe un point ξ de $[a, b]$, appelé* ***point fixe de la fonction*** *g, vérifiant $g(\xi) = \xi$.*

DÉMONSTRATION. Posons $f(x) = g(x) - x$. On a alors $f(a) = g(a) - a \geq 0$ et $f(b) = g(b) - b \leq 0$, puisque $g(x)$ appartient à $[a, b]$ pour tout x appartenant à $[a, b]$. Par conséquent, la fonction f, continue sur $[a, b]$, est telle que $f(a)f(b) \leq 0$. Ceci assure l'existence d'un point ξ dans $[a, b]$ tel que $0 = f(\xi) = g(\xi) - \xi$. □

Bien entendu, toute équation de la forme $f(x) = 0$ peut s'écrire sous la forme $x = g(x)$ en posant $g(x) = x + f(x)$, mais ceci ne garantit en rien que la fonction auxiliaire g ainsi définie satisfait les hypothèses du théorème 7. Il existe cependant de nombreuses façons de construire g à partir de f et il suffit donc de trouver une transformation adaptée.

Nous venons de montrer que, sous certaines conditions, approcher les zéros d'une fonction f revient à approcher les points fixes d'une fonction g, sans que l'on sache pour autant traiter ce nouveau problème. Une méthode courante pour la détermination de point fixe se résume à la construction d'une suite $(x^{(k)})_{k\in\mathbb{N}}$ par le procédé itératif suivant : étant donnée une valeur initiale $x^{(0)}$ (appartenant à $[a, b]$), on pose

$$\forall k \in \mathbb{N},\ x^{(k+1)} = g(x^{(k)}). \tag{1.8}$$

On dit que la relation (1.8) est une *itération de point fixe* (*fixed-point iteration* en anglais). La méthode d'approximation résultante est appelée *méthode de point fixe* ou bien encore *méthode des approximations successives*. Si la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par (1.8) converge, cela ne peut être que vers un point fixe de g. En effet, en posant $\lim_{k\to+\infty} x^{(k)} = \xi$, nous avons

$$\xi = \lim_{k\to+\infty} x^{(k+1)} = \lim_{k\to+\infty} g(x^{(k)}) = g\left(\lim_{k\to+\infty} x^{(k)}\right) = g(\xi),$$

la deuxième égalité provenant de la définition (1.8) de la suite récurrente et la troisième étant une conséquence de la continuité de g.

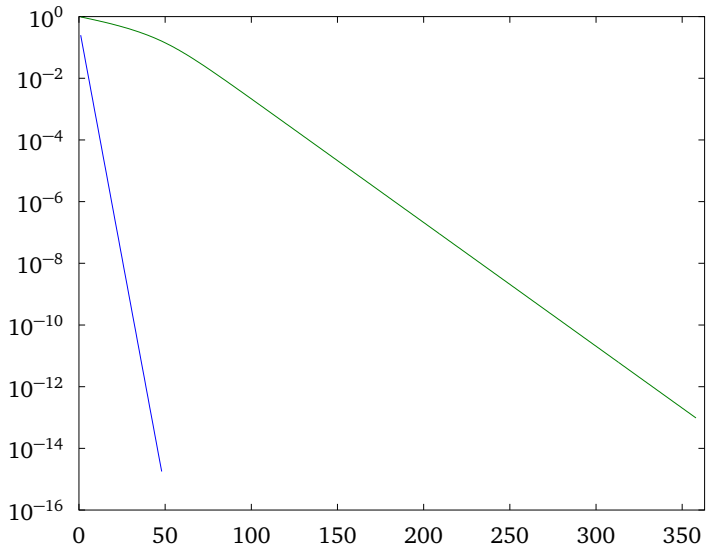

FIGURE 1.6: Tracés, en fonction du nombre d'itérations, des erreurs absolues de la méthode de dichotomie (en bleu) et de la méthode de la fausse position (en vert) utilisées pour la résolution approchée de l'équation $x^{10}-1=0$ à partir de l'intervalle d'encadrement initial $\left[0, \frac{3}{2}\right]$. Malgré l'accélération observée de la convergence de la méthode de la fausse position durant les premières itérations, la vitesse de cette dernière reste largement inférieure à celle de la méthode de dichotomie.

1.3.2 Quelques résultats de convergence

Le choix de la fonction g pour mettre en œuvre cette méthode n'étant pas unique, celui-ci est alors motivé par les exigences du théorème 9, qui donne des conditions *suffisantes* sur g pour avoir convergence de la méthode de point fixe définie par (1.8). Avant de l'énoncer, rappelons tout d'abord la notion d'application *contractante*.

Définition 8 (application contractante) *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et g une application de $[a,b]$ dans $\mathbb{R}$. On dit que g est une application* ***contractante*** *si et seulement si il existe une constante K telle que $0<K<1$ vérifiant*

$$\forall x \in [a,b],\ \forall y \in [a,b],\ |g(x)-g(y)| \leq K\,|x-y|\,. \tag{1.9}$$

On notera que la constante de Lipschitz de l'application g n'est autre que la plus petite constante K vérifiant la condition (1.9).

Théorème 9 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et g une application contractante sur $[a,b]$, telle que l'image de $[a,b]$ par g est incluse dans $[a,b]$. Alors, la fonction g possède un unique point fixe ξ dans $[a,b]$. De plus, la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par la relation* (1.8) *converge, pour toute initialisation $x^{(0)}$ dans $[a,b]$, vers ce point fixe et l'on a les estimations suivantes*

$$\forall k \in \mathbb{N},\ \left|x^{(k)}-\xi\right| \leq K^k \left|x^{(0)}-\xi\right|, \tag{1.10}$$

$$\forall k \in \mathbb{N}^*,\ \left|x^{(k)}-\xi\right| \leq \frac{K}{1-K}\left|x^{(k)}-x^{(k-1)}\right|. \tag{1.11}$$

DÉMONSTRATION. On commence par montrer que la suite $(x^{(k)})_{k\in\mathbb{N}}$ est une suite de Cauchy. En effet, on a

$$\forall k \in \mathbb{N}^*,\ \left|x^{(k+1)}-x^{(k)}\right| = \left|g(x^{(k)})-g(x^{(k-1)})\right| \leq K\left|x^{(k)}-x^{(k-1)}\right|,$$

par hypothèse, et on obtient par récurrence que

$$\forall k \in \mathbb{N},\ \left|x^{(k+1)}-x^{(k)}\right| \leq K^k\left|x^{(1)}-x^{(0)}\right|.$$

On en déduit, par une utilisation répétée de l'inégalité triangulaire, que

$$\begin{aligned}\forall k\in\mathbb{N},\ \forall p\in\mathbb{N},\ p>2,\ \left|x^{(k+p)}-x^{(k)}\right| &\leq \left|x^{(k+p)}-x^{(k+p-1)}\right|+\left|x^{(k+p-1)}-x^{(k+p-2)}\right|+\cdots+\left|x^{(k+1)}-x^{(k)}\right| \\ &\leq \left(K^{p-1}+K^{p-2}+\cdots+1\right)\left|x^{(k+1)}-x^{(k)}\right| \\ &\leq \frac{1-K^p}{1-K}K^k\left|x^{(1)}-x^{(0)}\right|,\end{aligned}$$

le dernier membre tendant vers zéro lorsque l'entier k tend vers l'infini. La suite réelle $(x^{(k)})_{k\in\mathbb{N}}$ converge donc vers une limite ξ dans $[a,b]$. L'application g étant continue [5], on déduit alors par un passage à la limite dans (1.8) que $\xi=g(\xi)$. Supposons à présent que g possède deux points fixes ξ et ζ dans l'intervalle $[a,b]$. On a alors

$$0\leq|\xi-\zeta|=|g(\xi)-g(\zeta)|\leq K|\xi-\zeta|,$$

d'où $\xi=\zeta$ puisque $K<1$.
La première estimation se prouve alors par un raisonnement par récurrence sur l'entier k, en écrivant que

$$\forall k\in\mathbb{N}^*,\ \left|x^{(k)}-\xi\right|=\left|g(x^{(k-1)})-g(\xi)\right|\leq K\left|x^{(k-1)}-\xi\right|,$$

et la seconde est obtenue en utilisant que

$$\forall k\in\mathbb{N}^*,\ \forall p\in\mathbb{N}^*,\ \left|x^{(k+p)}-x^{(k)}\right|\leq\frac{1-K^p}{1-K}\left|x^{(k+1)}-x^{(k)}\right|\leq\frac{1-K^p}{1-K}K\left|x^{(k)}-x^{(k-1)}\right|,$$

et en faisant tendre l'entier p vers l'infini. □

Sous les hypothèses du théorème 9, la convergence des itérations de point fixe est assurée quel que soit le choix de la valeur initiale $x^{(0)}$ dans l'intervalle $[a,b]$: c'est donc un nouvel exemple de convergence *globale*. Par ailleurs, un des intérêts de ce résultat est de donner une estimation de la vitesse de convergence de la suite vers sa limite, la première inégalité montrant en effet que la convergence est *géométrique*. La seconde inégalité s'avère particulièrement utile d'un point de vue pratique, car elle fournit à chaque étape un majorant de la distance à la limite (sans pour autant connaître cette dernière) en fonction d'une quantité connue.

Dans la pratique, vérifier que l'application g est K-lipschitzienne n'est pas toujours aisé. Lorsque g est une fonction de classe $\mathscr{C}^1$ sur l'intervalle $[a,b]$, il est cependant possible d'utiliser la caractérisation suivante.

Proposition 10 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et g une fonction de classe $\mathscr{C}^1$, définie de $[a,b]$ dans lui-même, vérifiant*

$$\forall x\in[a,b],\ \left|g'(x)\right|\leq K<1.$$

Alors, l'application g est contractante sur $[a,b]$.

DÉMONSTRATION. D'après le théorème des accroissements finis, pour tous x et y contenus dans l'intervalle $[a,b]$ et distincts, on sait qu'il existe un réel c strictement compris entre x et y tel que

$$|g(x)-g(y)|=\left|g'(c)\right||x-y|,$$

d'où le résultat. □

La dernière proposition permet alors d'affiner le résultat de convergence globale précédent dans ce cas particulier.

Théorème 11 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$ et g une application satisfaisant les hypothèses de la proposition 10. Alors, la fonction g possède un unique point fixe ξ dans $[a,b]$ et la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par* (1.8) *converge, pour toute initialisation $x^{(0)}$ dans $[a,b]$, vers ce point fixe. De plus, on a*

$$\lim_{k\to+\infty}\frac{x^{(k+1)}-\xi}{x^{(k)}-\xi}=g'(\xi),\tag{1.12}$$

la convergence est donc au moins linéaire.

5. C'est par hypothèse une application K-lipschitzienne.

DÉMONSTRATION. La proposition 10 établissant que g est une application contractante sur $[a,b]$, les conclusions du théorème 9 sont valides et il ne reste qu'à prouver l'égalité (1.12). En vertu du théorème des accroissements finis, il existe, pour tout entier naturel k, un réel $\eta^{(k)}$ strictement compris entre $x^{(k)}$ et ξ tel que

$$x^{(k+1)}-\xi = g(x^{(k)})-g(\xi) = g'(\eta^{(k)})(x^{(k)}-\xi).$$

La suite $(x^{(k)})_{k\in\mathbb{N}}$ convergeant vers ξ, cette égalité implique que

$$\lim_{k\to+\infty}\frac{x^{(k+1)}-\xi}{x^{(k)}-\xi} = \lim_{k\to+\infty} g'(\eta^{(k)}) = g'(\xi).$$

□

On notera que ce théorème assure une convergence *au moins linéaire* de la méthode de point fixe. Il apparaît que la quantité $|g'(\xi)|$ est la constante asymptotique d'erreur de la méthode.

En pratique, il est souvent difficile de déterminer *a priori* un intervalle $[a,b]$ sur lequel les hypothèses de la proposition 10 sont satisfaites. Il est néanmoins possible de se contenter d'hypothèses plus faibles, au prix d'un résultat moindre de convergence *locale*.

Théorème 12 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$, g une fonction continue de $[a,b]$ dans lui-même et ξ un point fixe de g dans $[a,b]$. On suppose de plus que g admet une dérivée continue dans un voisinage de ξ, avec $|g'(\xi)|<1$. Alors, la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par* (1.8) *converge vers ξ, pour toute initialisation $x^{(0)}$ choisie suffisamment proche de ξ.*

DÉMONSTRATION. Par hypothèse sur la fonction g, il existe un réel h strictement positif tel que la dérivée g' est continue sur l'intervalle $[\xi-h,\xi+h]$. Puisque $|g'(\xi)|<1$, on peut alors trouver un intervalle $I_\delta=[\xi-\delta,\xi+\delta]$, avec $0<\delta\le h$, tel que $|g'(x)|\le L$, avec $L<1$, pour tout réel x appartenant à I_δ. Pour cela, il suffit de poser $L=\frac{1}{2}(1+|g'(\xi)|)$ et d'utiliser la continuité de g' pour choisir $\delta\le h$ de manière à ce que

$$\forall x\in I_\delta,\ \left|g'(x)-g'(\xi)\right| \le \frac{1}{2}\left(1-\left|g'(\xi)\right|\right).$$

On en déduit alors que

$$\forall x\in I_\delta,\ \left|g'(x)\right| \le \left|g'(x)-g'(\xi)\right| + \left|g'(\xi)\right| \le \frac{1}{2}\left(1-\left|g'(\xi)\right|\right) + \left|g'(\xi)\right| = L.$$

Supposons à présent que, pour un entier naturel k donné, le terme $x^{(k)}$ de la suite définie par la relation de récurrence (1.8) appartienne à I_δ. On a alors, en vertu du théorème des accroissements finis,

$$x^{(k+1)}-\xi = g(x^{(k)})-\xi = g(x^{(k)})-g(\xi) = g'(\eta^{(k)})(x^{(k)}-\xi),$$

avec $\eta^{(k)}$ compris entre $x^{(k)}$ et ξ, d'où

$$\left|x^{(k+1)}-\xi\right| \le L\left|x^{(k)}-\xi\right|,$$

et $x^{(k+1)}$ appartient donc lui aussi à I_δ. On montre alors, en raisonnant par récurrence, que, si $x^{(0)}$ appartient à I_δ, alors tout terme de la suite appartient également à I_δ et

$$\forall k\in\mathbb{N},\ \left|x^{(k)}-\xi\right| \le L^k\left|x^{(0)}-\xi\right|,$$

ce qui implique que la suite $(x^{(k)})_{k\in\mathbb{N}}$ converge vers ξ. □

On peut observer que, si $|g'(\xi)|>1$ et si $x^{(k)}$ est suffisamment proche de ξ pour avoir $\left|g'(x^{(k)})\right|>1$, on obtient $\left|x^{(k+1)}-\xi\right| > \left|x^{(k)}-\xi\right|$ et la convergence ne peut alors avoir lieu (sauf si $x^{(k)}=\xi$). Dans le cas où $|g'(\xi)|=1$, il peut y avoir convergence ou divergence selon les cas considérés. Cette remarque et le précédent théorème conduisent à l'introduction des notions suivantes.

Définitions 13 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$, une fonction g continue de $[a,b]$ dans lui-même et ξ un point fixe de g dans $[a,b]$. On dit que ξ est un point fixe* ***attractif*** *si la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par l'itération de point fixe* (1.8) *converge pour toute initialisation $x^{(0)}$ suffisamment proche de ξ. Réciproquement, si cette suite ne converge pour aucune initialisation $x^{(0)}$ dans un voisinage de ξ, exceptée $x^{(0)}=\xi$, le point fixe est dit* ***répulsif****.*

Terminons cette section par un résultat sur l'ordre de convergence des méthodes de point fixe.

Proposition 14 *Soit $[a,b]$ un intervalle non vide de $\mathbb{R}$, g une fonction continue de $[a,b]$ dans lui-même et ξ un point fixe de g dans $[a,b]$. Si g est de classe $\mathscr{C}^{p+1}$, avec p un entier supérieur ou égal à 1, dans un voisinage de ξ et si $g^{(i)}(\xi)=0$ pour $i=1,\dots,p$ et $g^{(p+1)}(\xi)\neq 0$, alors toute suite convergente $(x^{(k)})_{k\in\mathbb{N}}$ définie par la méthode de point fixe* (1.8) *converge avec un ordre égal à $p+1$ et l'on a*

$$\lim_{k\to+\infty}\frac{x^{(k+1)}-\xi}{(x^{(k)}-\xi)^{p+1}}=\frac{g^{(p+1)}(\xi)}{(p+1)!}.$$

DÉMONSTRATION. Par la formule de Taylor–Lagrange à l'ordre p appliquée à la fonction g au voisinage du point ξ, on obtient

$$\forall k\in\mathbb{N},\ x^{(k+1)}-\xi=\sum_{i=0}^{p}\frac{g^{(i)}(\xi)}{i!}(x^{(k)}-\xi)^i+\frac{g^{(p+1)}(\eta^{(k)})}{(p+1)!}(x^{(k)}-\xi)^{p+1}-g(\xi),$$

avec $\eta^{(k)}$ compris entre $x^{(k)}$ et ξ. Il vient alors

$$\lim_{k\to+\infty}\frac{x^{(k+1)}-\xi}{(x^{(k)}-\xi)^{p+1}}=\lim_{k\to+\infty}\frac{g^{(p+1)}(\eta^{(k)})}{(p+1)!}=\frac{g^{(p+1)}(\xi)}{(p+1)!},$$

par convergence de la suite $(x^{(k)})_{k\in\mathbb{N}}$ et continuité de la fonction $g^{(p+1)}$. □

1.3.3 Exemple de la méthode de Newton–Raphson

En supposant que la fonction f est de classe $\mathscr{C}^1$ et que le zéro ξ est simple, c'est-à-dire que $f(\xi)=0$ et $f'(\xi)\neq 0$, la *méthode de Newton–Raphson* est définie par la relation de récurrence

$$\forall k\in\mathbb{N},\ x^{(k+1)}=x^{(k)}-\frac{f(x^{(k)})}{f'(x^{(k)})},\tag{1.13}$$

l'initialisation $x^{(0)}$ étant donnée.

Dans cette méthode, toute nouvelle approximation du zéro est construite au moyen d'une *linéarisation* de l'équation $f(x)=0$ autour de l'approximation précédente. En effet, si l'on remplace $f(x)$ au voisinage du point $x^{(k)}$ par l'approximation affine obtenue en tronquant au premier ordre le développement de Taylor de f en $x^{(k)}$ et qu'on résout l'équation linéaire résultante

$$f(x^{(k)})+(x-x^{(k)})f'(x^{(k)})=0,$$

en notant sa solution $x^{(k+1)}$, on retrouve l'égalité (1.13). Il en résulte que, géométriquement parlant, le point $x^{(k+1)}$ est l'abscisse du point d'intersection entre la tangente à la courbe de f au point $(x^{(k)},f(x^{(k)}))$ et l'axe des abscisses (voir la figure 1.7).

Par rapport à toutes les méthodes introduites jusqu'à présent, on pourra remarquer que la méthode de Newton nécessite à chaque itération l'évaluation des deux fonctions f et f' au point courant $x^{(k)}$. Cet effort est compensé par une vitesse de convergence accrue, puisque cette méthode est d'ordre deux si le zéro recherché est simple.

Théorème 15 (convergence locale de la méthode de Newton–Raphson) *Soit f une fonction réelle de classe $\mathscr{C}^2$ dans un voisinage d'un zéro simple ξ. Alors, la suite $(x^{(k)})_{k\in\mathbb{N}}$ définie par* (1.13) *converge au moins quadratiquement vers ξ, pour toute initialisation $x^{(0)}$ choisie suffisamment proche de ce zéro.*

DÉMONSTRATION. Nous allons tout d'abord prouver la convergence locale de la méthode et ensuite obtenir son ordre de convergence. À cette fin, introduisons, pour un réel δ strictement positif, l'ensemble $I_\delta=\{x\in\mathbb{R}\,|\,|x-\xi|\leq\delta\}$ et supposons que f soit de classe $\mathscr{C}^2$ dans ce voisinage de ξ. Définissons alors, pour δ suffisamment petit, la quantité

$$M(\delta)=\max_{\substack{s\in I_\delta\\ t\in I_\delta}}\left|\frac{f''(s)}{2f'(t)}\right|,$$

et supposons que δ soit tel que [6]

$$2\delta M(\delta)<1.\tag{1.14}$$

6. Notons que $\lim_{\delta\to 0}M(\delta)=\left|\frac{f''(\xi)}{2f'(\xi)}\right|<+\infty$, puisqu'on a fait l'hypothèse que ξ est un zéro simple de f. On peut donc bien satisfaire la condition (1.14) pour δ assez petit.

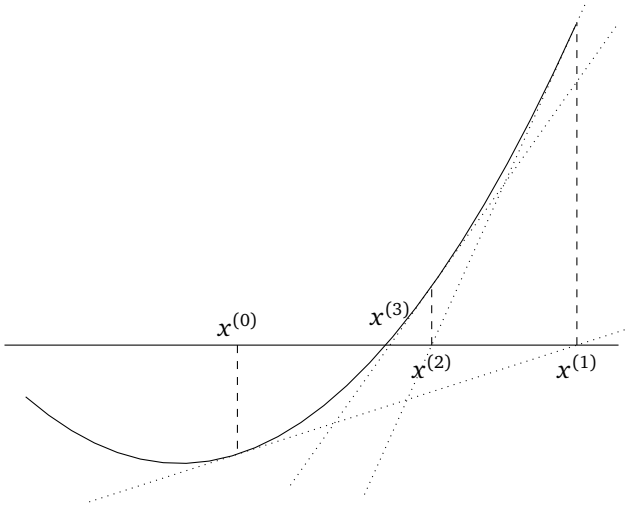

FIGURE 1.7: Construction des premiers itérés de la méthode de Newton–Raphson.

Montrons à présent que le réel ξ est l'unique zéro de f contenu dans I_δ. En appliquant la formule de Taylor–Lagrange à l'ordre un à la fonction f au voisinage du point ξ, on trouve que

$$f(x) = f(\xi) + (x-\xi)f'(\xi) + \frac{1}{2}(x-\xi)^2 f''(\eta),$$

avec η compris entre x et ξ. Par conséquent, si x appartient à I_δ, on a également que η appartient à I_δ et l'on obtient

$$f(x) = (x-\xi)f'(\xi)\left(1 + (x-\xi)\frac{f''(\eta)}{2f'(\xi)}\right).$$

Si x appartient à I_δ et $x \neq \xi$, les trois facteurs dans le membre de droite sont tous non nuls (le dernier parce que $\left|(x-\xi)\frac{f''(\eta)}{2f'(\xi)}\right| \leq \delta M(\delta) < \frac{1}{2}$) et la fonction f ne s'annule donc qu'en ξ sur l'intervalle I_δ. Prouvons d'autre part que la fonction f' ne s'annule pas sur I_δ. On a en effet

$$f'(x) = f'(\xi) + (x-\xi)f''(\mu),$$

avec μ compris entre x et ξ. Comme précédemment, si x appartient à I_δ, alors μ appartient à I_δ, d'où

$$\forall x \in I_\delta,\ \left|f'(x)\right| \geq \left|f'(\xi)\right| - \left|(x-\xi)f''(\mu)\right| \geq \left|f'(\xi)\right|\,(1-\delta M(\delta)) > \frac{1}{2}\left|f'(\xi)\right| > 0,$$

ce qui assure que la méthode de Newton donnée par (1.13) est bien définie quel que soit $x^{(k)}$ dans l'intervalle I_δ.

Montrons à présent que, pour tout choix de $x^{(0)}$ dans I_δ, la suite $(x^{(k)})_{k\in\mathbb{N}}$ construite par la méthode de Newton est contenue dans l'intervalle I_δ et converge vers ξ. Tout d'abord, si, pour un entier naturel k, $x^{(k)}$ appartient à I_δ, il découle de (1.13) et de la formule de Taylor–Lagrange que

$$x^{(k+1)} - \xi = (x^{(k)} - \xi)^2 \frac{f''(\eta^{(k)})}{2f'(x^{(k)})}, \tag{1.15}$$

avec $\eta^{(k)}$ compris entre $x^{(k)}$ et ξ, d'où

$$\left|x^{(k+1)} - \xi\right| < \frac{1}{2}\left|x^{(k)} - \xi\right| \leq \frac{\delta}{2}.$$

On en conclut que tous les termes de la suite $(x^{(k)})_{k\in\mathbb{N}}$ sont contenus dans I_δ en raisonnant par récurrence sur l'indice k. On obtient également l'estimation suivante

$$\forall k \in \mathbb{N},\ \left|x^{(k)} - \xi\right| \leq \frac{1}{2^k}\left|x^{(0)} - \xi\right| \leq \frac{\delta}{2^k},$$

ce qui implique la convergence de la méthode.

Pour établir le fait que la suite converge quadratiquement, on se sert de (1.15) pour trouver

$$\lim_{k\to+\infty} \frac{\left|x^{(k+1)} - \xi\right|}{\left|x^{(k)} - \xi\right|^2} = \lim_{k\to+\infty} \left|\frac{f''(\eta^{(k)})}{2f'(x^{(k)})}\right| = \left|\frac{f''(\xi)}{2f'(\xi)}\right|,$$

en vertu de la convergence de la suite $(x^{(k)})_{k\in\mathbb{N}}$ et de la continuité de f' et f'' sur l'intervalle I_δ. □

On notera que ce théorème ne garantit la convergence de la méthode de Newton–Raphson que si l'initialisation $x^{(0)}$ est *« suffisamment proche »* du zéro recherché. La méthode peut en effet diverger lorsque ce n'est pas le cas.

Il est également important d'ajouter que, bien que la méthode de Newton–Raphson converge quadratiquement vers un zéro simple, la notion d'ordre de convergence est *asymptotique*. De fait, on constate parfois que la méthode converge tout d'abord linéairement pour ensuite atteindre une convergence quadratique, une fois arrivé suffisamment près du zéro recherché. La figure 1.8 illustre ce phénomène.

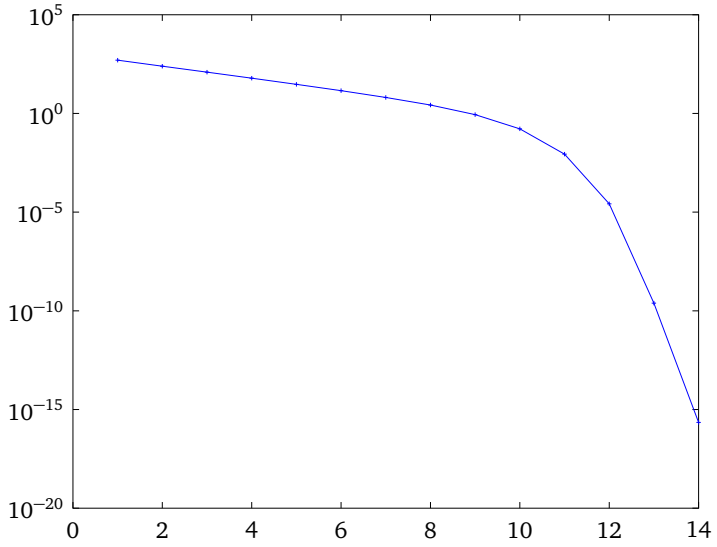

FIGURE 1.8: Tracé de l'erreur absolue en fonction du nombre d'itérations de la méthode de Newton–Raphson utilisée pour la détermination de la racine positive de l'équation $x^2 - 2 = 0$. On a choisi $x^{(0)} = 1000$, ce qui constitue évidemment une très mauvaise estimation initiale, mais permet de mettre en évidence une période transitoire durant laquelle la convergence de la méthode est seulement *linéaire*.

En conclusion, la méthode de Newton–Raphson est la méthode de choix en termes de vitesse de convergence, puisque, dans les cas favorables, les approximations successives du zéro recherché convergent de manière quadratique, ce qui se traduit *grosso modo* par un doublement du nombre de décimales exactes de l'approximation à chaque itération de l'algorithme. Elle nécessite pour cela que la dérivée de la fonction f puisse être évaluée en tout point donné. Si cela n'est pas le cas, on utilisera la méthode de la sécante (voir la prochaine section), dont la vitesse de convergence est moindre mais ne requiert pas que la dérivée de f soit connue.

La plus grande difficulté dans l'utilisation de la méthode de Newton–Raphson réside dans le caractère local de sa convergence. Si l'initialisation $x^{(0)}$ est trop éloignée du zéro, la méthode peut très bien diverger. Pour cette raison, il est courant dans les applications de l'associer à une méthode d'encadrement, comme la méthode de dichotomie, cette dernière permettant d'approcher, bien que lentement, le zéro recherché de manière à fournir une « bonne » initialisation pour la méthode de Newton–Raphson.

1.4 Méthode de la sécante

La *méthode de la sécante* (*secant method* en anglais) peut être considérée comme une modification de la méthode de la fausse position permettant de se passer de l'hypothèse sur le signe de la fonction f aux extrémités de l'intervalle d'encadrement initial (il n'y a d'ailleurs plus besoin de connaître un tel intervalle). On peut aussi la voir comme une *quasi*-méthode de Newton–Raphson, dans laquelle on aurait remplacé la valeur $f'(x^{(k)})$ par

une approximation obtenue par une différence finie. C'est l'une des méthodes que l'on peut employer lorsque la dérivée de f est compliquée, voire impossible [7], à calculer ou encore coûteuse à évaluer.

Plus précisément, à partir de la donnée de deux valeurs initiales $x^{(-1)}$ et $x^{(0)}$, telles que $x^{(-1)} \neq x^{(0)}$, la méthode de la sécante consiste en l'utilisation de la relation de récurrence

$$\forall k \in \mathbb{N},\ x^{(k+1)} = x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f(x^{(k)}) - f(x^{(k-1)})} f(x^{(k)}), \tag{1.16}$$

pour obtenir les approximations successives du zéro recherché. Elle tire son nom de l'interprétation géométrique de (1.16) : pour tout entier positif k, le point $x^{(k+1)}$ est le point d'intersection de l'axe des abscisses avec la droite passant par les points $(x^{(k-1)}, f(x^{(k-1)}))$ et $(x^{(k)}, f(x^{(k)}))$ de la courbe représentative de la fonction f (voir la figure 1.9).

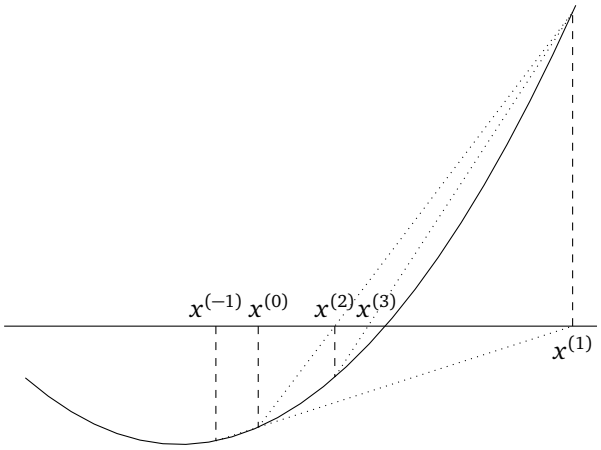

FIGURE 1.9: Construction des premiers itérés de la méthode de la sécante.

Bien que l'on doive disposer de deux estimations de ξ avant de pouvoir utiliser la relation de récurrence (1.16), cette méthode ne requiert à chaque étape qu'une seule évaluation de fonction, ce qui est un avantage par rapport à la méthode de Newton–Raphson, dont la relation de récurrence demande de connaître les valeurs de $f(x^{(k)})$ et de $f'(x^{(k)})$. En revanche, à la différence de la méthode de la fausse position, rien n'assure que, pour tout entier naturel k, au moins un zéro de f se trouve entre les points $x^{(k-1)}$ et $x^{(k)}$.

Le résultat suivant, dont la preuve est admise, montre que la convergence de la méthode est *superlinéaire*, mais seulement *locale*.

Théorème 16 (convergence locale de la méthode de la sécante) *Soit f une fonction de classe $\mathscr{C}^2$ dans un voisinage d'un zéro simple ξ. Alors, si les données $x^{(-1)}$ et $x^{(0)}$, avec $x^{(-1)} \neq x^{(0)}$, choisies dans ce voisinage, sont suffisamment proches de ξ, la suite définie par* (1.16) *converge vers ξ avec un ordre au moins égal à $\frac{1}{2}(1+\sqrt{5}) = 1,6180339887\ldots$*

On notera que la méthode de la sécante *n'est pas* une méthode de point fixe, la relation de récurrence (1.16) définissant la méthode ne pouvant s'écrire sous la forme $x^{(k+1)} = g(x^{(k)})$.

1.5 Critères d'arrêt

Mis à part dans le cas de la méthode de dichotomie, nous n'avons (volontairement) pas abordé la question du *critère d'arrêt* à utiliser en pratique. En effet, s'il y a convergence, la suite $(x^{(k)})_{k\in\mathbb{N}}$ construite par une méthode itérative tend vers le zéro ξ quand k tend vers l'infini, et il faut donc introduire un critère permettant d'interrompre le processus itératif lorsque l'approximation courante de ξ est jugée « satisfaisante ». Pour cela, on a principalement le choix entre deux types de critères « qualitatifs » (imposer un nombre maximum d'itérations constituant une troisième possibilité strictement « quantitative ») : l'un basé sur l'incrément et l'autre sur le résidu.

7. C'est le cas si la fonction f n'est connue qu'*implicitement*, par exemple lorsque que c'est la solution d'une équation différentielle et que x est un paramètre de la donnée initiale du problème associé.

Quel que soit le critère retenu, notons $\varepsilon > 0$ la tolérance fixée pour le calcul approché de ξ. Dans le cas d'un *contrôle de l'incrément*, les itérations s'achèveront dès que

$$\left|x^{(k+1)} - x^{(k)}\right| < \varepsilon, \tag{1.17}$$

alors qu'on mettra fin au calcul dès que

$$\left|f(x^{(k)})\right| < \varepsilon, \tag{1.18}$$

si l'on choisit de *contrôler le résidu*.

Selon les configurations, chacun de ces critères peut s'avérer plus ou moins bien adapté. Pour s'en convaincre, considérons la suite $(x^{(k)})_{k\in\mathbb{N}}$ produite par une méthode de point fixe, en supposant la fonction g continûment différentiable dans un voisinage de ξ. Par un développement au premier ordre, on obtient

$$\forall k \in \mathbb{N},\ x^{(k+1)} - \xi = g(x^{(k)}) - g(\xi) = g'(\eta^{(k)})(x^{(k)} - \xi),$$

avec $\eta^{(k)}$ un réel compris entre $x^{(k)}$ et ξ. On a alors

$$\forall k \in \mathbb{N},\ x^{(k+1)} - x^{(k)} = x^{(k+1)} - \xi - (x^{(k)} - \xi) = (g'(\eta^{(k)}) - 1)(x^{(k)} - \xi),$$

dont on déduit, en cas de convergence, le comportement asymptotique suivant

$$x^{(k)} - \xi \simeq \frac{1}{g'(\xi) - 1}(x^{(k+1)} - x^{(k)}).$$

Par conséquent, le critère d'arrêt (1.17), basé sur l'incrément, sera indiqué si $-1 < g'(\xi) \leq 0$ (il est d'ailleurs optimal pour une méthode de point fixe dont la convergence est au moins quadratique, c'est-à dire pour laquelle $g'(\xi) = 0$), mais très peu satisfaisant si $g'(\xi)$ est proche de 1.

Considérons maintenant le cas d'un critère basé sur le résidu, en supposant la fonction f continûment différentiable dans un voisinage d'un zéro simple ξ. En cas de convergence de la méthode et pour $k \geq 0$ assez grand, il vient, par la formule de Taylor–Young,

$$f(x^{(k)}) = f'(\xi)(\xi - x^{(k)}) + (\xi - x^{(k)})\epsilon(\xi - x^{(k)}),$$

avec $\epsilon(x)$ une fonction définie dans un voisinage de l'origine et tendant vers 0 quand x tend vers 0, dont on déduit l'estimation

$$\left|x^{(k)} - \xi\right| \lesssim \frac{\left|f(x^{(k)})\right|}{|f'(\xi)|}.$$

Le critère (1.18) fournira donc un test d'arrêt adéquat lorsque $|f'(\xi)| \simeq 1$, mais s'avérera trop restrictif si $|f'(\xi)| \gg 1$ ou en revanche trop optimiste si $|f'(\xi)| \ll 1$.

Chapitre 2

Interpolation polynomiale

Historiquement, le terme d'*interpolation* regroupe l'ensemble des techniques permettant de construire une courbe d'un type donné passant par un nombre fini de points donnés du plan. D'un point de vue applicatif, les ordonnées de ces points peuvent représenter les valeurs aux abscisses d'une fonction arbitraire, que l'on cherche dans ce cas à remplacer par une fonction plus simple à manipuler lors d'un calcul numérique, ou encore de données expérimentales, pour lesquelles on vise à obtenir empiriquement une loi de distribution lorsque leur nombre est important.

Nous nous limiterons ici à des problèmes d'interpolation *polynomiale*, ce qui signifie que la courbe que l'on cherche à obtenir est le graphe d'une fonction polynomiale (éventuellement par morceaux). En effet, les nombreuses propriétés analytiques et algébriques des polynômes, alliées à la facilité que l'on a à les dériver, les intégrer ou les évaluer numériquement en un point, en font une classe de fonctions extrêmement intéressante en pratique. L'interpolation polynomiale est pour cette raison un outil numérique de premier ordre pour l'*approximation polynomiale* des fonctions réelles d'une variable réelle.

2.1 Interpolation de Lagrange

Soit n un entier naturel. Dans l'ensemble de cette section, on suppose que la famille $\{(x_i, y_i)\}_{i=0,\dots,n}$, est un ensemble de $n+1$ points du plan euclidien dont les abscisses sont *toutes deux à deux distinctes*.

2.1.1 Définition du problème d'interpolation

Le problème d'interpolation de Lagrange s'énonce en ces termes : *étant donné une famille de $n+1$ couples (x_i, y_i), $i = 0, \dots, n$, distincts de nombres réels, trouver un polynôme Π_n de degré inférieur ou égal à n dont le graphe de la fonction polynomiale associée passe par les $n+1$ points du plan ainsi définis*. Plus concrètement, ceci signifie que le polynôme Π_n solution de ce problème, appelé *polynôme d'interpolation*, ou *interpolant*, *de Lagrange associé aux points* $\{(x_i, y_i)\}_{i=0,\dots,n}$, satisfait les contraintes

$$\Pi_n(x_i) = y_i, \ i = 0, \dots, n. \tag{2.1}$$

On dit encore qu'il *interpole* les quantités y_i aux *nœuds* x_i, $i = 0, \dots, n$.

Commençons par montrer que ce problème de détermination est bien posé, c'est-à-dire qu'il admet une unique solution.

Théorème 17 (existence et unicité du polynôme d'interpolation de Lagrange) *Soit n un entier naturel. Étant donné $n+1$ points distincts $x_0, \dots, x_n$ et $n+1$ valeurs $y_0, \dots, y_n$, il existe un unique polynôme Π_n de $\mathbb{P}_n$ satisfaisant* (2.1).

DÉMONSTRATION. Le polynôme Π_n recherché étant de degré n, on peut poser

$$\forall x \in \mathbb{R}, \ \Pi_n(x) = \sum_{j=0}^{n} a_j \, x^j, \tag{2.2}$$

et ramener le problème d'interpolation à la détermination des coefficients a_j, $j = 0, \ldots, n$. En utilisant les conditions $\Pi_n(x_i) = y_i$, $i = 0, \ldots, n$, on arrive à un système linéaire à $n+1$ équations et $n+1$ inconnues :

$$a_0 + a_1 x_i + \cdots + a_n {x_i}^n = y_i, \ i = 0, \ldots, n. \tag{2.3}$$

Ce système possède une unique solution si et seulement si la matrice carrée qui lui est associée est inversible. Or, il se trouve que cette dernière est une matrice de Vandermonde dont le déterminant vaut

$$\begin{vmatrix} 1 & x_0 & \ldots & {x_0}^n \\ 1 & x_1 & \ldots & {x_1}^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & {x_n}^n \end{vmatrix} = \prod_{0 \le i < j \le n} (x_j - x_i) = \prod_{i=0}^{n-1} \left(\prod_{j=i+1}^{n} (x_j - x_i) \right).$$

Les nœuds d'interpolation étant tous distincts, ce déterminant est non nul. □

On notera qu'il est également possible de prouver l'unicité du polynôme d'interpolation en supposant qu'il existe un autre polynôme Ψ_m, de degré m inférieur ou égal à n, tel que $\Psi_m(x_i) = y_i$ pour $i = 0, \ldots, n$. La différence $\Pi_n - \Psi_m$ s'annulant en $n+1$ points distincts, il découle du théorème fondamental de l'algèbre qu'elle est nulle.

Pour obtenir les coefficients du polynôme Π_n dans la base canonique de l'anneau des polynômes, il suffit donc de résoudre le système linéaire (2.3). On peut cependant montrer que les matrices de Vandermonde sont généralement très mal conditionnées, quel que soit le choix de nœuds d'interpolation et la résolution numérique des systèmes associés par une méthode directe est alors sujette à des problèmes de stabilité, en plus de s'avérer coûteuse lorsque le nombre de nœuds est important [1].

2.1.2 Différentes représentations du polynôme d'interpolation de Lagrange

Nous venons de voir comment obtenir le polynôme d'interpolation de Lagrange dans la base canonique de l'anneau des polynômes et les inconvénients de cette approche. Une autre possibilité consiste à utiliser une représentation différente de ce polynôme, de manière à ce que sa détermination soit rendue particulièrement aisée. C'est ce que l'on fait en adaptant la base choisie de façon à ce que la matrice du système linéaire associé au problème soit diagonale ou triangulaire.

Forme de Lagrange

Commençons par introduire les *polynômes de Lagrange* et leurs propriétés.

Définition 18 *Soit n un entier naturel non nul. On appelle* ***polynômes de Lagrange associés aux nœuds*** $\{x_i\}_{i=0,\ldots,n}$ *les $n+1$ polynômes $l_i \in \mathbb{P}_n$, $i = 0, \ldots, n$, définis par*

$$\forall i \in \{0, \ldots, n\}, \ \forall x \in \mathbb{R}, \ l_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}. \tag{2.4}$$

Bien que communément employée pour ne pas alourdir les écritures, la notation l_i, $i = 0, \ldots, n$, utilisée pour les polyômes de Lagrange ne fait pas explicitement apparaître leur degré, la valeur de l'entier n étant fixée et généralement claire compte tenu du contexte. Il faudra cependant garder cette remarque à l'esprit, puisque l'on peut être amené à faire tendre tendre cette valeur vers l'infini. Ajoutons que, si l'on a exigé que l'entier n soit supérieur ou égal à 1 dans la définition, le cas trivial $n = 0$ peut être inclus dans tout ce qui va suivre en posant $l_0 \equiv 1$ si $n = 0$.

Proposition 19 *Soit n un entier naturel. Les polynômes de Lagrange $\{l_i\}_{i=0,\ldots,n}$ sont de degré n, vérifient $l_i(x_k) = \delta_{ik}$, $i, k = 0, \ldots, n$, où δ_{ik} désigne le symbole de Kronecker, et forment une base de $\mathbb{P}_n$.*

1. On a vu précédemment que le coût de la résolution d'un système d'ordre n par la méthode d'élimination de Gauss était de l'ordre de $\frac{2}{3} n^3$ opérations arithmétiques.

DÉMONSTRATION. Le résultat est évident si l'entier n est nul. S'il est non nul, les deux premières propriétés découlent directement de la définition (2.4) des polynômes de Lagrange. On déduit ensuite de la deuxième propriété que, si le polynôme $\sum_{i=0}^{n} \lambda_i \, l_i$, les coefficients $\lambda_1, \ldots, \lambda_n$ étant réels, est identiquement nul, alors on a

$$\forall j \in \{1, \ldots, n\}, \ 0 = \sum_{i=0}^{n} \lambda_i \, l_i(x_j) = \lambda_j.$$

La famille $\{l_i\}_{i=0,\ldots,n}$ est donc libre et forme une base de $\mathbb{P}_n$. □

On déduit de la proposition 19 le résultat suivant.

Théorème 20 (« formule d'interpolation de Lagrange ») *Soit n un entier naturel. Étant donné $n+1$ points distincts $x_0, \ldots, x_n$ et $n+1$ valeurs $y_0, \ldots, y_n$, le polynôme d'interpolation Π_n de $\mathbb{P}_n$ tel que $\Pi_n(x_i) = y_i$, $i = 0, \ldots, n$, est donné par*

$$\forall x \in \mathbb{R}, \ \Pi_n(x) = \sum_{i=0}^{n} y_i \, l_i(x). \tag{2.5}$$

DÉMONSTRATION. Pour établir (2.5), on utilise que la famille de polynômes $\{l_i\}_{i=0,\ldots,n}$ forment une base de $\mathbb{P}_n$. La décomposition de Π_n dans cette base s'écrit $\Pi_n = \sum_{i=0}^{n} \mu_i \, l_i$, et on a alors

$$\forall j \in \{0, \ldots, n\}, \ y_j = \Pi_n(x_j) = \sum_{i=0}^{n} \mu_i \, l_i(x_j) = \mu_j.$$

□

L'évaluation du polynôme d'interpolation Π_n sous sa forme de Lagrange (2.5) en un point autre que l'un des nœuds d'interpolation demande d'évaluer chacun des polynômes de Lagrange l_i, $i = 0, \ldots, n$, en ce point et nécessite au total n additions, $\frac{1}{2}(n+2)(n+1)$ soustractions, $(2n+1)(n+1)$ multiplications et $n+1$ divisions. La « mise à jour [2] » de ce même polynôme, c'est-à-dire l'opération consistant à obtenir le polynôme Π_{n+1} associé à $n+2$ couples (x_i, y_i), $i = 0, \ldots, n+1$, à partir de la donnée du polynôme Π_n associé aux paires (x_i, y_i), $i = 0, \ldots, n$, et de celle du couple (x_{n+1}, y_{n+1}), est malaisée, car la base des polynômes de Lagrange servant à écrire Π_{n+1} est différente de celle utilisée pour Π_n. Pour ces raisons, la formule d'interpolation de Lagrange (2.5) est généralement considérée comme un outil théorique, peu utile en pratique, et le recours à la *forme de Newton* du polynôme d'interpolation est souvent recommandé.

Forme de Newton

La forme de Newton du polynôme d'interpolation offre une alternative à la formule (2.5) qui facilite à la fois l'évaluation et la mise à jour du polynôme d'interpolation après que certaines quantités, indépendantes du point auquel on évalue le polynôme, ont été calculées. Afin de l'expliciter, nous allons chercher à écrire le polynôme d'interpolation de Lagrange Π_n, avec n un entier strictement positif, associé aux nœuds d'interpolation $x_0, \ldots, x_n$, comme la somme du polynôme Π_{n-1}, tel que $\Pi_{n-1}(x_i) = y_i$ pour $i = 0, \ldots, n-1$, et d'un polynôme de degré n, qui dépendra des nœuds $x_0, \ldots, x_{n-1}$ et d'un seul autre coefficient que l'on devra déterminer. Posons ainsi

$$\Pi_n(x) = \Pi_{n-1}(x) + q_n(x), \tag{2.6}$$

où q_n appartient à $\mathbb{P}_n$. Puisque $q_n(x_i) = \Pi_n(x_i) - \Pi_{n-1}(x_i) = 0$ pour $i = 0, \ldots, n-1$, on a nécessairement

$$q_n(x) = a_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}).$$

Notons alors

$$\forall x \in \mathbb{R}, \ \omega_n(x) = \prod_{j=0}^{n-1} (x - x_j) \tag{2.7}$$

2. Cette opération est primordiale dans le cadre de l'approximation polynomiale d'une fonction par le biais de l'interpolation. En effet, lorsqu'on ne sait *a priori* pas combien de points sont nécessaires pour approcher une fonction donnée par son polynôme d'interpolation de Lagrange avec une précision fixée, il est particulièrement utile de pouvoir introduire, un à un, de nouveaux nœuds d'interpolation jusqu'à satisfaction.

le *polynôme de Newton de degré n associé aux nœuds* $\{x_i\}_{i=0,\dots,n-1}$ et déterminons le coefficient a_n. Puisque $\Pi_n(x_n) = y_n$, on déduit de (2.6) que

$$a_n = \frac{y_n - \Pi_{n-1}(x_n)}{\omega_n(x_n)}.$$

Le coefficient a_n donné par la formule ci-dessus est appelée la *n^e différence divisée de Newton* et se note généralement [3]

$$a_n = [x_0, x_1, \dots, x_n]\boldsymbol{y}.$$

On a par conséquent

$$\forall x \in \mathbb{R},\ \Pi_n(x) = \Pi_{n-1}(x) + [x_0, x_1, \dots, x_n]\boldsymbol{y}\ \omega_n(x). \tag{2.8}$$

En posant $[x_0]\boldsymbol{y} = y_0$ et $\omega_0 \equiv 1$, on obtient, à partir de (2.8) et en raisonnant par récurrence sur le degré n, que

$$\forall x \in \mathbb{R},\ \Pi_n(x) = \sum_{i=0}^{n} [x_0, \dots, x_i]\boldsymbol{y}\ \omega_i(x), \tag{2.9}$$

qui est, en vertu de l'unicité du polynôme d'interpolation, le même polynôme que celui défini par la formule (2.5). La forme (2.9) est appelée *formule des différences divisées de Newton du polynôme d'interpolation*. Ce n'est autre que l'écriture de Π_n dans la base [4] de $\mathbb{P}_n$ formée par la famille de polynômes de Newton $\{\omega_i\}_{i=0,\dots,n}$. On remarquera que, écrite dans la base des polynômes de Newton, la matrice du système linéaire associé au problème d'interpolation de Lagrange est

$$\begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ 1 & x_1 - x_0 & 0 & & & \vdots \\ 1 & x_2 - x_0 & (x_2 - x_0)(x_2 - x_1) & 0 & & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_{n-1} - x_0 & (x_{n-1} - x_0)(x_{n-1} - x_1) & \dots & \prod_{j=0}^{n-2}(x_{n-1} - x_j) & 0 \\ 1 & x_n - x_0 & (x_n - x_0)(x_n - x_1) & \dots & \prod_{j=0}^{n-2}(x_n - x_j) & \prod_{j=0}^{n-1}(x_n - x_j) \end{pmatrix}. \tag{2.10}$$

Les différences divisées $[x_0]\boldsymbol{y}, [x_0, x_1]\boldsymbol{y}, \dots, [x_0, \dots, x_n]\boldsymbol{y}$ sont donc solution d'un système triangulaire inférieur et peuvent par conséquent être obtenues au moyen d'une méthode de descente), après calcul des coefficients de la matrice ci-dessus.

Les différences divisées possèdent plusieurs propriétés algébriques. Tout d'abord, on peut vérifier, à titre d'exercice, que l'on a

$$\forall i \in \{0, \dots, n\},\ \forall x \in \mathbb{R},\ l_i(x) = \frac{\omega_{n+1}(x)}{\omega'_{n+1}(x_i)(x - x_i)}, \tag{2.11}$$

où ω_{n+1} est le polynôme de Newton de degré $n+1$ associé aux nœuds $x_0, \dots, x_n$, et la formule (2.5) se réécrit donc

$$\forall x \in \mathbb{R},\ \Pi_n(x) = \omega_{n+1}(x) \sum_{i=0}^{n} \frac{y_i}{\omega'_{n+1}(x_i)(x - x_i)}. \tag{2.12}$$

En utilisant alors l'écriture (2.9) pour identifier le coefficient $[x_0, \dots, x_n]\boldsymbol{y}$ avec celui qui lui correspond dans l'identité (2.12), on en déduit la forme explicite

$$[x_0, \dots, x_n]\boldsymbol{y} = \sum_{i=0}^{n} \frac{y_i}{\omega'_{n+1}(x_i)} = \sum_{i=0}^{n} \frac{y_i}{\prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j)} \tag{2.13}$$

3. On peut trouver dans la littérature de nombreuses notations pour les différences divisées. Celle choisie dans ces pages, à savoir $[\dots]\boldsymbol{y}$ ou, plus loin, $[\dots]f$ lorsque la différence divisée est appliquée aux valeurs prises aux nœuds par une fonction f continue, vise à mettre en avant le fait que $[\dots]$ désigne un opérateur, dépendant de l'ensemble des nœuds d'interpolation apparaissant en place de $\dots$.

4. On montre en effet par récurrence que $\{\omega_i\}_{i=0,\dots,n}$ est une famille de $n+1$ polynômes *échelonnée en degré* (*i.e.*, que le polynôme ω_i, $i = 0, \dots, n$, est de degré i).

pour la n^{e} différence divisée. Parmi toutes les conséquences de cette dernière expression, il en est une particulièrement importante pour la mise en œuvre de la forme de Newton du polynôme d'interpolation. En effet, par une manipulation algébrique, il vient

$$[x_0,\dots,x_n]\boldsymbol{y} = \frac{[x_1,\dots,x_n]\boldsymbol{y} - [x_0,\dots,x_{n-1}]\boldsymbol{y}}{x_n - x_0},$$

dont on déduit la formule de récurrence

$$[x_{i-k},\dots,x_i]\boldsymbol{y} = \frac{[x_{i-k+1},\dots,x_i]\boldsymbol{y} - [x_{i-k},\dots,x_{i-1}]\boldsymbol{y}}{x_i - x_{i-k}},\ i = k,\dots,n,\ k = 0,\dots,n, \tag{2.14}$$

de laquelle les différences divisées tirent leur nom et qui fournit un procédé pour leur calcul effectif. Ce dernier consiste en la construction du tableau suivant

$$\begin{array}{c|cccccc} x_0 & [x_0]\boldsymbol{y} & & & & \\ x_1 & [x_1]\boldsymbol{y} & [x_0,x_1]\boldsymbol{y} & & & \\ x_2 & [x_2]\boldsymbol{y} & [x_1,x_2]\boldsymbol{y} & [x_0,x_1,x_2]\boldsymbol{y} & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & \\ x_n & [x_n]\boldsymbol{y} & [x_{n-1},x_n]\boldsymbol{y} & [x_{n-2},x_{n-1},x_n]\boldsymbol{y} & \cdots & [x_0,\dots,x_n]\boldsymbol{y} \end{array} \tag{2.15}$$

au sein duquel les différences divisées sont disposées de manière à ce que leur évaluation se fasse de proche en proche en observant la règle suivante : la valeur d'une différence est obtenue en soustrayant à la différence placée immédiatement à sa gauche celle située au dessus de cette dernière, puis en divisant le résultat par la différence entre les deux points de l'ensemble $\{x_i\}_{i=0,\dots,n}$ situés respectivement sur la ligne de la différence à calculer et sur la dernière ligne atteinte en remontant diagonalement dans le tableau à partir de cette même différence.

Les différences divisées apparaissant dans la forme de Newton (2.9) du polynôme d'interpolation de Lagrange sont les $n+1$ coefficients diagonaux du tableau (2.15). Leur obtention requiert par conséquent $(n+1)n$ soustractions et $\frac{1}{2}(n+1)n$ divisions. Si ce coût est du même ordre que celui requis par la résolution d'un système linéaire triangulaire, il s'avère que la construction du tableau des différences divisées est bien moins susceptible de produire des débordements vers l'infini ou vers zéro en arithmétique à virgule flottante que le calcul des éléments de la matrice (2.10).

Exemple de calcul de la forme de Newton du polynôme d'interpolation. Calculons le polynôme d'interpolation de Lagrange prenant les valeurs 4, −1, 4 et 6 aux points respectifs −1, 1, 2 et 3, en tirant parti de (2.9) et de la méthode de calcul des différences divisées basée sur la formule (2.14). Nous avons

$$\begin{array}{r|cccc} -1 & 4 & & & \\ 1 & -1 & (-1-4)/(1-(-1)) = -5/2 & & \\ 2 & 4 & (4-(-1))/(2-1) = 5 & (5-(-5/2))/(2-(-1)) = 5/2 & \\ 3 & 6 & (6-4)(3-2) = 2 & (2-5)/(3-1) = -3/2 & (-3/2-5/2)/(3-(-1)) = -1 \end{array}$$

d'où

$$\Pi_3(x) = 4 - \frac{5}{2}(x+1) + \frac{5}{2}(x+1)(x-1) - (x+1)(x-1)(x-2).$$

Il découle enfin de la représentation (2.13) que les différences divisées sont des *fonctions symétriques de leurs arguments*. On a en effet

$$[x_0,\dots,x_n]\boldsymbol{y} = [x_{\sigma(0)},\dots,x_{\sigma(n)}]\boldsymbol{y}, \tag{2.16}$$

pour toute permutation σ de l'ensemble $\{0,\dots,n\}$.

Une fois les différences divisées calculées, l'évaluation du polynôme d'interpolation Π_n, sous sa forme de Newton, en un point autre que l'un des nœuds d'interpolation se fait au moyen d'une généralisation de la *méthode de Horner*, en remarquant que l'on a

$$\Pi_n(x) = (\dots([x_0,\dots,x_n]\boldsymbol{y}\,(x-x_{n-1}) + [x_0,\dots,x_{n-1}]\boldsymbol{y})(x-x_{n-2}) + \dots + [x_0,x_1]\boldsymbol{y})(x-x_0) + [x_0]\boldsymbol{y}.$$

Le calcul de la valeur de $\Pi_n(x)$ nécessite alors n additions, n soustractions et n multiplications. Pour mettre à jour le polynôme d'interpolation, il suffit simplement, disposant d'une valeur y_{n+1} associée à un nœud x_{n+1}, de calculer et d'ajouter la ligne supplémentaire $[x_{n+1}]\boldsymbol{y}\ [x_n,x_{n+1}]\boldsymbol{y}\ \cdots\ [x_0,\dots,x_{n+1}]\boldsymbol{y}$ au tableau des différences divisées existant, ce qui nécessite $2(n+1)$ soustractions et $n+1$ divisions.

2.1.3 Interpolation polynomiale d'une fonction

L'intérêt de remplacer une fonction quelconque par un polynôme l'approchant aussi précisément que voulu sur un intervalle donné est évident d'un point de vue numérique et informatique, puisqu'il est très aisé de stocker et de manipuler, c'est-à-dire additionner, multiplier, dériver ou intégrer, des polynômes dans un calculateur. Pour ce faire, il semble naturel de chercher à utiliser un polynôme d'interpolation de Lagrange associé aux valeurs prises par la fonction en des nœuds choisis.

Polynôme d'interpolation de Lagrange d'une fonction

Cette dernière idée conduit à l'introduction de la définition suivante.

Définition 21 *Soit n un entier naturel, $x_0, \ldots, x_n$ $n+1$ nœuds distincts et f une fonction réelle donnée, définie en chacun des nœuds. On appelle* ***polynôme d'interpolation*** *(ou* ***interpolant) de Lagrange de degré*** *n* ***de la fonction*** *f, et on note $\Pi_n f$, le polynôme d'interpolation de Lagrange de degré n associé aux points $(x_i, f(x_i))_{i=0,\ldots,n}$.*

Exemple de polynôme d'interpolation de Lagrange d'une fonction. Construisons le polynôme d'interpolation de Lagrange de degré deux de la fonction $f(x) = e^x$ sur l'intervalle $[-1, 1]$, avec comme nœuds d'interpolation les points $x_0 = -1$, $x_1 = 0$ et $x_2 = 1$. Nous avons tout d'abord

$$l_0(x) = \frac{1}{2}x(x-1),\ l_1(x) = 1-x^2 \text{ et } l_2(x) = \frac{1}{2}x(x+1),$$

la forme de Lagrange du polynôme d'interpolation est donc la suivante

$$\Pi_2 f(x) = \frac{1}{2}x(x-1)e^{-1} + (1-x^2) + \frac{1}{2}x(x+1)e.$$

Pour la forme de de Newton de ce même polynôme d'interpolation, il vient

$$\omega_0(x) = 1,\ \omega_1(x) = (x+1) \text{ et } \omega_2(x) = (x+1)x,$$

ainsi que, en étendant quelque peu la notation utilisée pour les différences divisées,

$$[x_0]f = e^{-1},\ [x_0, x_1]f = 1-e^{-1} \text{ et } [x_0, x_1, x_2]f = \frac{1}{2}(e-2+e^{-1}) = \cosh(1)-1,$$

d'où

$$\Pi_2 f(x) = e^{-1} + (1-e^{-1})(x+1) + (\cosh(1)-1)(x+1)x.$$

On remarquera que $\Pi_2 f$ s'écrit encore

$$\Pi_2 f(x) = 1 + \sinh(1)\,x + (\cosh(1)-1)\,x^2$$

en utilisant les fonction polynomiales associées aux éléments de la base canonique de l'anneau des polynômes.

Erreur d'interpolation polynomiale

En termes de théorie de l'approximation, on peut voir le polynôme d'interpolation de Lagrange de la fonction f aux nœuds x_i, $i = 0, \ldots, n$, comme le polynôme de degré n minimisant l'*erreur d'approximation* $\|f - p_n\|$, $p_n \in \mathbb{P}_n$, mesurée avec la semi-norme

$$\|f\| = \sum_{i=0}^{n} |f(x_i)|.$$

Bien que les valeurs de f et de son polynôme d'interpolation coïncident aux nœuds d'interpolation, elles diffèrent en général en tout autre point et il convient donc d'étudier l'*erreur d'interpolation* $f - \Pi_n f$ sur l'intervalle auquel appartiennent les nœuds d'interpolation. En supposant la fonction f suffisamment régulière, on peut établir le résultat suivant, qui donne une estimation de cette différence.

Théorème 22 *Soit n un entier naturel, $[a,b]$ un intervalle non vide borné de $\mathbb{R}$, f une fonction de classe $\mathscr{C}^{n+1}$ sur $[a,b]$ et $n+1$ nœuds distincts $x_0,\ldots,x_n$, contenus dans l'intervalle $[a,b]$. Alors, pour tout réel x appartenant à $[a,b]$, il existe un point ξ dans I_x, le plus petit intervalle contenant $x_0,\ldots,x_n$ et x, tel que l'erreur d'interpolation au point x est donnée par*

$$f(x)-\Pi_n f(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\,\omega_{n+1}(x), \tag{2.17}$$

où ω_{n+1} est le polynôme de Newton de degré $n+1$ associé à la famille $\{x_i\}_{i=0,\ldots,n}$.

DÉMONSTRATION. Si le point x coïncide avec l'un des nœuds d'interpolation, les deux membres de (2.17) sont nuls et l'égalité est trivialement vérifiée. Supposons donc que x est un point distinct des nœuds $x_0,\ldots,x_n$ et introduisons la fonction auxiliaire

$$\forall t\in I_x,\ \varphi(t)=f(t)-\Pi_n f(t)-\frac{f(x)-\Pi_n f(x)}{\omega_{n+1}(x)}\,\omega_{n+1}(t).$$

Celle-ci est de classe $\mathscr{C}^{n+1}$ sur I_x (en vertu des hypothèses sur la fonction f) et s'annule en $n+2$ points (puisque $\varphi(x)=\varphi(x_0)=\cdots=\varphi(x_n)=0$). D'après le théorème de Rolle, la fonction φ' possède au moins $n+1$ zéros distincts dans l'intervalle I_x et, en raisonnant par récurrence, on en déduit que, pour tout entier naturel j inférieur ou égal à $n+1$, la dérivée $\varphi^{(j)}$ admet au moins $n+2-j$ zéros distincts. Par conséquent, il existe un réel ξ appartenant à I_x tel que $\varphi^{(n+1)}(\xi)=0$, ce qui s'écrit encore

$$f^{(n+1)}(\xi)-\frac{f(x)-\Pi_n f(x)}{\omega_{n+1}(x)}\,(n+1)!=0$$

et dont on déduit (2.17). □

En utilisant la continuité de $f^{(n+1)}$ et en considérant les bornes supérieures des valeurs absolues des deux membres de (2.17) sur l'intervalle $[a,b]$, on obtient comme corollaire immédiat

$$\|f-\Pi_n f\|_\infty\le\frac{1}{(n+1)!}\,\|f^{(n+1)}\|_\infty\|\omega_{n+1}\|_\infty\le\frac{1}{(n+1)!}\,\|f^{(n+1)}\|_\infty(b-a)^{n+1}. \tag{2.18}$$

La première majoration montre en particulier que l'amplitude de l'erreur d'interpolation dépend à la fois de la quantité $\|f^{(n+1)}\|_\infty$, qui peut être importante si la fonction f est très oscillante, et de la quantité $\|\omega_{n+1}\|_\infty$, dont la valeur est liée à la distribution des nœuds d'interpolation dans l'intervalle $[a,b]$.

Convergence des polynômes d'interpolation et contre-exemple de Runge

Intéressons-nous à la question de la convergence uniforme du polynôme d'interpolation d'une fonction vers cette dernière lorsque le nombre de nœuds d'interpolation tend vers l'infini. Comme ce polynôme dépend de la distribution des nœuds d'interpolation, il est nécessaire de formuler ce problème de manière plus précise. Nous supposerons ici que l'on fait le choix, particulièrement simple, d'une répartition *uniforme* des nœuds (on dit que les nœuds sont *équirépartis* ou encore *équidistribués*) sur un intervalle $[a,b]$ non vide de $\mathbb{R}$, en posant

$$\forall n\in\mathbb{N}^*,\ x_i=a+i\,\frac{(b-a)}{n},\ i=0,\ldots,n.$$

Au regard de l'estimation (2.18), il apparaît clairement que la convergence de la suite $(\Pi_n f)_{n\in\mathbb{N}^*}$ des polynômes d'interpolation d'une fonction f de classe $\mathscr{C}^\infty$ sur $[a,b]$ est liée au comportement de $\|f^{(n+1)}\|_\infty$ lorsque n augmente. En effet, si

$$\lim_{n\to+\infty}\frac{1}{(n+1)!}\,\|f^{(n+1)}\|_\infty\|\omega_{n+1}\|_\infty=0,$$

il vient immédiatement que

$$\lim_{n\to+\infty}\|f-\Pi_n f\|_\infty=0,$$

c'est-à-dire qu'on a convergence vers f, uniformément sur $[a,b]$, de la suite des polynômes d'interpolation de f associés à des nœuds équirépartis sur l'intervalle $[a,b]$ quand n tend vers l'infini.

Malheureusement, il existe des fonctions, que l'on qualifiera de « pathologiques », pour lesquelles le produit $\|f^{(n+1)}\|_\infty\|\omega_{n+1}\|_\infty$ tend vers l'infini *plus rapidement* que $(n+1)!$ lorsque n tend vers l'infini. Un exemple célèbre, dû à Runge, considère la convergence du polynôme d'interpolation de la fonction

$$f(x)=\frac{1}{1+x^2} \tag{2.19}$$

degré n	$\max_{x\in[-5,5]} \lvert f(x)-\Pi_n f(x)\rvert$
2	0,64623
4	0,43836
6	0,61695
8	1,04518
10	1,91566
12	3,66339
14	7,19488
16	14,39385
18	29,19058
20	59,82231
22	123,62439
24	257,21305

TABLE 2.1: Valeur (arrondie à la cinquième décimale) de l'erreur d'interpolation de Lagrange à nœuds équirépartis en norme de la convergence uniforme en fonction du degré d'interpolation pour la fonction de Runge $f(x)=\frac{1}{1+x^2}$ sur l'intervalle $[-5,5]$.

à nœuds équirépartis sur l'intervalle $[-5,5]$. Les valeurs du maximum de la valeur absolue de l'erreur d'interpolation pour cette fonction sont présentées dans la table 2.1 pour quelques valeurs paires du degré d'interpolation n. On observe une croissance exponentielle de l'erreur avec l'entier n. Le tracé des graphes de la fonction f et des polynômes d'interpolation associés à des nœuds équirépartis sur l'intervalle $[-5,5]$ met visuellement en évidence un phénomène de divergence de l'erreur d'interpolation au voisinage des extrémités de l'intervalle, parfois nommé *phénomène de Runge* (*Runge's phenomenon* en anglais).

Ce comportement de la suite des polynômes d'interpolation n'a rien à voir avec un éventuel « défaut » de régularité de la fonction que l'on interpole, qui est de classe $\mathscr{C}^\infty$ sur $\mathbb{R}$. Il est en revanche lié au fait que, vue comme une fonction d'une variable complexe, la fonction f, bien qu'analytique sur l'axe réel, possède deux pôles sur l'axe imaginaire en $z=\pm\mathrm{i}$.

D'autres choix de nœuds d'interpolation permettent néanmoins d'établir un résultat de convergence uniforme du polynôme d'interpolation dans ce cas. C'est, par exemple, le cas des points de Chebyshev (voir la table 2.2), donnés sur tout intervalle $[a,b]$ non vide de $\mathbb{R}$ par

$$\forall n\in\mathbb{N}^*,\ x_i=\frac{a+b}{2}+\frac{b-a}{2}\cos\left(\frac{2i+1}{2(n+1)}\pi\right),\ i=0,\dots,n.$$

2.2 Interpolation de Lagrange par morceaux

Jusqu'à présent, nous n'avons envisagé le problème de l'approximation d'une fonction f sur un intervalle $[a,b]$ par l'interpolation de Lagrange qu'en un sens *global*, c'est-à-dire en cherchant à n'utiliser qu'une seule expression analytique de l'interpolant (un seul polynôme) sur $[a,b]$. Pour obtenir une approximation que l'on espère plus précise, on n'a alors d'autre choix que d'augmenter le degré du polynôme d'interpolation. L'exemple de Runge évoqué dans la section précédente montre que la convergence uniforme de la suite $(\Pi_n f)_{n\in\mathbb{N}}$ vers f n'est cependant pas garantie pour toute distribution arbitraire des nœuds d'interpolation.

Une alternative à cette première approche est de construire une partition de l'intervalle $[a,b]$ en sous-intervalles sur chacun desquels une interpolation polynomiale de bas degré est employée. On parle alors d'*interpolation polynomiale par morceaux*. L'idée naturelle suivie est que toute fonction peut être approchée de manière arbitrairement précise par des polynômes de bas degré (un ou même zéro par exemple), de manière à limiter les phénomènes d'oscillations observés avec l'interpolation de haut degré, sur des intervalles *suffisamment petits*.

Désignons par $[a,b]$ un intervalle non vide borné de $\mathbb{R}$ et par f une application de $[a,b]$ dans $\mathbb{R}$. On considère également un entier naturel non nul m et $m+1$ points $x_0,\dots,x_m$, tels que $a=x_0<x_1<\cdots<x_m=b$, réalisant une partition $\mathscr{T}_h$ de $[a,b]$ en m sous-intervalles $[x_{j-1},x_j]$ de longueur $h_j=x_j-x_{j-1}$, $1\leq j\leq m$, dont on

degré n	$\max_{x\in[-5,5]} \lvert f(x)-\Pi_n f(x)\rvert$
2	0,60060
4	0,20170
6	0,15602
8	0,17083
10	0,10915
12	0,06921
14	0,04660
16	0,03261
18	0,02249
20	0,01533
22	0,01036
24	0,00695

TABLE 2.2: Valeur (arrondie à la cinquième décimale) de l'erreur d'interpolation de Lagrange utilisant les points de Chebyshev en norme de la convergence uniforme en fonction du degré d'interpolation pour la fonction de Runge $f(x)=\frac{1}{1+x^2}$ sur l'intervalle $[-5,5]$.

caractérise la « *finesse* » par

$$h=\max_{1\leq j\leq m} h_j.$$

L'interpolation de Lagrange par morceaux d'une fonction f donnée relativement à une partition $\mathscr{T}_h$ d'un intervalle $[a,b]$ consiste en la construction d'un *polynôme d'interpolation par morceaux* coïncidant sur chacun des sous-intervalles $[x_{j-1},x_j]$, $j=1,\ldots,m$, de $\mathscr{T}_h$ avec le polynôme d'interpolation de Lagrange de f en des nœuds fixés de ce sous-intervalle. La fonction interpolante ainsi obtenue est, en général, simplement continue sur $[a,b]$.

On notera qu'on peut *a priori* choisir un polynôme d'interpolation de degré différent sur chaque sous-intervalle (il en va de même la répartition des nœuds lui correspondant). Cependant, en pratique, on utilise très souvent la même interpolation, de bas degré, sur tous les sous-intervalles pour des raisons de commodité. Dans ce cas, on note $\Pi_h^n f$ le polynôme d'interpolation par morceaux obtenu en considérant sur chaque sous-intervalle $[x_{j-1},x_j]$, $j=1,\ldots,m$, d'une partition $\mathscr{T}_h$ de $[a,b]$ une interpolation de Lagrange de f aux nœuds $x_j^{(0)},\ldots,x_j^{(n)}$, par exemple équirépartis, avec n un entier naturel non nul et « petit ». Puisque la restriction de $\Pi_h^n f$ à chaque sous-intervalle $[x_{j-1},x_j]$, $j=1,\ldots,m$, est le polynôme d'interpolation de Lagrange de f de degré n associé aux nœuds $x_j^{(0)},\ldots,x_j^{(n)}$, on déduit aisément, si f est de classe $\mathscr{C}^{n+1}$ sur $[a,b]$, du théorème 22 une majoration de l'erreur d'interpolation $\left\lvert f(x)-\Pi_h^n f(x)\right\rvert$ sur chaque sous-intervalle de $\mathscr{T}_h$, conduisant à une estimation d'erreur globale de la forme

$$\|f-\Pi_h^n f\|_\infty \leq C\,h^{n+1}\|f^{(n+1)}\|_\infty,$$

avec C une constante strictement positive dépendant de l'entier n. On observe alors qu'on peut rendre arbitrairement petite l'erreur d'interpolation dès lors que la partition $\mathscr{T}_h$ de $[a,b]$ est suffisamment fine (*i.e.*, le réel h est suffisamment petit).

Chapitre 3

Formules de quadrature

L'évaluation d'une intégrale définie de la forme

$$I(f\,;a,b)=\int_a^b f(x)\,\mathrm{d}x,$$

où $[a,b]$ est un intervalle non vide et borné de $\mathbb{R}$ et f est une fonction d'une variable réelle, continue sur $[a,b]$, à valeurs réelles, est un problème classique. Nous nous intéressons dans le présent chapitre à l'utilisation de *formules de quadrature* (*quadrature rules* en anglais) qui approchent la valeur de l'intégrale par une somme pondérée finie de valeurs de la fonction f en des points choisis. Les formules que nous considérons pour ce faire sont essentiellement de la forme

$$I_n(f\,;a,b)=\sum_{i=0}^{n}\alpha_i\,f(x_i), \tag{3.1}$$

où n est un entier naturel, les coefficients $\{\alpha_i\}_{i=0,\ldots,n}$ sont réels et les points $\{x_i\}_{i=0,\ldots,n}$ appartiennent à l'intervalle $[a,b]$.

Les raisons conduisant à une évaluation seulement approchée d'une intégrale comme $I(f)$ sont variées. Tout d'abord, si l'on a, en vertu du théorème fondamental de l'analyse, que

$$I(f\,;a,b)=F(b)-F(a),$$

où la fonction F est une primitive de f, on ne sait pas toujours, même en ayant recours à des techniques plus ou moins sophistiquées telles que le changement de variable ou l'intégration par parties, exprimer F en termes de fonctions algébriques, trigonométriques, exponentielles ou logarithmiques. Lorsque c'est toutefois le cas, l'évaluation numérique d'une telle intégrale peut encore s'avérer difficile et coûteuse en pratique, tout en n'étant réalisée qu'avec une certaine exactitude en arithmétique en précision finie et donc, au final, approchée. Enfin, le recours à une évaluation approchée est obligatoire lorsque l'intégrande est solution d'une équation fonctionnelle (une équation différentielle par exemple) que l'on ne sait pas explicitement résoudre.

3.1 Formules de quadrature interpolatoires

En dehors d'une première sous-section générale sur les *formules de quadrature interpolatoires* (*interpolatory quadrature rules* en anglais), nous limitons dans l'exposé aux *formules de Newton–Cotes*, qui sont un cas particulier de formules de quadrature basées sur l'interpolation de Lagrange introduite dans le précédent chapitre.

3.1.1 Généralités

On notera tout d'abord que les points x_i et les coefficients α_i, $i=0,\ldots,n$, dans l'expression (3.1) sont respectivement appelés *nœuds* (*nodes* en anglais) et *poids* (*weights* en anglais) de la formule de quadrature. Les nœuds de quadrature appartiennent en général à l'intervalle d'intégration $[a,b]$, mais cela n'est pas toujours le cas.

Comme pour les problèmes d'interpolation étudiés au chapitre précédent, la précision d'une formule de quadrature pour une fonction f continue sur l'intervalle $[a,b]$ donnée se mesure notamment en évaluant l'*erreur de quadrature*

$$E_n(f;a,b) = I(f;a,b) - I_n(f;a,b). \tag{3.2}$$

On définit par ailleurs le *degré (algébrique) d'exactitude*[1] (*(algebraic) degree of exactness* en anglais) d'une formule de quadrature $I_n(f)$ comme le plus grand entier $r \geq 0$ pour lequel

$$\forall m \in \{0,\dots,r\},\ \forall f \in \mathbb{P}_m,\ I(f;a,b) = I_n(f;a,b).$$

Ce degré ne dépend de l'intervalle d'intégration sur lequel la formule de quadrature est considérée.

Enfin, une formule de quadrature interpolatoire est obtenue en remplaçant la fonction f dans l'intégrale par son polynôme d'interpolation. Dans le cas du polynôme d'interpolation de Lagrange (voir le précédent chapitre), on pose ainsi

$$I_n(f;a,b) = \int_a^b \Pi_n f(x)\,\mathrm{d}x, \tag{3.3}$$

où $\Pi_n f$ désigne le polynôme d'interpolation de Lagrange de f associé à un ensemble de nœuds $\{x_i\}_{i=0,\dots,n}$ donné. En vertu de la définition du polynôme d'interpolation de Lagrange d'une fonction, on a alors, par la propriété de linéarité de l'intégrale,

$$I_n(f;a,b) = \int_a^b \left(\sum_{i=0}^n f(x_i)\,l_i(x)\right)\mathrm{d}x = \sum_{i=0}^n f(x_i)\int_a^b l_i(x)\,\mathrm{d}x,$$

et, en identifiant avec (3.1), on trouve que les poids de quadrature sont simplement les intégrales respectives des polynômes de Lagrange $\{l_i\}_{i=0,\dots,n}$ sur l'intervalle $[a,b]$, c'est-à-dire

$$\alpha_i = \int_a^b l_i(x)\,\mathrm{d}x,\ i = 0,\dots,n. \tag{3.4}$$

Pour certaines familles de formules, il arrive que les valeurs des nœuds et poids de quadrature soient données relativement à l'intervalle d'intégration « standard » $[-1,1]$. Dans ce cas, on utilise le changement de variable affine

$$t \mapsto \frac{1}{2}(b-a)t + \frac{1}{2}(a+b),$$

pour obtenir les valeurs des nœuds correspondant à un intervalle borné $[a,b]$ arbitraire, et l'on multiplie les valeurs des poids par $\frac{1}{2}(b-a)$ en vertu de la formule de changement de variable dans une intégrale définie.

On a la caractérisation suivante pour les formules de quadrature interpolatoires de la forme (3.1).

Théorème 23 *Soit n un entier naturel. Toute formule de quadrature utilisant $n+1$ nœuds distincts et donnée par* (3.1) *est interpolatoire si et seulement si son degré d'exactitude au moins égal à n.*

DÉMONSTRATION. Soit un intervalle non vide $[a,b]$ de $\mathbb{R}$. Pour toute formule de quadrature interpolatoire à $n+1$ nœuds distincts x_i, $i = 0,\dots,n$, on déduit d'une égalité précédemment obtenue pour l'erreur d'interpolation ponctuelle que, pour toute fonction polynomiale f de degré inférieur ou égal à n, l'erreur de quadrature $E_n(f;a,b)$ est nulle.

Réciproquement, si le degré d'exactitude de la formule de quadrature est au moins égal à n, les poids de quadrature α_i, $i = 0,\dots,n$, doivent vérifier les relations

$$\begin{aligned} \sum_{i=0}^n \alpha_i &= b-a, \\ \sum_{i=0}^n \alpha_i\, x_i &= \frac{1}{2}(b^2-a^2), \\ &\vdots \\ \sum_{i=0}^n \alpha_i\, {x_i}^n &= \frac{1}{n+1}(b^{n+1}-a^{n+1}), \end{aligned} \tag{3.5}$$

1. On trouve parfois aussi le terme de *degré (algébrique) de précision* (*(algebraic) degree of precision* en anglais).

qui constituent un système linéaire de $n+1$ équations à $n+1$ inconnues admettant une unique solution (le déterminant qui lui est associé étant de Vandermonde et les nœuds x_i, $i=0,\dots,n$, étant supposés distincts). On remarque alors que la formule de quadrature de Lagrange est par définition exacte pour toute fonction polynomiale de degré inférieur ou égal à n, et plus particulièrement pour tout polynôme d'interpolation de degré n associé aux nœuds x_i, $i=0,\dots,n$. Le choix des poids de quadrature de Lagrange satisfait donc chacune des équations de (3.5). □

3.1.2 Formules de Newton–Cotes

Les formules de quadrature de Newton–Cotes (*Newton–Cotes formulas* ou *Newton–Cotes quadrature rules* en anglais) sont basées sur l'interpolation de Lagrange à nœuds *équirépartis* dans l'intervalle $[a,b]$; ce sont donc des cas particuliers de formules de quadrature interpolatoires de Lagrange. Pour n un entier positif fixé, notons $x_i = x_0 + ih$, $i=0,\dots,n$, les nœuds de quadrature. On peut définir deux types de formules de Newton–Cotes :

- les formules *fermées* (*closed formulae* en anglais), pour lesquelles les extrémités de l'intervalle $[a,b]$ font partie des nœuds, c'est-à-dire $x_0=a$, $x_n=b$ et $h=\frac{b-a}{n}$ $(n\geq 1)$, et dont les règles bien connues *du trapèze* $(n=1)$ et *de Simpson* $(n=2)$ sont des cas particuliers,
- les formules *ouvertes* (*open formulae* en anglais), pour lesquelles $x_0=a+h$, $x_n=b-h$ et $h=\frac{b-a}{n+2}$ $(n\geq 0)$, auxquelles appartient la *règle du point milieu* $(n=0)$.

Une propriété intéressante de ces formules est que leurs poids de quadrature ne dépendent explicitement que de n et h et non de l'intervalle d'intégration $[a,b]$; ceux-ci peuvent donc être calculés *a priori*. En effet, en introduisant, dans le cas des formules fermées, le changement de variable

$$x = x_0 + th = a + th, \text{ avec } t\in[0,n],$$

on vérifie que, pour tout $(i,j)\in\{0,\dots,n\}^2$, $i\neq j$, $n\geq 1$,

$$\forall t\in[0,n],\ \frac{x-x_j}{x_i-x_j} = \frac{a+th-(a+jh)}{a+ih-(a+jh)} = \frac{t-j}{i-j},$$

et donc

$$\forall i\in\{0,\dots,n\},\ l_i(x) = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j}{i-j},$$

d'où l'expression suivante pour les poids de quadrature

$$\forall i\in\{0,\dots,n\},\ \alpha_i = \int_a^b l_i(x)\,\mathrm{d}x = h\int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j}{i-j}\,\mathrm{d}t.$$

On obtient ainsi que

$$I_n(f;a,b) = h\sum_{i=0}^{n} w_i f(x_i),\ \text{avec } w_i = \int_0^n \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j}{i-j}\,\mathrm{d}t.$$

En procédant de manière analogue pour les formules ouvertes, on trouve que

$$I_n(f;a,b) = h\sum_{i=0}^{n} w_i f(x_i),\ \text{avec } w_i = \int_{-1}^{n+1} \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{t-j}{i-j}\,\mathrm{d}t,$$

en posant $x_{-1} = x_0 - h = a$ et $x_{n+1} = x_0 + (n+1)h = b$. Dans le cas particulier où $n=0$, on a $w_0 = 2$ puisque $l_0(x)=1$.

Ajoutons par ailleurs, en vertu d'une propriété de symétrie des polynômes de Lagrange, que les poids w_i et w_{n-i} sont égaux pour $i=0,\dots,n$ pour les formules fermées $(n\geq 1)$ et ouvertes $(n\geq 0)$. Pour cette raison, on ne tabule les valeurs des poids que pour $0\leq i\leq \lfloor \frac{n}{2} \rfloor$ (voir la table 3.1), où $\lfloor \frac{n}{2} \rfloor$ désigne la partie entière de $\frac{n}{2}$.

n	w_0	w_1	w_2	w_3	
1	$\frac{1}{2}$				(règle du trapèze)
2	$\frac{1}{3}$	$\frac{4}{3}$			(règle de Simpson)
3	$\frac{3}{8}$	$\frac{9}{8}$			(règle des trois huitièmes)
4	$\frac{14}{45}$	$\frac{64}{45}$	$\frac{24}{45}$		(règle de Boole)
5	$\frac{95}{288}$	$\frac{375}{288}$	$\frac{125}{144}$		
6	$\frac{3}{10}$	$\frac{3}{2}$	$\frac{3}{10}$	$\frac{9}{5}$	(règle de Weddle)

n	w_0	w_1	w_2	
0	2			(règle du point milieu)
1	$\frac{3}{2}$			
2	$\frac{8}{3}$	$-\frac{4}{3}$		
3	$\frac{55}{24}$	$\frac{5}{24}$		
4	$\frac{33}{10}$	$-\frac{21}{5}$	$\frac{39}{5}$	

TABLE 3.1: Poids des formules de Newton–Cotes fermées (à gauche) et ouvertes (à droite) pour quelques valeurs de l'entier n.

On remarque la présence de poids négatifs pour certaines formules ouvertes[2], ce qui peut conduire à des instabilités dues aux erreurs d'annulation lors de l'évaluation numérique de la formule. La convergence de la suite des intégrales approchées par une formule de Newton-Cotes à $n+1$ nœuds n'est par ailleurs pas assurée lorsque n tend vers l'infini, même si l'intégrande est une fonction analytique sur l'intervalle d'intégration (ce comportement est à relier à la divergence de l'interpolation de Lagrange avec nœuds équirépartis pour la fonction du contre-exemple de Runge). Pour ces deux raisons, l'utilisation des formules de Newton–Cotes (fermées ou ouvertes) utilisant plus de huit nœuds reste délicate et est généralement déconseillée en pratique. Ainsi, si l'on souhaite améliorer la précision de l'approximation d'une intégrale obtenue par une formule de quadrature de Newton–Cotes donnée, on fera plutôt appel à des formules composées ou encore aux *formules de Gauss*.

Présentons maintenant de quelques exemples importants de formules de quadrature de Newton–Cotes.

Règle du point milieu (*midpoint rule* en anglais). Cette formule, aussi appelée *règle du rectangle*[3] (*rectangle rule* en anglais), est obtenue en remplaçant dans l'intégrale la fonction f par la valeur qu'elle prend au milieu de l'intervalle $[a,b]$ (voir la figure 3.1), d'où

$$I_0(f;a,b) = (b-a)f\left(\frac{a+b}{2}\right). \tag{3.8}$$

Le poids de quadrature vaut donc $\alpha_0 = b-a$ et le nœud est $x_0 = \dfrac{a+b}{2}$.

En supposant la fonction f de classe $\mathscr{C}^2$ sur $[a,b]$, on peut utiliser le théorème 24 pour montrer que l'erreur de quadrature de cette formule vaut

$$E_0(f;a,b) = \frac{(b-a)^3}{24} f''(\eta), \text{ avec } \eta \in]a,b[.$$

Son degré d'exactitude est par conséquent égal à un.

Règle du trapèze (*trapezoidal rule* en anglais). On obtient cette formule en remplaçant dans l'intégrale la fonction f par son polynôme d'interpolation de Lagrange de degré un aux points a et b (voir la figure 3.2). Il

2. La formule fermée à neuf nœuds ainsi que toutes les formules fermées à plus de onze points présentent elles aussi au moins un poids négatif.

3. D'autres formules de quadrature interpolatoires sont basées sur un polynôme d'interpolation de Lagrange de degré nul (et donc un seul nœud de quadrature) : ce sont les règles du *rectangle à gauche*

$$I_0(f;a,b) = (b-a)f(a), \tag{3.6}$$

et *du rectangle à droite*

$$I_0(f;a,b) = (b-a)f(b), \tag{3.7}$$

dont le degré d'exactitude est égal à zéro. Elles ne font cependant pas partie des formules de Newton–Cotes.

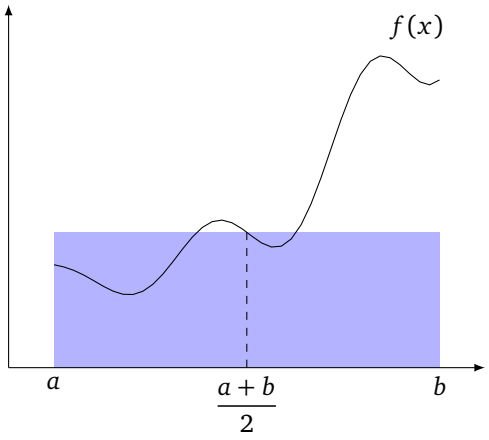

FIGURE 3.1: Illustration de la règle du point milieu. La valeur approchée de l'intégrale $I(f;a,b)$ correspond à l'aire colorée en bleu.

vient alors

$$I_1(f;a,b) = \frac{b-a}{2}\left(f(a)+f(b)\right). \tag{3.9}$$

Les poids de quadrature valent $\alpha_0 = \alpha_1 = \dfrac{b-a}{2}$, tandis que les nœuds sont $x_0 = a$ et $x_1 = b$.

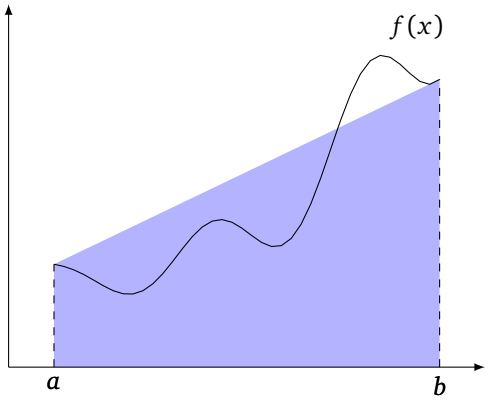

FIGURE 3.2: Illustration de la règle du trapèze. La valeur approchée de l'intégrale $I(f;a,b)$ correspond à l'aire colorée en bleu.

En supposant f de classe $\mathscr{C}^2$ sur $[a,b]$, on obtient la valeur suivante pour l'erreur de quadrature

$$E_1(f;a,b) = -\frac{(b-a)^3}{12} f''(\xi), \text{ avec } \xi \in]a,b[.$$

et l'on en déduit que cette formule à un degré d'exactitude égal à un.

Règle de Simpson (*Simpson's rule* en anglais). Cette dernière formule est obtenue en substituant dans l'intégrale à la fonction f son polynôme d'interpolation de Lagrange de degré deux aux nœuds $x_0 = a$, $x_1 = \dfrac{a+b}{2}$ et $x_2 = b$ (voir la figure 3.3) et s'écrit

$$I_2(f;a,b) = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right). \tag{3.10}$$

Les poids de quadrature sont donnés par $\alpha_0 = \alpha_2 = \dfrac{b-a}{6}$ et $\alpha_1 = 2\dfrac{b-a}{3}$.

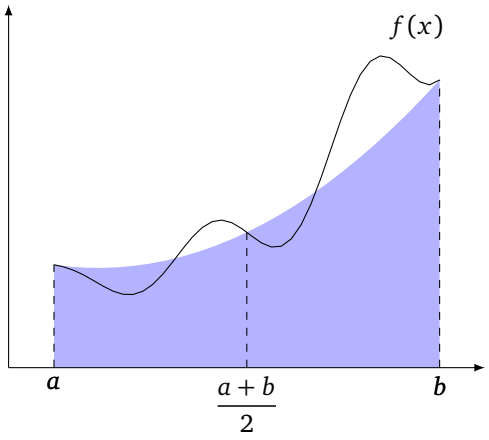

FIGURE 3.3: Illustration de la règle de Simpson. La valeur approchée de l'intégrale $I(f;a,b)$ correspond à l'aire colorée en bleu.

On montre, grâce au théorème 24, que, si la fonction f est de classe $\mathscr{C}^4$ sur l'intervalle $[a,b]$, l'erreur de quadrature peut s'écrire

$$E_2(f;a,b) = -\frac{(b-a)^5}{2880} f^{(4)}(\xi), \text{ avec } \xi \in]a,b[. \tag{3.11}$$

Cette formule a donc un degré d'exactitude égal à trois.

3.1.3 Estimations d'erreur

Cette section est consacrée à l'établissement d'expressions permettant d'arriver à une estimation de l'erreur d'une formule de Newton–Cotes. Le résultat suivant, dont la preuve technique est admise, fournit les estimations des erreurs de quadrature des règles du point milieu, du trapèze et de Simpson annoncées plus haut.

Théorème 24 *Soit $[a,b]$ un intervalle non vide et borné de $\mathbb{R}$, n un entier positif et f une fonction de $\mathscr{C}^{n+2}([a,b])$ si n est pair, de $\mathscr{C}^{n+1}([a,b])$ si n est impair. Alors, l'erreur de quadrature pour les formules de Newton–Cotes fermées est donnée par*

$$E_n(f;a,b) = \begin{cases} \dfrac{K_n}{(n+2)!} f^{(n+2)}(\xi),\ K_n = \displaystyle\int_a^b x\,\omega_{n+1}(x)\,\mathrm{d}x < 0, & \text{si } n \text{ est pair,} \\ \dfrac{K_n}{(n+1)!} f^{(n+1)}(\xi),\ K_n = \displaystyle\int_a^b \omega_{n+1}(x)\,\mathrm{d}x < 0, & \text{si } n \text{ est impair,} \end{cases}$$

avec $a < \xi < b$, et par

$$E_n(f;a,b) = \begin{cases} \dfrac{K'_n}{(n+2)!} f^{(n+2)}(\eta),\ K'_n = \displaystyle\int_a^b x\,\omega_{n+1}(x)\,\mathrm{d}x > 0, & \text{si } n \text{ est pair,} \\ \dfrac{K'_n}{(n+1)!} f^{(n+1)}(\eta),\ K'_n = \displaystyle\int_a^b \omega_{n+1}(x)\,\mathrm{d}x > 0, & \text{si } n \text{ est impair,} \end{cases}$$

avec $a < \eta < b$, pour les formules de Newton–Cotes ouvertes.

Ce théorème montre que le degré d'exactitude d'une formule de Newton–Cotes à $n+1$ nœuds est égal à $n+1$ lorsque n est pair et n lorsque n est impair, que la formule soit fermée ou ouverte. Il est donc en général préférable d'employer une formule avec un nombre *impair* de nœuds.

Notons que l'on peut également chercher à faire apparaître la dépendance de l'erreur de quadrature par rapport au pas h et utilisant le changement de variable $x = x_0 + th$. On obtient ainsi facilement le résultat suivant.

Corollaire 25 *Sous les hypothèses du théorème 24, on a les expressions suivantes pour les erreurs de quadrature respectives des formules de Newton–Cotes fermées et ouvertes*

$$E_n(f;a,b)=\begin{cases}\dfrac{M_n}{(n+2)!}\,h^{n+3}f^{(n+2)}(\xi),\ M_n=\displaystyle\int_0^n t\,\pi_{n+1}(t)\,\mathrm{d}t<0, & \textit{si } n \textit{ est pair,}\\ \dfrac{M_n}{(n+1)!}\,h^{n+2}f^{(n+1)}(\xi),\ M_n=\displaystyle\int_0^n \pi_{n+1}(t)\,\mathrm{d}t<0, & \textit{si } n \textit{ est impair,}\end{cases}$$

avec $a<\xi<b$,

$$E_n(f;a,b)=\begin{cases}\dfrac{M_n'}{(n+2)!}\,h^{n+3}f^{(n+2)}(\eta),\ M_n'=\displaystyle\int_{-1}^{n+1} t\,\pi_{n+1}(t)\,\mathrm{d}t>0, & \textit{si } n \textit{ est pair,}\\ \dfrac{M_n'}{(n+1)!}\,h^{n+2}f^{(n+1)}(\eta),\ M_n'=\displaystyle\int_{-1}^{n+1} \pi_{n+1}(t)\,\mathrm{d}t>0, & \textit{si } n \textit{ est impair,}\end{cases}$$

avec $a<\eta<b$.

3.2 Formules de quadrature composées

Les formules de quadrature introduites jusqu'à présent ont toutes été obtenues en substituant à l'intégrande son polynôme d'interpolation de Lagrange à nœuds équirépartis sur l'intervalle d'intégration, la valeur de l'intégrale considérée étant alors approchée par la valeur de l'intégrale du polynôme. Pour améliorer la précision de cette approximation, on est donc tenté d'augmenter le degré de l'interpolation polynomiale utilisée. Le phénomène de Runge (voir le précédent chapitre) montre cependant qu'un polynôme d'interpolation de degré élevé peut, lorsque les nœuds d'interpolation équirépartis, fournir une approximation catastrophique d'une fonction pourtant très régulière, ce qui a des conséquences désastreuses lorsque l'on cherche, par exemple, à approcher l'intégrale

$$\int_{-5}^{5}\frac{\mathrm{d}x}{1+x^2}=2\,\arctan(5)$$

par une formule de Newton–Cotes. Il a d'ailleurs été démontré que les formules de Newton–Cotes ne convergent généralement pas lorsque l'entier n tend vers l'infini, même lorsque la fonction à intégrer est analytique. Ceci, allié à l'observation que les poids de quadrature n'ont pas tous le même signe à partir de $n=2$ pour les formules ouvertes et $n=8$ pour les formules fermées, ce qui qui pose des problèmes de stabilité numérique, conduit les praticiens à ne pas, ou peu, utiliser les formules de quadrature de Newton–Cotes dont le nombre de nœuds est supérieur ou égal à huit.

Il est cependant possible construire très simplement des formules de quadrature dont la mise en œuvre est aisée et dont la précision pourra être aussi grande que souhaitée : ce sont les *formules de quadrature composées* (*composite quadrature rules* ou *compound quadrature rules* en anglais). Celles-ci utilisent la technique de l'interpolation polynomiale par morceaux introduite dans le précédent chapitre, qui consiste une interpolation polynomiale à nœuds équirépartis de bas degré sur des sous-intervalles obtenus en partitionnant l'intervalle d'intégration. On peut construire de nombreuses classes de formules de quadrature interpolatoires composées, mais nous ne présentons ici que les plus courantes, en lien avec les formules de Newton–Cotes qui viennent d'être étudiées.

Étant donné un entier m supérieur ou égal a 1, on pose

$$H=\frac{b-a}{m}$$

et l'on introduit une partition de l'intervalle $[a,b]$ en m sous-intervalles $[x_{j-1},x_j]$, $j=1,\ldots,m$, de longueur H, avec $x_i=a+iH$, $i=0,\ldots,m$. Comme

$$I(f;a,b)=\int_a^b f(x)\,\mathrm{d}x=\sum_{j=1}^{m}\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x,$$

il suffit d'approcher chacune des intégrales apparaissant dans le membre de droite de l'égalité ci-dessus en utilisant une formule de quadrature interpolatoire, généralement la même sur chaque sous-intervalle, pour obtenir une formule composée, conduisant à une approximation de $I(f;a,b)$ de la forme

$$I_{m,n}(f;a,b)=\sum_{j=1}^{m} I_n(f;x_{j-1},x_j)=\sum_{j=1}^{m}\sum_{i=0}^{n}\alpha_{i,j}\,f(x_{i,j}), \tag{3.12}$$

où les coefficients $\alpha_{i,j}$ et les points $x_{i,j}$, $i=0,\dots,n$, désignent respectivement les poids et les nœuds de la formule de quadrature interpolatoire utilisée sur le j^{e} sous-intervalle, $j=1,\dots,m$, de la partition de $[a,b]$. Dans les *formules de Newton–Cotes composées*, la formule de quadrature utilisée sur chaque sous-intervalle est une même formule de Newton–Cotes, fermée ou ouverte, à $n+1$ nœuds équirépartis, $n\geq 0$, et l'on a par conséquent

$$x_{i,j}=x_{j-1}+i\,h,\ i=0,\dots,n,\ j=1,\dots,m,$$

avec $h=\frac{H}{n}$, $n\geq 1$, pour une formule fermée, et $h=\frac{H}{n+2}$, $n\geq 0$, pour une formule ouverte, les poids de quadrature $\alpha_{i,j}=h\,w_i$ étant indépendants de j.

En notant que l'erreur de quadrature d'une formule composée, notée $E_{m,n}(f;a,b)$, se décompose de la manière suivante

$$E_{m,n}(f;a,b)=I(f;a,b)-I_{m,n}(f;a,b)=\sum_{j=1}^{m}\left(\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x-\sum_{i=0}^{n}\alpha_{i,j}\,f(x_{i,j})\right),$$

on obtient sans mal, grâce à l'analyse d'erreur réalisée dans la précédente section, le résultat suivant pour les formules de Newton–Cotes composées.

Théorème 26 *Soit $[a,b]$ un intervalle non vide et borné de $\mathbb{R}$, n un entier positif et f une fonction de $\mathscr{C}^{n+2}([a,b])$ si n est pair, de $\mathscr{C}^{n+1}([a,b])$ si n est impair. Alors, en conservant les notations du corollaire 25, l'erreur de quadrature pour les formules de Newton–Cotes composées vaut*

$$E_{m,n}(f;a,b)=\begin{cases}\dfrac{M_n}{(n+2)!}\dfrac{b-a}{n^{n+3}}\,H^{n+2}f^{(n+2)}(\xi) & \text{si } n \text{ est pair,}\\[2ex] \dfrac{M_n}{(n+1)!}\dfrac{b-a}{n^{n+2}}\,H^{n+1}f^{(n+1)}(\xi) & \text{si } n \text{ est impair,}\end{cases}$$

avec $a<\xi<b$, pour les formules fermées et

$$E_{m,n}(f;a,b)=\begin{cases}\dfrac{M'_n}{(n+2)!}\dfrac{b-a}{(n+2)^{n+3}}\,H^{n+2}f^{(n+2)}(\eta) & \text{si } n \text{ est pair,}\\[2ex] \dfrac{M'_n}{(n+1)!}\dfrac{b-a}{(n+2)^{n+2}}\,H^{n+1}f^{(n+1)}(\eta) & \text{si } n \text{ est impair,}\end{cases}$$

avec $a<\eta<b$.

On déduit de ce théorème que, à n fixé, l'erreur de quadrature d'une formule de Newton–Cotes composée tend vers 0 lorsque m tend vers l'infini, c'est-à-dire lorsque H tend vers 0, ce qui assure la convergence de la valeur approchée de l'intégrale vers sa valeur exacte. De plus, le degré d'exactitude d'une formule composée coïncide avec celui de la formule dont elle dérive. En pratique, on utilise généralement des formules de Newton–Cotes composées basées sur des formules à peu de nœuds ($n\leq 2$), ce qui garantit que tous les poids de quadrature sont positifs. À cet égard, les formules dans l'exemple qui suit sont très couramment employées.

Exemples de formules de Newton–Cotes composées. On présente, avec leur erreur de quadrature (obtenue via le théorème 26) ci-dessous trois formules Newton–Cotes composées parmi les plus utilisées. La *règle du point milieu composée* (voir la figure 3.4) fait partie des formules ouvertes et ne possède qu'un nœud de quadrature dans chaque sous-intervalle de la partition de l'intervalle $[a,b]$. Dans ce cas, on a $h=\frac{H}{2}$ et

$$I(f;a,b)=H\sum_{j=1}^{m} f\left(a+\left(j-\frac{1}{2}\right)H\right)+\frac{b-a}{24}\,H^2 f''(\eta),$$

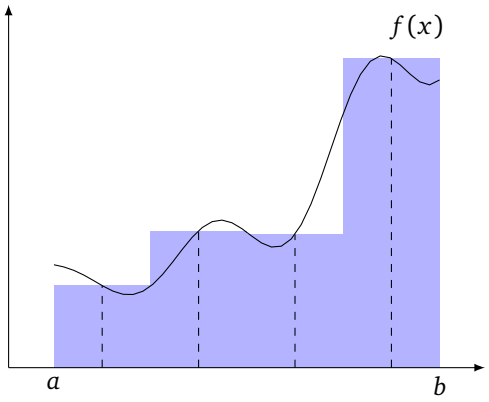

FIGURE 3.4: Illustration de la règle du point milieu composée à quatre sous-intervalles sur l'intervalle $[a,b]$. La valeur approchée de l'intégrale $I(f;a,b)$ correspond à l'aire colorée en bleu.

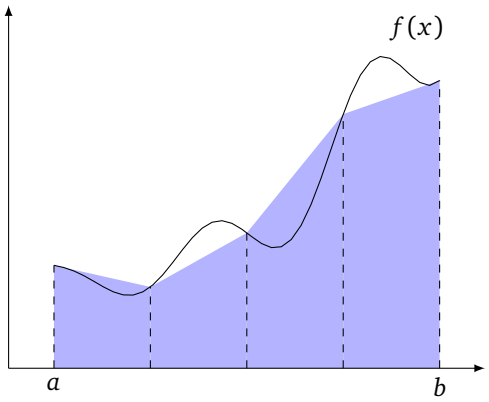

FIGURE 3.5: Illustration de la règle du trapèze composée à quatre sous-intervalles sur l'intervalle $[a,b]$. La valeur approchée de l'intégrale $I(f;a,b)$ correspond à l'aire colorée en bleu.

avec $\eta \in]a,b[$.

La *règle du trapèze composée* (voir la figure 3.5) est une formule fermée qui a pour nœuds de quadrature les extrémités de chaque sous-intervalle. On a alors $h = H$ et

$$I(f;a,b) = \frac{H}{2}\left(f(a) + 2\sum_{j=1}^{m-1} f(a+jH) + f(b)\right) - \frac{b-a}{12} H^2 f''(\xi),$$

avec $\xi \in]a,b[$. Enfin, la *règle de Simpson composée* (voir la figure 3.6) utilise comme nœuds de quadrature les extrémités et le milieu de chaque sous-intervalle, d'où $h = \frac{H}{2}$ et

$$I(f;a,b) = \frac{H}{6}\left(f(a) + 2\sum_{j=1}^{m-1} f(a+jH) + 4\sum_{j=1}^{m} f\left(a+\left(j-\frac{1}{2}\right)H\right) + f(b)\right) - \frac{b-a}{2880} H^4 f^{(4)}(\xi),$$

avec, là encore, $\xi \in]a,b[$.

On peut établir la convergence d'une formule de quadrature composée sous des hypothèses bien moins restrictives que celles du théorème 26. C'est l'objet du résultat suivant, dont la preuve est admise.

Théorème 27 *Soit $[a,b]$ un intervalle non vide et borné de $\mathbb{R}$, f une fonction continue sur $[a,b]$, $\{x_j\}_{j=0,\ldots,m}$ l'ensemble des nœuds d'une partition de $[a,b]$ en m sous-intervalles et une formule de quadrature composée de la forme* (3.12) *relativement à cette partition, de degré exactitude égal à r et dont les poids de quadrature $\alpha_{i,j}$, $i = 0,\ldots,n$, $j = 1,\ldots,m$, sont positifs. Alors, on a*

$$\lim_{m\to+\infty} I_{m,n}(f;a,b) = I(f;a,b).$$

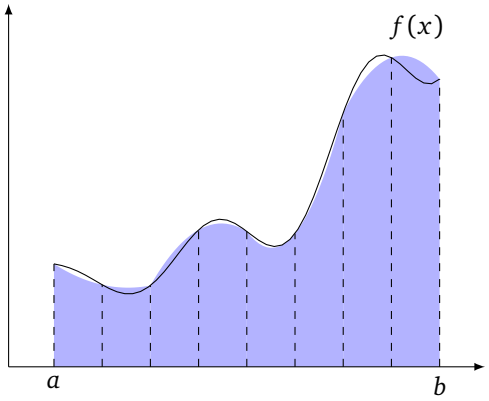

FIGURE 3.6: Illustration de la règle de Simpson composée à quatre sous-intervalles sur l'intervalle $[a, b]$. La valeur approchée de l'intégrale $I(f\,; a, b)$ correspond à l'aire colorée en bleu.

Chapitre 4

Méthodes directes de résolution des systèmes linéaires

On considère la résolution du système linéaire

$$A\boldsymbol{x} = \boldsymbol{b}, \tag{4.1}$$

avec A une matrice d'ordre n à coefficients réels inversible et $\boldsymbol{b}$ une matrice colonne de $M_{n,1}(\mathbb{R})$, par des méthodes dites *directes*, c'est-à-dire fournissant, en l'absence d'erreurs d'arrondi, la solution *exacte* en un nombre *fini* [1] d'opérations élémentaires. On verra que ces méthodes consistent en la construction d'une matrice inversible M telle que MA soit une matrice triangulaire, le système linéaire équivalent (au sens où il possède la même solution) obtenu,

$$MA\boldsymbol{x} = M\boldsymbol{b},$$

étant alors « facile » à résoudre (on verra ce que l'on entend précisément par là). Une telle idée est par exemple à la base de la célèbre *méthode d'élimination de Gauss*, qui permet de ramener la résolution d'un système linéaire quelconque à celle d'un système triangulaire supérieur.

Après avoir présenté quelques cas pratiques d'application de ces méthodes et donné des éléments sur la résolution numérique des systèmes triangulaires, nous introduisons dans le détail la méthode d'élimination de Gauss. Ce procédé est ensuite réinterprété en termes d'opérations matricielles, donnant lieu à une méthode de *factorisation* (*factorization* ou *decomposition* en anglais) des matrices. Les propriétés d'une telle décomposition sont explorées, notamment dans le cas de matrices particulières. Le chapitre se conclut sur la présentation de quelques autres méthodes de factorisation.

4.1 Remarques sur la résolution des systèmes triangulaires

Observons tout d'abord que la solution du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$, avec A une matrice inversible, *ne s'obtient pas* en inversant A, puis en calculant le vecteur $A^{-1}\boldsymbol{b}$, mais en réalisant plutôt des combinaisons linéaires sur les lignes du système et des substitutions. En effet, on peut facilement voir que le calcul de la matrice A^{-1} équivaut à résoudre n systèmes linéaires [2], ce qui s'avère bien plus coûteux que la résolution du *seul* système dont on cherche la solution.

Considérons à présent un système linéaire (4.1) dont la matrice est triangulaire inférieure, c'est-à-dire de la

1. On oppose ici ce type de méthodes avec les méthodes dites *itératives*, qui nécessitent *a priori* un nombre infini d'opérations pour fournir la solution. Celles-ci sont l'objet du chapitre 5.
2. Ces systèmes sont

$$A\boldsymbol{x_i} = \boldsymbol{e_i},\ 1 \leq i \leq n,$$

où $\boldsymbol{e}_i$ désigne le i^{e} vecteur de la base canonique de $M_{n,1}(\mathbb{R})$.

forme

$$\begin{array}{ccccccccc} a_{11}x_1 & & & & & & & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & & & & & = & b_2 \\ \vdots & & \vdots & & \ddots & & & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \ldots & + & a_{nn}x_n & = & b_n \end{array}.$$

Si la matrice A est inversible, ses termes diagonaux a_{ii}, $i = 1, \ldots, n$, sont tous non nuls [3] et la résolution du système est alors extrêmement simple : on calcule x_1 par une division, que l'on substitue ensuite dans la deuxième équation pour obtenir x_2, et ainsi de suite... Cette méthode, dite *de « descente »* (*"forward substitution"* en anglais), s'écrit

$$\begin{aligned} x_1 &= \frac{b_1}{a_{11}} \\ x_i &= \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j \right),\ i = 2, \ldots, n. \end{aligned} \tag{4.2}$$

L'algorithme mis en œuvre pour cette résolution effectue $\frac{1}{2}n(n-1)$ soustractions, $\frac{1}{2}n(n-1)$ multiplications et n divisions pour calculer la solution, soit un nombre d'opérations global de l'ordre de n^2.

Exemple de résolution d'un système triangulaire inférieur. Appliquons une approche orientée colonne pour la résolution du système

$$\begin{pmatrix} 2 & 0 & 0 \\ 1 & 5 & 0 \\ 7 & 9 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \\ 5 \end{pmatrix}.$$

On trouve que $x_1 = 3$ et l'on considère ensuite le système à deux équations et deux inconnues

$$\begin{pmatrix} 5 & 0 \\ 9 & 8 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \end{pmatrix} - 3 \begin{pmatrix} 1 \\ 7 \end{pmatrix},$$

pour lequel on trouve $x_2 = -\dfrac{1}{5}$. On a enfin

$$8x_3 = -16 + \frac{9}{5},$$

soit $x_3 = -\dfrac{71}{40}$.

Le cas d'un système linéaire dont la matrice est inversible et triangulaire supérieure se traite de manière analogue, par la méthode dite *de « remontée »* (*"back substitution"* en anglais) suivante

$$\begin{aligned} x_n &= \frac{b_n}{a_{nn}} \\ x_i &= \frac{1}{a_{ii}}\left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j \right),\ i = n-1, \ldots, 1, \end{aligned} \tag{4.3}$$

et dont la complexité est également de l'ordre de n^2 opérations.

4.2 Méthode d'élimination de Gauss

Une technique de choix pour ramener la résolution d'un système linéaire quelconque à celle d'un système triangulaire est la *méthode d'élimination de Gauss* (*Gaussian elimination* en anglais). Celle-ci consiste en premier lieu à transformer, par des opérations simples sur les équations, le système en un système équivalent, c'est-à-dire ayant la (ou les) même(s) solution(s), $MA\boldsymbol{x} = M\boldsymbol{b}$, dans lequel MA est une matrice triangulaire supérieure [4] (on dit encore que la matrice du système est sous forme *échelonnée*). Cette étape de mise à zéro d'une partie des coefficients de la matrice est qualifiée d'*élimination* et utilise de manière essentielle le fait qu'on ne modifie pas la solution d'un système linéaire en ajoutant à une équation donnée une combinaison linéaire des autres équations. Lorsque la matrice du système est inversible, la solution du système peut ensuite être obtenue par une méthode de remontée, mais le procédé d'élimination est très général et s'applique à des matrices rectangulaires.

3. On a en effet $a_{11}a_{22}\ldots a_{nn} = \det(A) \neq 0$.
4. Il faut bien noter qu'on ne calcule en pratique jamais explicitement la matrice d'élimination M, mais seulement les produits MA et $M\boldsymbol{b}$.

4.2.1 Élimination sans échange

Commençons par décrire étape par étape la méthode dans sa forme de base, dite *sans échange*, en considérant le système linéaire (4.1), avec A étant une matrice inversible d'ordre n. Supposons de plus que le terme a_{11} de la matrice A est non nul. Nous pouvons alors éliminer l'inconnue x_1 de la deuxième à la n^{e} ligne du système en leur retranchant respectivement la première ligne multipliée par le coefficient $\frac{a_{i1}}{a_{11}}$, $i = 2, \ldots, n$. En notant $A^{(1)}$ et $\boldsymbol{b}^{(1)}$ la matrice et le vecteur second membre résultant de ces opérations [5], on a alors

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}}\, a_{1j} \text{ et } b_i^{(1)} = b_i - \frac{a_{i1}}{a_{11}}\, b_1,\ i = 2, \ldots, n, j = 1, \ldots, n,$$

les coefficients de la première ligne restant les mêmes, et le système $A^{(1)}\boldsymbol{x} = \boldsymbol{b}^{(1)}$ est équivalent au système de départ. En supposant non nul le coefficient diagonal $a_{22}^{(1)}$ de $A^{(1)}$, on peut ensuite procéder à l'élimination de l'inconnue x_2 de la troisième à la n^{e} ligne de ce nouveau système, et ainsi de suite. On obtient, sous l'hypothèse $a_{k+1\,k+1}^{(k)} \neq 0$, $k = 0, \ldots, n-2$, une suite finie de matrices $A^{(k)}$, $1 \leq k \leq n-1$, de la forme

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & \cdots & \cdots & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & & & & a_{2n}^{(k)} \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \cdots & 0 & a_{k+1\,k+1}^{(k)} & \cdots & a_{k+1\,n}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{n\,k+1}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix},$$

telle que le système $A^{(n-1)}\boldsymbol{x} = \boldsymbol{b}^{(n-1)}$ est triangulaire supérieur. Les quantités $a_{k+1\,k+1}^{(k)}$, $k = 0, \ldots, n-2$, sont appelées les *pivots* et l'on a supposé qu'elles étaient non nulles à chaque étape, les formules permettant de passer du système linéaire de matrice $A^{(k)}$ et de second membre $\boldsymbol{b}^{(k)}$ au suivant se résumant à

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik+1}^{(k)}}{a_{k+1\,k+1}^{(k)}}\, a_{k+1\,j}^{(k)} \text{ et } b_i^{(k+1)} = b_i^{(k)} - \frac{a_{ik+1}^{(k)}}{a_{k+1\,k+1}^{(k)}}\, b_{k+1}^{(k)},\ i = k+2, \ldots, n, j = k+1, \ldots, n,$$

les autres coefficients étant inchangés. En pratique, pour une résolution « à la main » d'un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ par cette méthode, il est commode d'appliquer l'élimination à la matrice « *augmentée* » $\begin{pmatrix} A & \boldsymbol{b} \end{pmatrix}$.

Exemple d'application de la méthode d'élimination de Gauss sans échange. Considérons la résolution par la méthode d'élimination de Gauss sans échange du système linéaire suivant

$$\left\{ \begin{array}{rcrcrcrcl} x_1 & + & 2x_2 & + & 3x_3 & + & 4x_4 & = & 11 \\ 2x_1 & + & 3x_2 & + & 4x_3 & + & x_4 & = & 12 \\ 3x_1 & + & 4x_2 & + & x_3 & + & 2x_4 & = & 13 \\ 4x_1 & + & x_2 & + & 2x_3 & + & 3x_4 & = & 14 \end{array} \right. .$$

À la première étape, le pivot vaut 1 et on soustrait de la deuxième (resp. troisième (resp. quatrième)) équation la première équation multipliée par 2 (resp. 3 (resp. 4)) pour obtenir

$$\left\{ \begin{array}{rcrcrcrcl} x_1 & + & 2x_2 & + & 3x_3 & + & 4x_4 & = & 11 \\ & - & x_2 & - & 2x_3 & - & 7x_4 & = & -10 \\ & - & 2x_2 & - & 8x_3 & - & 10x_4 & = & -20 \\ & - & 7x_2 & - & 10x_3 & - & 13x_4 & = & -3 \end{array} \right. .$$

Le pivot vaut -1 à la deuxième étape. On retranche alors à la troisième (resp. quatrième) équation la deuxième équation multipliée par -2 (resp. -7), d'où le système

$$\left\{ \begin{array}{rcrcrcrcl} x_1 & + & 2x_2 & + & 3x_3 & + & 4x_4 & = & 11 \\ & - & x_2 & - & 2x_3 & - & 7x_4 & = & -10 \\ & & & - & 4x_3 & + & 4x_4 & = & 0 \\ & & & & 4x_3 & + & 36x_4 & = & 40 \end{array} \right. .$$

5. On pose $A^{(0)} = A$ et $\boldsymbol{b}^{(0)} = \boldsymbol{b}$ pour être consistant.

À la dernière étape, le pivot est égal à -4 et on soustrait à la dernière équation l'avant-derniére multipliée par -1 pour arriver à

$$\left\{\begin{array}{rcrcrcrcr} x_1 & + & 2x_2 & + & 3x_3 & + & 4x_4 & = & 11 \\ & - & x_2 & - & 2x_3 & - & 7x_4 & = & -10 \\ & & & - & 4x_3 & + & 4x_4 & = & 0 \\ & & & & & & 40x_4 & = & 40 \end{array}\right. .$$

Ce système triangulaire, équivalent au système d'origine, est enfin résolu par remontée :

$$\left\{\begin{array}{l} x_4 = 1 \\ x_3 = x_4 = 1 \\ x_2 = 10-2-7 = 1 \\ x_1 = 11-2-3-4 = 2 \end{array}\right. .$$

Comme on l'a vu, la méthode d'élimination de Gauss, dans sa forme sans échange, ne peut s'appliquer que si tous les pivots $a^{(k)}_{k+1\,k+1}$, $k = 0, \ldots, n-2$, sont non nuls, ce qui rejette de fait des matrices inversibles aussi simples que

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

De plus, le fait que la matrice soit inversible n'empêche aucunement l'apparition d'un pivot nul durant l'élimination, comme le montre l'exemple ci-dessous.

Exemple de mise en échec de la méthode d'élimination de Gauss sans échange. Considérons la matrice inversible

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{pmatrix} = A^{(0)}.$$

On a alors

$$A^{(1)} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & -1 \\ 0 & -6 & -12 \end{pmatrix},$$

et l'élimination s'interrompt à l'issue de la première étape, le pivot $a^{(1)}_{22}$ étant nul.

Il apparaît donc que des conditions plus restrictives que l'inversibilité de la matrice sont nécessaires pour assurer la bonne exécution de cette méthode. Celles-ci sont fournies par le théorème 29. Indiquons qu'il existe des catégories de matrices pour lesquelles la méthode de Gauss sans échange peut-être utilisée sans aucun risque. Parmi celles-ci, on trouve les matrices *à diagonale dominante par lignes ou par colonnes* et les matrices *symétriques définies positives*.

4.2.2 Élimination de Gauss avec échange

Dans sa forme générale, la méthode d'élimination de Gauss permet de transformer un système linéaire dont la matrice est carrée (inversible ou non) ou même rectangulaire en un système échelonné équivalent. En considérant le cas d'une matrice A carrée inversible, nous allons maintenant décrire les modifications à apporter à la méthode déjà présentée pour mener l'élimination à son terme. Dans tout ce qui suit, les notations de la section 4.2.1 sont conservées.

À la première étape, au moins l'un des coefficients de la première colonne de la matrice $A^{(0)}(=A)$ est non nul, faute de quoi la matrice A ne serait pas inversible. On choisit [6] un de ces éléments comme premier pivot d'élimination et l'on échange alors la première ligne du système avec celle du pivot avant de procéder à l'élimination de la première colonne de la matrice résultante, c'est-à-dire l'annulation de tous les éléments de la première colonne de la matrice (permutée) du système situés sous la diagonale. On note $A^{(1)}$ et $\boldsymbol{b}^{(1)}$ la matrice et le second membre du système obtenu et l'on réitère ce procédé. Par la suite, à l'étape $k+1$, $1 \leq k \leq n-2$, la matrice $A^{(k)}$ est inversible [7], et donc l'un au moins des éléments $a^{(k)}_{ik+1}$, $k+1 \leq i \leq n$, est différent de zéro. Après avoir choisi comme pivot l'un

6. Pour l'instant, on ne s'intéresse pas au choix *effectif* du pivot, qui est cependant d'une importance cruciale pour la stabilité numérique de la méthode. Ce point est abordé dans la section 4.2.3.

7. On a en effet que $\det(A^{(k)}) = \pm\det(A)$, $k = 0, \ldots, n-1$. On renvoie à la section 4.3.1 pour une justification de ce fait.

de ces coefficients non nuls, on effectue l'échange de la ligne de ce pivot avec la $k+1^{\text{e}}$ ligne de la matrice $A^{(k)}$, puis l'élimination conduisant à la matrice $A^{(k+1)}$. Ainsi, on arrive après $n-1$ étapes à la matrice $A^{(n-1)}$, dont le coefficient $a_{nn}^{(n-1)}$ est non nul.

En raison de l'échange de lignes qui a éventuellement lieu avant chaque étape d'élimination, on parle de méthode d'élimination de Gauss *avec échange*.

Exemple d'application de la méthode d'élimination de Gauss avec échange. Considérons la résolution du système linéaire $A\boldsymbol{x} = \boldsymbol{b}$, avec

$$A = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 3 & 6 & 1 & -2 \\ -1 & 1 & 2 & 3 \\ 1 & 1 & -4 & 1 \end{pmatrix} \text{ et } \boldsymbol{b} = \begin{pmatrix} 0 \\ -7 \\ 4 \\ 2 \end{pmatrix},$$

par application de la méthode d'élimination de Gauss avec échange. On trouve successivement

$$A^{(1)} = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 0 & 0 & 7 & -\frac{7}{2} \\ 0 & 3 & 0 & \frac{7}{2} \\ 0 & -1 & -2 & \frac{1}{2} \end{pmatrix} \text{ et } \boldsymbol{b}^{(1)} = \begin{pmatrix} 0 \\ -7 \\ 4 \\ 2 \end{pmatrix},$$

$$A^{(2)} = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 0 & 3 & 0 & \frac{7}{2} \\ 0 & 0 & 7 & -\frac{7}{2} \\ 0 & 0 & -2 & \frac{5}{3} \end{pmatrix} \text{ et } \boldsymbol{b}^{(2)} = \begin{pmatrix} 0 \\ 4 \\ -7 \\ \frac{10}{3} \end{pmatrix},$$

et

$$A^{(3)} = \begin{pmatrix} 2 & 4 & -4 & 1 \\ 0 & 3 & 0 & \frac{7}{2} \\ 0 & 0 & 7 & -\frac{7}{2} \\ 0 & 0 & 0 & \frac{2}{3} \end{pmatrix} \text{ et } \boldsymbol{b}^{(3)} = \begin{pmatrix} 0 \\ 4 \\ -7 \\ \frac{4}{3} \end{pmatrix},$$

d'où la solution $\boldsymbol{x} = \begin{pmatrix} 1 & -1 & 0 & 2 \end{pmatrix}^\top$. On note que l'on a procédé au cours de la deuxième étape à l'échange des deuxième et troisième lignes.

On pourra remarquer que si la matrice A est non inversible, alors tous les éléments $a_{ik+1}^{(k)}$, $k+1 \leq i \leq n$, seront nuls pour au moins une valeur de k entre 0 et $n-1$. Si cela vient à se produire alors que $k < n-1$, on n'a pas besoin de réaliser l'élimination dans la $k+1^{\text{e}}$ colonne (puisque celle-ci est déjà nulle) et l'on passe simplement à l'étape suivante en posant $A^{(k+1)} = A^{(k)}$ et $\boldsymbol{b}^{(k+1)} = \boldsymbol{b}^{(k)}$. L'élimination est donc bien possible pour une matrice carrée non inversible et l'on a démontré le résultat suivant.

Théorème 28 *Soit A une matrice carrée, inversible ou non. Il existe au moins une matrice inversible M telle que la matrice MA soit triangulaire supérieure.*

Il reste à compter le nombre d'opérations élémentaires que requiert l'application de la méthode d'élimination de Gauss pour la résolution d'un système linéaire de n équations à n inconnues. Tout d'abord, pour passer de la matrice $A^{(k)}$ à la matrice $A^{(k+1)}$, $0 \leq k \leq n-2$, on effectue $(n-k-1)^2$ soustractions, $(n-k-1)^2$ multiplications et $n-k-1$ divisions, ce qui correspond à un total de $\frac{1}{6}(2n-1)n(n-1)$ soustractions, $\frac{1}{6}(2n-1)n(n-1)$ multiplications et $\frac{1}{2}n(n-1)$ divisions pour l'élimination complète. Pour la mise à jour du second membre à l'étape $k+1$, on a besoin de $n-k-1$ soustractions et autant de multiplications, soit en tout $\frac{1}{2}n(n-1)$ soustractions et $\frac{1}{2}n(n-1)$ multiplications. Enfin, il faut faire $\frac{1}{2}n(n-1)$ soustractions, autant de multiplications et n divisions pour résoudre le système final par une méthode de remontée.

En tout, la résolution du système par la méthode d'élimination de Gauss nécessite donc environ $\frac{n^3}{3}$ soustractions, $\frac{n^3}{3}$ multiplications et $\frac{n^2}{2}$ divisions. À titre de comparaison, le calcul de la solution du système par la règle de Cramer (voir la proposition 35) requiert, en utilisant un développement « brutal » par ligne ou colonne pour le calcul des déterminants, environ $(n+1)!$ additions ou soustractions, $(n+2)!$ multiplications et n divisions. Ainsi, pour $n = 10$ par exemple, on obtient un compte d'approximativement 700 opérations pour la méthode d'élimination de Gauss contre près de 479000000 opérations pour la règle de Cramer !

4.2.3 Choix du pivot

Revenons à présent sur le choix des pivots lors de l'élimination. À la $k+1^{\text{e}}$ étape du procédé, si l'élément $a_{k+1\,k+1}^{(k)}$ est non nul, il semble naturel de l'utiliser comme pivot (c'est d'ailleurs ce que l'on fait dans la méthode de Gauss sans échange). Cependant, à cause de la présence d'erreurs d'arrondi en pratique, cette manière de procéder est en général à proscrire, comme l'illustre l'exemple d'instabilité numérique suivant.

Exemple d'application numérique. Supposons que les calculs soient effectués en virgule flottante dans le système décimal, avec une mantisse à trois chiffres, et considérons le système

$$\begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

dont la solution « exacte » est $x_1 = 1,\overline{0001}$ et $x_2 = 0,\overline{9998}$. En choisissant le nombre 10^{-4} comme pivot à la première étape de l'élimination de Gauss, on obtient le système triangulaire

$$\begin{pmatrix} 10^{-4} & 1 \\ 0 & -9990 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -9990 \end{pmatrix},$$

car les nombres $-10^4 + 1 = -9999$ et $-10^4 + 2 = -9998$ sont tous deux arrondis au nombre -9990. La solution numérique calculée est alors

$$x_1 = 0 \text{ et } x_2 = 1,$$

ce qui très différent de la véritable solution du système. Si, par contre, on commence par échanger les deux équations du système pour utiliser le nombre 1 comme pivot, on trouve

$$\begin{pmatrix} 1 & 1 \\ 0 & 0,999 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0,999 \end{pmatrix},$$

puisque les nombres $-10^{-4}+1 = 0,9999$ et $-2\ 10^{-4}+1 = 0,9998$ sont chacun arrondis au nombre $0,999$. La solution calculée vaut alors

$$x_1 = 1 \text{ et } x_2 = 1,$$

ce qui est cette fois très satisfaisant.

En général, le changement de pivot n'a pas un effet aussi spectaculaire que dans cet exemple, mais il n'en demeure pas moins essentiel lorsque les calculs sont effectués en arithmétique à virgule flottante. De fait, pour éviter la propagation d'erreurs et obtenir une meilleure stabilité numérique de la méthode, il faut chercher, même lorsque le pivot « naturel » est non nul, à choisir le plus grand pivot en valeur absolue. On peut pour cela suivre, au début de la $k+1^{\text{e}}$ étape, $0 \leq k \leq n-2$, de l'élimination,

- soit une stratégie de *pivot partiel* (*partial pivoting* en anglais) dans laquelle le pivot est l'élément $a_{rk+1}^{(k)}$ de la $k+1^{\text{e}}$ colonne de la matrice $A^{(k)}$ situé sous la diagonale ayant la plus grande valeur absolue,

$$|a_{rk+1}^{(k)}| = \max_{k+1 \leq i \leq n} |a_{ik+1}^{(k)}|,$$

- soit une stratégie de *pivot total* (*complete pivoting* en anglais), plus coûteuse, dans laquelle le pivot est l'élément $a_{rs}^{(k)}$ de la sous-matrice $(a_{ij}^{(k)})_{k+1 \leq i,j \leq n}$ le plus grand en valeur absolue,

$$|a_{rs}^{(k)}| = \max_{k+1 \leq i,j \leq n} |a_{ij}^{(k)}|,$$

Dans le derniers cas, si le pivot n'est pas dans la $k+1^{\text{e}}$ colonne, il faut procéder à un échange de colonnes en plus d'un éventuel échange de lignes.

Quelle que soit la stratégie adoptée, la recherche des pivots doit également être prise en compte dans l'évaluation du coût global de la méthode d'élimination de Gauss. Celle-ci demande de l'ordre de n^2 comparaisons au total pour la stratégie de pivot partiel et de l'ordre de n^3 comparaisons pour celle de pivot total, la première étant privilégiée en raison de sa complexité algorithmique moindre et de performances généralement très bonnes.

Exemple d'application de la méthode d'élimination de Gauss–Jordan pour l'inversion d'une matrice. Soit la matrice $A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$. La matrice augmentée est alors

$$\begin{pmatrix} A & I_n \end{pmatrix} = \left(\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 1 & 0 \\ 0 & -1 & 2 & 0 & 0 & 1 \end{array}\right),$$

et l'on trouve successivement

$$k=0, \quad \left(\begin{array}{ccc|ccc} 1 & -1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 3/2 & -1 & 1/2 & 1 & 0 \\ 0 & -1 & 2 & 0 & 0 & 1 \end{array}\right),$$

$$k=1, \quad \left(\begin{array}{ccc|ccc} 1 & 0 & -1/3 & 2/3 & 1/3 & 0 \\ 0 & 1 & -2/3 & 1/3 & 2/3 & 0 \\ 0 & 0 & 4/3 & 1/3 & 2/3 & 1 \end{array}\right),$$

$$k=2, \quad \left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 3/4 & 1/2 & 1/4 \\ 0 & 1 & 0 & 1/2 & 1 & 1/2 \\ 0 & 0 & 1 & 1/4 & 1/2 & 3/4 \end{array}\right),$$

d'où $A^{-1} = \frac{1}{4}\begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}$.

4.3 Interprétation matricielle de l'élimination de Gauss : la factorisation LU

Nous allons maintenant montrer que la méthode de Gauss dans sa forme sans échange est équivalente à la décomposition de la matrice A sous la forme d'un produit de deux matrices, $A = LU$, avec L une matrice triangulaire inférieure (*lower triangular* en anglais), qui est l'inverse de la matrice M des transformations successives appliquées à la matrice A lors de l'élimination de Gauss sans échange, et U une matrice triangulaire supérieure (*upper triangular* en anglais), avec $U = A^{(n-1)}$ en reprenant la notation utilisée dans la section 4.2.1.

4.3.1 Formalisme matriciel

Chacune des opérations que nous avons effectuées pour transformer le système linéaire lors de l'élimination de Gauss, que ce soit l'échange de deux lignes ou l'annulation d'une partie des coefficients d'une colonne de la matrice $A^{(k)}$, $0 \leq k \leq n-2$, peut se traduire matriciellement par la multiplication de la matrice et du second membre du système linéaire courant par une matrice inversible particulière. L'introduction de ces matrices va permettre de traduire le procédé d'élimination dans un formalisme matriciel débouchant sur une factorisation remarquable de la matrice A.

Matrices des transformations élémentaires

Soit $(m,n) \in (\mathbb{N}\backslash\{0,1\})^2$ et $A = (a_{ij})_{1\leq i\leq m,\ 1\leq j\leq n} \in M_{m,n}(\mathbb{R})$. On appelle *opérations élémentaires sur les lignes de A* les transformations suivantes :

- l'échange (entre elles) des i^{e} et j^{e} lignes de A ;
- la multiplication de la i^{e} ligne de A par un scalaire $\lambda \in \mathbb{R}\backslash\{0\}$;
- le remplacement de la i^{e} ligne de A par la somme de cette même ligne avec la j^{e} ligne de A multipliée par un scalaire λ, où $\lambda \in \mathbb{R}\backslash\{0\}$.

Explicitons à présent les opérations matricielles correspondant à chacunes de ces opérations. Tout d'abord, échanger les i^{e} et j^{e} lignes, $(i,j) \in \{1\ldots,m\}^2$, de la matrice A revient à multiplier à gauche cette matrice par la

matrice de permutation (*permutation matrix* en anglais)

$$P_{ij} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & & & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & & & & & & \vdots \\ \vdots & & \ddots & 0 & 0 & \cdots & 0 & 1 & & & \vdots \\ \vdots & & & 0 & 1 & \ddots & & 0 & & & \vdots \\ \vdots & & & \vdots & \ddots & \ddots & \ddots & \vdots & & & \vdots \\ \vdots & & & 0 & & \ddots & 1 & 0 & & & \vdots \\ \vdots & & & 1 & 0 & \dots & 0 & 0 & & & \vdots \\ \vdots & & & & & & & & 1 & & \vdots \\ \vdots & & & & & & & & & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix} = I_m + (E_{ij} + E_{ji} - E_{ii} - E_{jj}) \in M_m(\mathbb{R}).$$

Cette matrice est symétrique et orthogonale, de déterminant valant -1.

La multiplication de la i^{e} ligne de la matrice A par un scalaire non nul λ s'effectue en multipliant à gauche cette matrice par la *matrice de dilatation* (*directional scaling matrix* en anglais)

$$D_i(\lambda) = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & & \vdots \\ \vdots & & \ddots & \lambda & \ddots & & \vdots \\ \vdots & & & \ddots & 1 & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix} = I_m + (\lambda - 1)E_{ii} \in M_m(\mathbb{R}).$$

Cette matrice est inversible et $D_i(\lambda)^{-1} = D_i\left(\frac{1}{\lambda}\right)$.

Enfin, le remplacement de la i^{e} ligne de A par la somme de la i^{e} ligne et de la j^{e} ligne, $i \neq j$, multipliée par un scalaire non nul λ est obtenu en multipliant à gauche la matrice A par la *matrice de transvection* (*shear matrix* en anglais)

$$T_{ij}(\lambda) = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & & & \vdots \\ \vdots & \ddots & 1 & \cdots & 0 & & \vdots \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ \vdots & & \lambda & \cdots & 1 & \ddots & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix} = I_m + \lambda\, E_{ij} \in M_m(\mathbb{R})$$

(on a supposé dans l'exemple que $j < i$). Cette matrice a pour inverse[8] $T_{ij}(-\lambda)$. Pour la suite, il est important de remarquer que, étant donné trois entiers naturels i, k et l, tels que $1 \leq k < i < l \leq m$, et deux scalaires non nuls λ et μ, le produit des matrices de tranvection $T_{ik}(\lambda)$ et $T_{lk}(\mu)$ de $M_m(\mathbb{R})$ est commutatif et vaut

$$T_{ik}(\lambda)T_{lk}(\mu) = I_m + \lambda\, E_{ik} + \mu\, E_{lk}.$$

On effectue de manière analogue des opérations élémentaires *sur les colonnes* de la matrice A en multipliant *à droite* cette dernière par les matrices d'ordre n correspondantes.

8. L'ensemble des matrices de transvection de $M_m(\mathbb{R})$ engendre le *groupe spécial linéaire* (*special linear group* en anglais) $\mathrm{SL}_m(\mathbb{R})$, constitué des matrices d'ordre m dont le déterminant vaut 1, et, avec l'ensemble des matrices de dilatation de $M_m(\mathbb{R})$, le groupe général linéaire $\mathrm{GL}_m(\mathbb{R})$.

Exemple d'action d'une matrice de permutation. Soit les matrices $A=\begin{pmatrix}1&2&3\\4&5&6\\7&8&9\end{pmatrix}$ et $P_{23}=\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}$. On a

$$P_{23}A=\begin{pmatrix}1&2&3\\7&8&9\\4&5&6\end{pmatrix} \text{ et } AP_{23}=\begin{pmatrix}1&3&2\\4&6&5\\7&9&8\end{pmatrix}.$$

Factorisation LU

Si l'élimination arrive à son terme sans qu'il y ait besoin d'échanger des lignes du système linéaire, la matrice inversible M du théorème 28 est unique et égale au produit

$$M=E^{(n-1)}\dots E^{(2)}E^{(1)}$$

de $n-1$ *matrices d'élimination* définies par

$$E^{(k)}=\prod_{i=k+1}^{n}T_{ik}\left(-\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}\right)=\begin{pmatrix}1&0&\dots&\dots&\dots&\dots&\dots&0\\0&1&\ddots&&&&&\vdots\\\vdots&\ddots&\ddots&\ddots&&&&\vdots\\\vdots&&0&1&\ddots&&&\vdots\\\vdots&&\vdots&-\frac{a_{k+1k}^{(k-1)}}{a_{kk}^{(k-1)}}&1&\ddots&&\vdots\\\vdots&&\vdots&-\frac{a_{k+2k}^{(k-1)}}{a_{kk}^{(k-1)}}&0&\ddots&\ddots&\vdots\\\vdots&&\vdots&\vdots&\vdots&\ddots&\ddots&0\\0&\dots&0&-\frac{a_{nk}^{(k-1)}}{a_{kk}^{(k-1)}}&0&\dots&0&1\end{pmatrix},\ 1\le k\le n-1. \tag{4.4}$$

Par construction, la matrice M est triangulaire inférieure et son inverse est donc également une matrice triangulaire inférieure. Il en résulte que la matrice A s'écrit comme le produit

$$A=LU, \tag{4.5}$$

dans lequel $L=M^{-1}$ et $U=MA=A^{(n-1)}$ est une matrice triangulaire supérieure. Fait remarquable, la matrice L se calcule de manière immédiate à partir des matrices $E^{(k)}$, $1\le k\le n-1$, alors qu'il n'existe pas d'expression simple pour M. En effet, chacune des matrices d'élimination définies par (4.4) étant produit de matrices de transvection, il est facile de vérifier que son inverse vaut

$$(E^{(k)})^{-1}=\prod_{i=k+1}^{n}T_{ik}\left(\frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}}\right)=\begin{pmatrix}1&0&\dots&\dots&\dots&\dots&\dots&0\\0&1&\ddots&&&&&\vdots\\\vdots&\ddots&\ddots&\ddots&&&&\vdots\\\vdots&&0&1&\ddots&&&\vdots\\\vdots&&\vdots&\frac{a_{k+1k}^{(k-1)}}{a_{kk}^{(k-1)}}&1&\ddots&&\vdots\\\vdots&&\vdots&\frac{a_{k+2k}^{(k-1)}}{a_{kk}^{(k-1)}}&0&\ddots&\ddots&\vdots\\\vdots&&\vdots&\vdots&\vdots&\ddots&\ddots&0\\0&\dots&0&\frac{a_{nk}^{(k-1)}}{a_{kk}^{(k-1)}}&0&\dots&0&1\end{pmatrix},\ 1\le k\le n-1,$$

et l'on a [9]

$$L = (E^{(1)})^{-1}(E^{(2)})^{-1}\dots(E^{(n-1)})^{-1} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \frac{a_{21}^{(0)}}{a_{11}^{(0)}} & 1 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \frac{a_{k+1\,k}^{(k-1)}}{a_{kk}^{(k-1)}} & 1 & \ddots & \vdots \\ \vdots & & \vdots & \ddots & \ddots & 0 \\ \frac{a_{n1}^{(0)}}{a_{11}^{(0)}} & \dots & \frac{a_{nk}^{(k-1)}}{a_{kk}^{(k-1)}} & \dots & \frac{a_{n\,n-1}^{(n-2)}}{a_{n-1\,n-1}^{(n-2)}} & 1 \end{pmatrix}.$$

Si des échanges de lignes ont eu lieu lors de l'élimination, la [10] matrice M s'écrit

$$M = E^{(n-1)}P^{(n-1)}\dots E^{(2)}P^{(2)}E^{(1)}P^{(1)},$$

où la matrice $P^{(k)}$, $1 \leq k \leq n-1$, est soit la matrice de permutation correspondant à l'échange de lignes effectué à la k^{e} étape, soit la matrice identité si le pivot « naturel » est utilisé. En écrivant que

$$M = E^{(n-1)}(P^{(n-1)}E^{(n-2)}P^{(n-1)})\dots(P^{(n-1)}\dots P^{(2)}E^{(1)}P^{(2)}\dots P^{(n-1)})(P^{(n-1)}\dots P^{(2)}P^{(1)}),$$

et en posant $P = P^{(n-1)}\dots P^{(2)}P^{(1)}$, on obtient $L = PM^{-1}$ et $U = (MP^{-1})PA$, d'où

$$PA = LU.$$

Enfin, si la stratégie de choix de pivot employée conduit à des échanges de colonnes en plus des éventuels échanges de lignes, la factorisation prend la forme

$$PAQ = LU,$$

les matrices P et Q rendant respectivement compte des échanges de lignes et de colonnes.

Exemple d'application de la factorisation $PA = LU$. Revenons à l'exemple de mise en échec de la méthode d'élimination de Gauss, pour lequel le pivot « naturel » est nul à la seconde étape. La recherche d'un pivot partiel conduit à l'échange de la deuxième ligne avec la troisième et l'on arrive à

$$A^{(2)} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & -1 \end{pmatrix} = U.$$

Les matrices d'élimination aux deux étapes effectuées sont respectivement

$$E^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -7 & 0 & 1 \end{pmatrix} \text{ et } E^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

et les matrices d'échange sont respectivement

$$P^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ et } P^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

d'où

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 7 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}.$$

et la matrice de permutation

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

9. La vérification est laissée en exercice.
10. Il n'y a pas forcément unicité de la matrice dans ce cas, en raison de possibles multiples choix de pivots.

On a

$$LU = \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 2 & 4 & 5 \end{pmatrix} = PA.$$

La résolution du système linéaire (4.1) connaissant la factorisation LU de la matrice A se ramène à celle du système triangulaire inférieur

$$L\boldsymbol{y} = \boldsymbol{b} \tag{4.6}$$

par une méthode de descente, suivie de celle du système triangulaire supérieur

$$U\boldsymbol{x} = \boldsymbol{y} \tag{4.7}$$

par une méthode de remontée, ce que l'on accomplit en effectuant $n(n-1)$ additions, $n(n-1)$ multiplications et $2n$ divisions. Dans le cas d'une factorisation de type $PA = LU$ (resp. $PAQ = LU$), il faudra appliquer la matrice de permutation P au vecteur $\boldsymbol{b}$ de manière à tout d'abord résoudre le système $L\boldsymbol{y} = P\boldsymbol{b}$ avant de considérer le système $U\boldsymbol{x} = \boldsymbol{y}$ (resp. $UQ\boldsymbol{x} = \boldsymbol{y}$).

Exemple d'application de la factorisation LU pour la résolution d'un système linéaire. Considérons la matrice

$$A = \begin{pmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 10 \end{pmatrix}.$$

En appliquant de l'algorithme de factorisation, on arrive à

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

Si $\boldsymbol{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, la solution de $L\boldsymbol{y} = \boldsymbol{b}$ est $\boldsymbol{y} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ et celle de $U\boldsymbol{x} = \boldsymbol{y}$ est $\boldsymbol{x} = \frac{1}{3}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$.

On observera que ce type de factorisation est particulièrement avantageux lorsque l'on doit résoudre plusieurs systèmes linéaires ayant tous la même matrice et des seconds membres différents. En effet, une fois obtenue la factorisation LU de la matrice, la résolution de chaque système ne nécessite que de l'ordre de n^2 opérations.

Terminons cette section en indiquant que la méthode de factorisation LU fournit une manière rapide de calculer le déterminant de la matrice A, qui n'est autre, au signe près, que le produit des pivots, puisque [11]

$$\det(PA) = \det(LU) = \det(L)\det(U) = \det(U) = \left(\prod_{i=1}^{n} u_{ii}\right),$$

et

$$\det(A) = \frac{\det(PA)}{\det(P)} = \begin{cases} \det(PA) & \text{si on a effectué un nombre pair d'échanges de lignes,} \\ -\det(PA) & \text{si on a effectué un nombre impair d'échanges de lignes,} \end{cases}$$

le déterminant d'une matrice de permutation étant égal à -1.

4.3.2 Condition d'existence de la factorisation LU

Commençons par donner une condition suffisante assurant qu'il n'y aura pas d'échange de lignes durant l'élimination de Gauss, ce qui conduira bien à une factorisation de la forme (4.5) de la matrice. On va à cette occasion aussi établir que cette décomposition est unique si l'on impose la valeur 1 aux éléments diagonaux de L (c'est précisément la valeur obtenue avec la construction par élimination de Gauss).

11. On laisse le lecteur traiter le cas d'une factorisation de la forme $PAQ = LU$.

Théorème 29 (condition suffisante d'existence et d'unicité de la factorisation LU) *Soit A une matrice d'ordre n. La factorisation LU de A, avec $l_{ii} = 1$ pour $i = 1, \ldots, n$, existe et est unique si toutes les sous-matrices principales*

$$A_k = \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \ldots & a_{kk} \end{pmatrix}, \ 1 \leq k \leq n, \tag{4.8}$$

extraites de A sont inversibles.

DÉMONSTRATION. Il est possible de montrer l'existence de la factorisation LU de manière constructive, en utilisant le procédé d'élimination de Gauss. En supposant que les n sous-matrices principales extraites de A sont inversibles, on va ici prouver en même temps l'existence et l'unicité par un raisonnement par récurrence [12].

Pour $k = 1$, on a

$$A_1 = a_{11} \neq 0,$$

et il suffit de poser $L_1 = 1$ et $U_1 = a_{11}$. Montrons à présent que s'il existe une unique factorisation de la sous-matrice A_{k-1}, $2 \leq k \leq n$, de la forme $A_{k-1} = L_{k-1}U_{k-1}$, avec $(L_{k-1})_{ii} = 1$, $i = 1, \ldots, k-1$, alors il existe une unique factorisation de ce type pour A_k. Pour cela, décomposons A_k en blocs

$$A_k = \begin{pmatrix} A_{k-1} & \boldsymbol{b} \\ \boldsymbol{c}^\top & d \end{pmatrix},$$

avec $\boldsymbol{b}$ et $\boldsymbol{c}$ des vecteurs de $\mathbb{R}^{k-1}$ et d un nombre réel, et cherchons une factorisation de A_k de la forme

$$\begin{pmatrix} A_{k-1} & \boldsymbol{b} \\ \boldsymbol{c}^\top & d \end{pmatrix} = \begin{pmatrix} L_{k-1} & \boldsymbol{0} \\ \boldsymbol{l}^\top & 1 \end{pmatrix} \begin{pmatrix} U_{k-1} & \boldsymbol{u} \\ \boldsymbol{0}^\top & \mu \end{pmatrix}$$

où $\boldsymbol{0}$ désigne le vecteur nul de $\mathbb{R}^{k-1}$, $\boldsymbol{l}$ et $\boldsymbol{u}$ sont des vecteurs de $\mathbb{R}^{k-1}$ et μ est un nombre réel. En effectuant le produit de matrices et en identifiant par blocs avec A_k, on obtient

$$L_{k-1}U_{k-1} = A_{k-1}, \ L_{k-1}\boldsymbol{u} = \boldsymbol{b}, \ \boldsymbol{l}^\top U_{k-1} = \boldsymbol{c}^\top \text{ et } \boldsymbol{l}^\top \boldsymbol{u} + \mu = d.$$

Si la première de ces égalités n'apporte aucune nouvelle information, les trois suivantes permettent de déterminer les vecteurs $\boldsymbol{l}$ et $\boldsymbol{u}$ et le scalaire μ. En effet, on a par hypothèse $0 \neq \det(A_{k-1}) = \det(L_{k-1})\det(U_{k-1})$, les matrices L_{k-1} et U_{k-1} sont donc inversibles. Par conséquent, les vecteurs $\boldsymbol{l}$ et $\boldsymbol{u}$ existent et sont uniques et $\mu = d - \boldsymbol{l}^\top \boldsymbol{u}$. Ceci achève la preuve par récurrence. □

Dans cette preuve, on utilise de manière fondamentale le fait les termes diagonaux de la matrice L sont tous égaux à 1. On aurait tout aussi bien pu choisir d'imposer d'autres valeurs (non nulles) ou encore décider de fixer les valeurs des éléments diagonaux de la matrice U. Ceci implique que plusieurs factorisations LU existent, chacune pouvant être déduite d'une autre par multiplication par une matrice diagonale convenable (voir la section 4.4.1).

On remarque également que la condition du théorème n'est que *suffisante*. Il n'est en effet pas nécessaire que la matrice A soit inversible pour que sa factorisation LU existe et soit unique (ce cas étant cependant le seul ayant vraiment un intérêt pratique). Nous laissons au lecteur le soin d'adapter et de compléter la démonstration précédente pour obtenir le résultat ci-après.

Théorème 30 (condition nécessaire et suffisante d'existence et d'unicité de la factorisation LU) *Soit A une matrice d'ordre n et de rang non nul r. La factorisation LU de A, avec $l_{ii} = 1$ pour $i = 1, \ldots, n$, existe et est unique si et seulement si les sous-matrices principales A_k d'ordre $k = 1, \ldots, r$ extraites de A sont inversibles.*

Pour toute matrice A inversible, il est possible de se ramener à la condition suffisante du théorème 29 après des échanges préalable de lignes de la matrice (comme on l'a vu lors de la description de l'élimination de Gauss avec échange). En ce sens, la factorisation LU des matrices inversibles est toujours possible. Si une stratégie de pivot partiel ou de pivot total est appliquée à l'élimination de Gauss, on a plus précisément le résultat suivant.

Théorème 31 *Soit A une matrice d'ordre n inversible. Alors, il existe une matrice P (resp. des matrices P et Q) tenant compte d'une stratégie de pivot partiel (resp. de pivot total), une matrice triangulaire inférieure L, dont les éléments sont inférieurs ou égaux à 1 en valeur absolue, et une matrice triangulaire supérieure U telles que*

$$PA = LU \quad (\textit{resp. } PAQ = LU).$$

12. Notons que ce procédé de démonstration permet aussi de prouver *directement* (c'est-à-dire sans faire appel à un résultat sur la factorisation LU) l'existence et l'unicité de la factorisation de Cholesky d'une matrice symétrique définie positive (voir le théorème 34).

4.4 Autres méthodes de factorisation

Nous présentons dans cette dernière section d'autres types de factorisation, adaptés à des matrices particulières. Il s'agit de la *factorisation LDM*$^\top$ d'une matrice carrée, qui devient la *factorisation LDL*$^\top$ lorsque cette matrice est symétrique et de la *factorisation de Cholesky*, pour une matrice *symétrique définie positive*.

4.4.1 Factorisation LDM$^\top$

Cette méthode considère une décomposition sous la forme d'un produit d'une matrice triangulaire inférieure, d'une matrice diagonale et d'une matrice triangulaire supérieure. Une fois obtenue la factorisation de la matrice A (d'un coût identique à celui de la factorisation LU), la résolution du système linéaire (4.1) fait intervenir la résolution d'un système triangulaire inférieur (par une méthode de descente), puis celle (triviale) d'un système diagonal et enfin la résolution d'un système triangulaire supérieur (par une méthode de remontée), ce qui représente un coût de n^2+n opérations.

Proposition 32 *Sous les hypothèses du théorème 29, il existe une unique matrice triangulaire inférieure L, une unique matrice diagonale D et une unique matrice triangulaire supérieure $M^\top$, les éléments diagonaux de L et M étant tous égaux à 1, telles que*

$$A = LDM^\top.$$

DÉMONSTRATION. Les hypothèses du théorème 29 étant satisfaites, on sait qu'il existe une unique factorisation LU de la matrice A. En choisissant les éléments diagonaux de la matrice D égaux à u_{ii}, $1 \leq i \leq n$, (tous non nuls puisque la matrice U est inversible), on a

$$A = LU = LDD^{-1}U.$$

Il suffit alors de poser $M^\top = D^{-1}U$ pour obtenir l'existence de la factorisation. Son unicité est une conséquence de l'unicité de la factorisation LU. □

Lorsque la matrice A considérée est inversible, la factorisation LDM$^\top$ permet également de démontrer simplement le résultat suivant, sans qu'il y ait besoin d'avoir recours au théorème 29.

Proposition 33 *Soit A une matrice carrée d'ordre n inversible admettant une factorisation LU. Alors, sa transposée $A^\top$ admet une factorisation LU.*

DÉMONSTRATION. Puisque A admet une factorisation LU, elle admet aussi une factorisation LDM$^\top$ et l'on a

$$A^\top = (LDM^\top)^\top = (M^\top)^\top D^\top L^\top = MDL^\top.$$

La matrice $A^\top$ admet donc elle aussi une factorisation LDM$^\top$ et, par suite, une factorisation LU. □

L'intérêt de la factorisation LDM$^\top$ devient clair lorsque la matrice A est symétrique, puisque $M = L$ dans ce cas. La factorisation résultante peut alors être calculée avec un coût et un stockage environ deux fois moindres que ceux d'une factorisation LU classique. Cependant, comme pour cette dernière méthode, il n'est pas conseillé [13], pour des questions de stabilité numérique, d'utiliser cette factorisation si la matrice A n'est pas symétrique définie positive ou à diagonale dominante. De manière générale, tout système linéaire pouvant être résolu au moyen de la factorisation de Cholesky (introduite dans la section 4.4.2 ci-après) peut également l'être par la factorisation LDL$^\top$ et, lorsque la matrice de ce système est une matrice bande (par exemple tridiagonale), il s'avère plus avantageux de préférer la seconde méthode, les extractions de racines carrées requises par la première représentant, dans ce cas particulier, une fraction importante du nombre d'opérations arithmétiques effectuées.

4.4.2 Factorisation de Cholesky

Une matrice symétrique définie positive vérifiant les hypothèses du théorème 29 en vertu du critère de Sylvester, elle admet une factorisation LDL$^\top$, dont la matrice D est de plus à termes diagonaux *strictement positifs*. Cette observation conduit à une factorisation ne faisant intervenir qu'une seule matrice triangulaire inférieure, appelée *factorisation de Cholesky*. Plus précisément, on a le résultat suivant.

13. Dans les autres situations, on se doit de faire appel à des stratégies de choix de pivot conservant le caractère symétrique de la matrice à factoriser, c'est-à-dire trouver une matrice de permutation P telle que la factorisation LDL$^\top$ de $PAP^\top$ soit stable. Nous renvoyons aux notes de fin de chapitre pour plus de détails sur les approches possibles.

Théorème 34 (« factorisation de Cholesky ») *Soit A une matrice symétrique définie positive. Alors, il existe une unique matrice triangulaire inférieure B, dont les éléments diagonaux sont strictement positifs, telle que*

$$A = BB^\top.$$

DÉMONSTRATION. Supposons que la matrice A est d'ordre n. On sait, par le théorème 36, que les déterminants des sous-matrices principales extraites A_k, $1 \leq k \leq n$, de A (définies par (4.8)), sont strictement positifs et les conditions du théorème 29 sont vérifiées. La matrice A admet donc une unique factorisation LU. Les éléments diagonaux de la matrice U sont de plus strictement positifs, car on a [14]

$$\prod_{i=1}^{k} u_{ii} = \det(A_k) > 0,\ 1 \leq k \leq n.$$

En introduisant la matrice diagonale Δ définie par $(\Delta)_{ii} = \sqrt{u_{ii}}$, $1 \leq i \leq n$, la factorisation se réécrit

$$A = L\Delta\Delta^{-1}U.$$

En posant $B = L\Delta$ et $C = \Delta^{-1}U$, la symétrie de A entraîne que $BC = C^\top B^\top$, d'où $C(B^\top)^{-1} = B^{-1}C^\top = I_n$ (une matrice étant triangulaire supérieure, l'autre triangulaire inférieure et toutes deux à coefficients diagonaux égaux à 1) et donc $C = B^\top$. On a donc montré l'existence d'au moins une factorisation de Cholesky. Pour montrer l'unicité de cette décomposition, on suppose qu'il existe deux matrices triangulaires inférieures B_1 et B_2 telles que

$$A = B_1{B_1}^\top = B_2{B_2}^\top,$$

d'où ${B_2}^{-1}B_1 = {B_2}^\top(B_1^\top)^{-1}$. Il existe donc une matrice diagonale D telle que ${B_2}^{-1}B_1 = D$ et, par conséquent, $B_1 = B_2D$. Finalement, on a

$$B_2{B_2}^\top = B_1{B_1}^\top = B_2DD^\top{B_2}^\top,$$

et donc $D^2 = I_n$. Les coefficients diagonaux d'une matrice de factorisation de Cholesky étant par hypothèse positifs, on a nécessairement $D = I_n$ et donc $B_1 = B_2$. □

Pour la mise en œuvre de cette factorisation, on procède de la manière suivante. On pose $B = (b_{ij})_{1\leq i,j\leq n}$ avec $b_{ij} = 0$ si $i < j$ et l'on déduit alors de l'égalité $A = BB^\top$ que

$$a_{ij} = \sum_{k=1}^{n} b_{ik}b_{jk} = \sum_{k=1}^{\min(i,j)} b_{ik}b_{jk},\ 1 \leq i,j \leq n.$$

La matrice A étant symétrique, il suffit, par exemple, que les relations ci-dessus soient vérifiées pour $j \leq i$ et l'on construit alors les colonnes de la matrice B à partir de celles de A. En fixant l'indice j à 1 et en faisant varier l'indice i de 1 à n, on trouve

$$\begin{array}{rcl@{\quad}rcl}
a_{11} &=& (b_{11})^2, & \text{d'où} & b_{11} &=& \sqrt{a_{11}},\\
a_{21} &=& b_{11}b_{21}, & \text{d'où} & b_{21} &=& \dfrac{a_{21}}{b_{11}},\\
\vdots & & & & \vdots & & \\
a_{n1} &=& b_{11}b_{n1}, & \text{d'où} & b_{n1} &=& \dfrac{a_{n1}}{b_{11}},
\end{array}$$

ce qui permet la détermination de la première colonne de B. Les coefficients de la j^{e} colonne de B, $2 \leq j \leq n$, s'obtiennent en utilisant les relations

$$\begin{array}{rcl@{\quad}rcl}
a_{jj} &=& (b_{j1})^2 + (b_{j2})^2 + \cdots + (b_{jj})^2, & \text{d'où} & b_{jj} &=& \sqrt{a_{jj} - \sum_{k=1}^{j-1}(b_{jk})^2},\\
a_{j+1j} &=& b_{j1}b_{j+11} + b_{j2}b_{j+12} + \cdots + b_{jj}b_{j+1j}, & \text{d'où} & b_{j+1j} &=& \dfrac{a_{j+1j} - \sum_{k=1}^{j-1} b_{jk}b_{j+1k}}{b_{jj}},\\
\vdots & & & & \vdots & & \\
a_{nj} &=& b_{j1}b_{n1} + b_{j2}b_{n2} + \cdots + b_{jj}b_{nj}, & \text{d'où} & b_{nj} &=& \dfrac{a_{nj} - \sum_{k=1}^{j-1} b_{jk}b_{nk}}{b_{jj}},
\end{array}$$

14. Ces égalités découlent de la preuve du théorème 29.

après avoir préalablement déterminé les $j-1$ premières colonnes, le théorème 34 assurant que les quantités sous les racines carrées sont strictement positives. En pratique, on ne vérifie d'ailleurs pas que la matrice A est définie positive, mais simplement qu'elle est symétrique, avant de débuter la factorisation. En effet, si la valeur trouvée à la k^{e} étape, $1 \leq k \leq n$, pour la quantité $(b_{kk})^2$ est négative ou nulle, c'est que A n'est pas définie positive. Au contraire, si l'algorithme de factorisation arrive à son terme, cela prouve que A est définie positive, car, pour toute matrice inversible B et tout vecteur $\boldsymbol{v}$ non nul, on a

$$\left(BB^\top \boldsymbol{v}, \boldsymbol{v}\right) = \|B^\top \boldsymbol{v}\|_2 > 0.$$

Il est à noter que le calcul du déterminant d'une matrice dont on connaît la factorisation de Cholesky est immédiat, puisque

$$\det(A) = \det(BB^\top) = (\det(B))^2 = \left(\prod_{i=1}^{n} b_{ii}\right)^2.$$

Le nombre d'opérations nécessaires pour effectuer la factorisation de Cholesky d'une matrice A symétrique définie positive d'ordre n par les formules ci-dessus est de $\frac{1}{6}(n^2-1)n$ additions et soustractions, $\frac{1}{6}(n^2-1)n$ multiplications, $\frac{1}{2}n(n-1)$ divisions et n extractions de racines carrées, soit une complexité favorable par rapport à la factorisation LU de la même matrice. Si l'on souhaite résoudre un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ associé, il faut ajouter $n(n-1)$ additions et soustractions, $n(n-1)$ multiplications et $2n$ divisions pour la résolution des systèmes triangulaires $B\boldsymbol{y} = \boldsymbol{b}$ et $B^\top \boldsymbol{x} = \boldsymbol{y}$.

Exemple d'application de la factorisation de Cholesky. Considérons la matrice symétrique définie positive

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 10 \\ 3 & 10 & 26 \end{pmatrix}.$$

En appliquant de l'algorithme de factorisation de Cholesky, on obtient

$$A = BB^\top = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 4 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{pmatrix}.$$

4.A Annexe du chapitre

Proposition 35 (« règle de Cramer ») *On suppose que les vecteurs $\boldsymbol{a}_j$, $j = 1, \ldots, n$, de $\mathbb{K}^n$ désignent les colonnes d'une matrice inversible A de $M_n(\mathbb{K})$. Les composante de la solution du système $A\boldsymbol{x} = \boldsymbol{b}$ sont données par*

$$x_i = \frac{\det(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{b}, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n)}{\det(A)},\ i = 1, \ldots, n.$$

DÉMONSTRATION. Le déterminant étant une forme multilinéaire alternée, on a

$$\forall (i,j) \in \{1, \ldots, n\}^2,\ i \neq j,\ \forall (\lambda, \mu) \in \mathbb{K}^2,\ \det(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \lambda\, \boldsymbol{a}_i + \mu\, \boldsymbol{a}_j, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n) = \lambda \det(A).$$

Or, si le vecteur $\boldsymbol{x}$ est solution de $A\boldsymbol{x} = \boldsymbol{b}$, ses composantes sont les composantes du vecteur $\boldsymbol{b}$ dans la base de $\mathbb{K}^n$ formée par les colonnes de A, c'est-à-dire

$$\boldsymbol{b} = \sum_{j=1}^{n} x_j\, \boldsymbol{a}_j.$$

On en déduit que

$$\det\left(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \textstyle\sum_{j=1}^{n} x_j\, \boldsymbol{a}_j, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n\right) = \det(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{i-1}, \boldsymbol{b}, \boldsymbol{a}_{i+1}, \ldots, \boldsymbol{a}_n) = x_i \det(A),\ i = 1, \ldots, n.$$

d'où la formule. □

Théorème 36 (« critère de Sylvester ») *Une matrice A symétrique ou hermitienne d'ordre n est définie positive si et seulement tous ses mineurs principaux dominants sont strictement positifs, c'est-à-dire si toutes les sous-matrices principales*

$$A_k = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix},\ 1 \leq k \leq n,$$

extraites de A ont un déterminant strictement positif.

DÉMONSTRATION. On démontre le théorème dans le cas réel, l'extension au cas complexe ne posant aucune difficulté, par récurrence sur l'ordre n de la matrice.

Pour $n = 1$, la matrice A est un nombre réel, $A = (a_{11})$, et, pour tout réel x, $\langle Ax, x\rangle_{M_1(\mathbb{R})} = a_{11}\, x^2$ est par conséquent positif si et seulement si $a_{11} > 0$, a_{11} étant par ailleurs le seul mineur principal.
Supposons maintenant le résultat vrai pour des matrices symétriques d'ordre $n-1$, $n \geq 2$, et prouvons-le pour celles d'ordre n. Soit A une telle matrice. On note respectivement λ_i et $\boldsymbol{v}_i$, $1 \leq i \leq n$ les valeurs et vecteurs propres de A, l'ensemble $\{\boldsymbol{v}_i\}_{1\leq i\leq n}$ formant par ailleurs une base orthonormée de $M_{n,1}(\mathbb{R})$.

Puisque $\langle A\boldsymbol{x}, \boldsymbol{x}\rangle_{M_{n,1}(\mathbb{R})} > 0$ pour tout vecteur $\boldsymbol{x}$ non nul de $M_{n,1}(\mathbb{R})$, ceci est donc en particulier vrai pour tous les vecteurs de la forme

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{pmatrix}.$$

Observons alors que

$$\left\langle A\begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{pmatrix}, \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{pmatrix} \right\rangle_{M_{n,1}(\mathbb{R})} = \left\langle A_{n-1}\begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \end{pmatrix}, \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \end{pmatrix} \right\rangle_{M_{n-1,1}(\mathbb{R})}$$

Par conséquent, la matrice A_{n-1} est définie positive et tous ses mineurs principaux dominants, qui ne sont autres que les $n-1$ mineurs principaux dominants de A, sont strictement positifs. Le fait que A soit définie positive impliquant que ses valeurs propres sont strictement positives, on a que $\det(A) = \prod_{i=1}^{n} \lambda_i > 0$ et l'on vient donc de montrer le sens direct de l'équivalence.

Réciproquement, si tous les mineurs principaux dominants de A sont strictement positifs, on applique l'hypothèse de récurrence pour en déduire que la sous-matrice A_{n-1} est définie positive. Comme $\det(A) > 0$, on a l'alternative suivante : soit toutes les valeurs propres de A sont strictement positives (et donc A est définie positive), soit au moins deux d'entre elles, λ_i et λ_j, sont strictement négatives. Dans ce dernier cas, il existe au moins une combinaison linéaire $\alpha\,\boldsymbol{v}_i + \beta\,\boldsymbol{v}_j$, avec α et β tous deux non nuls, ayant zéro pour dernière composante. Puisqu'on a démontré que A_{n-1} était définie positive, il s'ensuit que $\left\langle A(\alpha\,\boldsymbol{v}_i + \beta\,\boldsymbol{v}_j), \alpha\,\boldsymbol{v}_i + \beta\,\boldsymbol{v}_j\right\rangle_{M_{n,1}(\mathbb{R})} > 0$. Mais, on a par ailleurs

$$\left\langle A(\alpha\,\boldsymbol{v}_i + \beta\,\boldsymbol{v}_j), \alpha\,\boldsymbol{v}_i + \beta\,\boldsymbol{v}_j\right\rangle_{M_{n,1}(\mathbb{R})} = \alpha^2\lambda_i + \beta^2\lambda_j < 0,$$

d'où une contradiction. □

Chapitre 5

Méthodes itératives de résolution des systèmes linéaires

L'idée des méthodes itératives (*iterative methods* en anglais) de résolution de systèmes linéaires est de construire une suite convergente $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ de vecteurs vérifiant

$$\lim_{k\to+\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x}, \tag{5.1}$$

où $\boldsymbol{x}$ est solution du système (4.1). Dans ce chapitre, nous présentons des méthodes itératives parmi les plus simples à mettre en œuvre, à savoir les *méthodes de Jacobi*, *de Gauss–Seidel* et *de Richardson*, ainsi que leurs variantes. Dans celles-ci, partant d'un vecteur initial arbitraire $\boldsymbol{x}^{(0)}$, la suite $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ vérifie une relation de récurrence de la forme

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}, \tag{5.2}$$

la matrice carrée B, appelée *matrice d'itération* (*iteration matrix* en anglais) de la méthode, et le vecteur $\boldsymbol{c}$ dépendant de la matrice A et du second membre $\boldsymbol{b}$ du système à résoudre. Ces méthodes suivent ainsi le même principe que les méthodes de point fixe, dont l'application à la résolution d'équations non linéaires est présentée dans le chapitre 1.

Pour une matrice d'ordre n pleine, le coût de calcul induit par une telle méthode, dite *linéaire stationnaire du premier degré* [1] (*linear stationary method of first degree* en anglais), est de l'ordre de n^2 opérations à *chaque* itération. On a vu au chapitre 4 que le coût *total* d'une méthode directe pour la résolution d'un système linéaire à n équations et n inconnues est de l'ordre de n^3 opérations. Ainsi, une méthode itérative de ce type ne sera compétitive que si elle est en mesure de fournir une solution approchée [2] suffisamment précise en un nombre d'itérations indépendant de l'entier n, ou bien croissant de manière sous-linéaire avec n. Les méthodes directes pouvant cependant s'avérer coûteuses, notamment en termes d'allocation d'espace mémoire, dans certains cas particuliers (un exemple est celui des grandes matrices creuses, en raison du phénomène de remplissage évoqué dans le précédent chapitre), les méthodes itératives constituent souvent une alternative intéressante pour la résolution des systèmes linéaires.

Avant d'aborder la description des méthodes mentionnées plus haut, on donne quelques résultats généraux de convergence et de stabilité, ainsi que des principes de comparaison (en terme de *« vitesse » de convergence*), relatifs à la classe de méthodes itératives de la forme (5.2). Des résultats plus précis, traitant l'utilisation d'une méthode donnée pour la résolution d'un système linéaire possédant des propriétés ou une structure particulières, sont ensuite établis.

1. La méthode itérative est *linéaire* car ni la matrice B ni le vecteur $\boldsymbol{c}$ ne dépendent du vecteur $\boldsymbol{x}^{(k)}$ et *stationnaire* car ni B ni $\boldsymbol{c}$ ne dépendent de l'entier k. Enfin, elle est *du premier degré* car le vecteur $\boldsymbol{x}^{(k+1)}$ ne dépend explicitement que du vecteur $\boldsymbol{x}^{(k)}$.

2. On comprend en effet que, à la différence d'une méthode directe, une méthode itérative ne conduit à la solution exacte du problème qu'après avoir, en théorie, effectué un nombre *infini* d'opérations.

5.1 Généralités

Dans cette section, nous abordons quelques aspects généraux des méthodes itératives linéaires stationnaires du premier degré. Dans toute la suite, on désigne par n un entier naturel non nul et l'on considère la résolution d'un système linéaire dont la matrice est d'ordre n à coefficients complexes, les résultats énoncés restant valables *mutatis mutandis* dans le cas réel.

Nous commençons par introduire la notion fondamentale de *convergence* (*convergence* en anglais) d'une méthode itérative.

Définition 37 (convergence d'une méthode itérative linéaire stationnaire du premier degré) *On dit qu'une méthode itérative définie par* (5.2) *est* ***convergente*** *(****convergent*** *en anglais) si, pour tout choix d'initialisation* $\boldsymbol{x}^{(0)}$*, la suite* $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ *possède une limite indépendante de* $\boldsymbol{x}^{(0)}$*.*

Cette première définition ne fait aucunement intervenir le système linéaire (4.1) que l'on cherche à résoudre. Il convient par conséquent de préciser en quel sens la matrice d'itération B et le vecteur $\boldsymbol{c}$, intervenant dans la relation de récurrence définissant la méthode, sont liés à la matrice A et au second membre $\boldsymbol{b}$ de ce système. On utilise pour cela la notion de *consistance* (*consistency* en anglais) d'une méthode itérative.

Définitions 38 (consistance d'une méthode itérative linéaire stationnaire du premier degré) *Une méthode itérative définie par* (5.2) *est dite* ***consistante*** *(****consistent*** *en anglais) avec le système linéaire* (4.1) *si toute solution* $\boldsymbol{x}$ *de ce dernier satisfait l'égalité*

$$\boldsymbol{x} = B\boldsymbol{x} + \boldsymbol{c}. \tag{5.3}$$

Elle est dite ***réciproquement consistante*** *(****reciprocally consistent*** *en anglais) avec* (4.1) *si toute solution de* (5.3) *satisfait* (4.1) *et* ***complètement consistante*** *(****completely consistent*** *en anglais) avec* (4.1) *si elle est à la fois consistante et réciproquement consistante avec* (4.1)*.*

Les précédentes définitions n'exigent pas que les matrices A et $I_n - B$ soient inversibles, mais que les systèmes linéaires qui leur sont respectivement associés possèdent au moins une solution. On peut montrer qu'une méthode itérative définie par (5.2) est complètement consistante avec le système (4.1) si et seulement s'il existe une matrice M inversible telle que

$$B = I_n - M^{-1}A \text{ et } \boldsymbol{c} = M^{-1}\boldsymbol{b}.$$

Comme nous l'avons fait dans le précédent chapitre, nous supposons dorénavant que la matrice du système linéaire (4.1) est inversible. On a dans ce cas le résultat suivant.

Théorème 39 *Soit A une matrice d'ordre n inversible. Une méthode itérative définie par* (5.2) *est*

- *consistante avec le système* (4.1) *si et seulement si*

$$\boldsymbol{c} = (I_n - B)A^{-1}\boldsymbol{b},$$

- *complètement consistante avec le système* (4.1) *si et seulement si elle est consistante et la matrice* $I_n - B$ *est inversible.*

DÉMONSTRATION. La matrice A étant inversible, on a $\boldsymbol{x} = A^{-1}\boldsymbol{b}$. Si la méthode itérative est consistante, ce vecteur est solution du système (5.3), d'où $\boldsymbol{c} = (I_n - B)A^{-1}\boldsymbol{b}$. Réciproquement, si $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ et $\boldsymbol{c} = (I_n - B)A^{-1}\boldsymbol{b}$, alors $\boldsymbol{c} = (I_n - B)\boldsymbol{x}$ et la méthode est consistante.

Supposons à présent que la méthode itérative est consistante et que la matrice $I_n - B$ est inversible, la matrice A étant inversible. D'après la première assertion du théorème, on a alors $\boldsymbol{c} = (I_n - B)A^{-1}\boldsymbol{b}$, d'où $\boldsymbol{b} = A(I_n - B)^{-1}\boldsymbol{c}$. On en déduit que la méthode est réciproquement consistante (il suffit pour cela de procéder comme dans la preuve de la première assertion en échangeant A et $\boldsymbol{b}$ avec $I_n - B$ et $\boldsymbol{c}$) et donc bien complètement consistante. Réciproquement, si la méthode est complètement consistante, elle est en particulier consistante. Le système linéaire (4.1) admettant une unique solution, il va de même pour le système (5.3) et la matrice $I_n - B$ est donc inversible. □

La consistance complète d'une méthode itérative est une propriété essentielle si l'on veut que celle-ci converge vers la solution du système linéaire (4.1), comme le souligne le résultat suivant.

Théorème 40 *Soit A une matrice d'ordre n inversible,* $\boldsymbol{x}$ *l'unique vecteur solution du système linéaire* (4.1) *et une méthode itérative définie par* (5.2)*. Si la suite* $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ *construite par la méthode converge vers* $\boldsymbol{x}$ *pour toute initialisation* $\boldsymbol{x}^{(0)}$*, alors la méthode est complètement consistante. Réciproquement, si la méthode est complètement consistante et convergente, alors la suite* $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ *a pour limite* $\boldsymbol{x}$*.*

DÉMONSTRATION. Si la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ converge vers $\boldsymbol{x}$, alors, par passage à la limite dans la relation (5.2), on obtient que $\boldsymbol{x}$ est solution du système (5.3) et la méthode itérative est consistante. Supposons maintenant qu'il existe une solution $\boldsymbol{z}$ du système (5.3), différente de $\boldsymbol{x}$. Si $\boldsymbol{x}^{(0)}=\boldsymbol{z}$, on a $\boldsymbol{x}^{(1)}=\boldsymbol{x}^{(2)}=\cdots=\boldsymbol{z}$, ce qui vient contredire le fait que la méthode est convergente. Le vecteur $\boldsymbol{x}$ est donc l'unique solution de (5.3) et la méthode est complètement consistante.

D'autre part, si la méthode est convergente, la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$ a pour limite une solution du système (5.3). Si la méthode est de plus complètement consistante, elle est en particulier réciproquement consistante et $\boldsymbol{x}$ est alors l'unique solution de (5.3). □

Pour caractériser la convergence d'une méthode itérative, on utilise deux quantités particulières.

Définitions 41 (erreur et résidu d'une méthode itérative de résolution de système linéaire) *Soit k un entier naturel. On appelle **erreur** (**error** en anglais) à l'itération k d'une méthode itérative le vecteur $\boldsymbol{e}^{(k)}=\boldsymbol{x}^{(k)}-\boldsymbol{x}$, où $\boldsymbol{x}=A^{-1}\boldsymbol{b}$ est la solution de* (4.1). *Le **résidu** (**residual** en anglais) à l'itération k d'une méthode itérative est le vecteur $\boldsymbol{r}^{(k)}=\boldsymbol{b}-A\boldsymbol{x}^{(k)}$.*

On déduit des précédentes définitions qu'une méthode itérative converge si et seulement si $\lim\limits_{k\to+\infty}\boldsymbol{e}^{(k)}=\boldsymbol{0}$ (soit encore si $\lim\limits_{k\to+\infty}\boldsymbol{r}^{(k)}=-\lim\limits_{k\to+\infty}A\boldsymbol{e}^{(k)}=\boldsymbol{0}$) pour tout choix d'initialisation $\boldsymbol{x}^{(0)}$.

La seule propriété de consistance complète d'une méthode itérative ne suffisant pas à assurer qu'elle converge [3], le résultat ci-après fournit un critère fondamental de convergence.

Théorème 42 *Soit A une matrice inversible. Une méthode itérative linéaire stationnaire du premier ordre, définie par* (5.2) *et complètement consistante avec le système* (4.1)*, est convergente si et seulement si $\rho(B)<1$, où $\rho(B)$ désigne le rayon spectral de la matrice d'itération de la méthode.*

DÉMONSTRATION. Soit k un entier naturel. La méthode étant supposée complètement consistante avec (4.1), l'erreur à la $k+1^{\text{e}}$ itération vérifie la relation

$$\boldsymbol{e}^{(k+1)}=\boldsymbol{x}^{(k+1)}-\boldsymbol{x}=B\boldsymbol{x}^{(k)}-\boldsymbol{c}-(B\boldsymbol{x}-\boldsymbol{c})=B\left(\boldsymbol{x}^{(k)}-\boldsymbol{x}\right)=B\boldsymbol{e}^{(k)}.$$

Le résultat se déduit alors du théorème 56. □

En pratique, le rayon spectral d'une matrice peut être difficile à calculer, mais on déduit du théorème 55 que le rayon spectral d'une matrice B est strictement inférieur à 1 s'il existe au moins une norme matricielle pour laquelle $\|B\|<1$. L'étude de convergence des méthodes itératives de résolution de systèmes linéaires de la forme (5.2) repose donc sur la détermination de $\rho(B)$ ou, de manière équivalente, la recherche d'une norme matricielle pour laquelle $\|B\|<1$.

Une autre question se posant lorsque l'on est en présence de méthodes itératives convergentes est de savoir laquelle converge le plus rapidement. Une réponse est fournie par le résultat suivant, que l'on peut résumer ainsi : la méthode la plus « rapide » est celle dont la matrice a le plus petit rayon spectral.

Théorème 43 *Soit $\|\cdot\|$ une norme vectorielle quelconque. On considère deux méthodes itératives complètement consistantes avec le système* (4.1)*,*

$$\forall k\in\mathbb{N},\ \boldsymbol{x}^{(k+1)}=B\boldsymbol{x}^{(k)}+\boldsymbol{c}\ \textit{et}\ \tilde{\boldsymbol{x}}^{(k+1)}=\tilde{B}\tilde{\boldsymbol{x}}^{(k)}+\tilde{\boldsymbol{c}},$$

avec $\boldsymbol{x}^{(0)}=\tilde{\boldsymbol{x}}^{(0)}$ et $\rho(B)<\rho(\tilde{B})$. Alors, pour tout réel strictement positif ε, il existe un entier N tel que

$$k\geq N\implies\sup_{\substack{\boldsymbol{x}^{(0)}\in M_{n,1}(\mathbb{C})\\ \|\boldsymbol{x}^{(0)}-\boldsymbol{x}\|=1}}\left(\frac{\|\tilde{\boldsymbol{x}}^{(k)}-\boldsymbol{x}\|}{\|\boldsymbol{x}^{(k)}-\boldsymbol{x}\|}\right)^{1/k}\geq\frac{\rho(\tilde{B})}{\rho(B)+\varepsilon},$$

où $\boldsymbol{x}$ désigne la solution de (4.1).

3. Un exemple particulièrement simple de ce fait est le suivant. Soit le système linéaire

$$2I_n\boldsymbol{x}=\boldsymbol{b}$$

et la méthode itérative ayant pour relation de récurrence

$$\forall k\in\mathbb{N},\ \boldsymbol{x}^{(k+1)}=-\boldsymbol{x}^{(k)}+\boldsymbol{b}.$$

On a $A=2I_n$, $B=-I_n$ et $\boldsymbol{c}=\boldsymbol{b}=(I_n-B)A^{-1}\boldsymbol{b}$, la méthode est donc complètement consistante d'après le théorème 39. On trouve cependant que

$$\forall k\in\mathbb{N},\ \boldsymbol{x}^{(2k)}=\boldsymbol{x}^{(0)}\text{ et }\boldsymbol{x}^{(2k+1)}=-\boldsymbol{x}^{(0)}+\boldsymbol{b}.$$

La suite $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ n'est donc convergente que si $\boldsymbol{x}^{(0)}=\boldsymbol{x}=\frac{1}{2}\boldsymbol{b}$.

DÉMONSTRATION. D'après la formule de Gelfand (voir le théorème 58), étant donné un réel ε strictement positif, il existe un entier naturel N, dépendant de ε, tel que

$$\forall k \in \mathbb{N},\ k \geq N \implies \sup_{\substack{\boldsymbol{e}^{(0)} \in M_{n,1}(\mathbb{C}) \\ \|\boldsymbol{e}^{(0)}\|=1}} \|B^k \boldsymbol{e}^{(0)}\|^{1/k} \leq (\rho(B) + \varepsilon).$$

Par ailleurs, pour tout entier naturel k supérieur ou égal à N, il existe un vecteur $\boldsymbol{e}^{(0)}$, dépendant de k, tel que

$$\|\boldsymbol{e}^{(0)}\| = 1 \text{ et } \|\tilde{B}^k \boldsymbol{e}^{(0)}\|^{1/k} = \|\tilde{B}^k\|^{1/k} \geq \rho(\tilde{B}),$$

en vertu du théorème 55 et en notant également $\|\cdot\|$ la norme matricielle subordonnée à la norme vectorielle considérée. Ceci achève de démontrer l'assertion. □

On peut voir la méthode itérative (5.2) comme un cas particulier de méthode de point fixe. Il a aussi été remarqué que les méthodes itératives inventées avant les années 1950 appartenaient à une même famille (voir le tableau 5.1), obtenue en considérant une « décomposition » (on parle en anglais de *"splitting"*) de la matrice du système à résoudre de la forme

$$A = M - N, \tag{5.4}$$

avec M une matrice inversible, appelée dans certains contextes le *préconditionneur* (*preconditioner* en anglais) de la méthode, et en posant

$$\forall k \in \mathbb{N},\ M\boldsymbol{x}^{(k+1)} = N\boldsymbol{x}^{(k)} + \boldsymbol{b}. \tag{5.5}$$

Il découle des résultats donnés plus haut qu'une méthode itérative basée sur une telle décomposition est, par construction, complètement consistante avec le système (4.1) dès que la matrice A est inversible.

Pour que la dernière formule de récurrence soit d'utilité en pratique, il convient que tout système linéaire ayant M pour matrice puisse être résolu simplement et à faible coût. Les méthodes de Jacobi, de Gauss–Seidel et de Richardson stationnaire, basées sur la décomposition (5.4) et présentées ci-après, correspondent ainsi au choix d'une matrice inversible M soit diagonale, soit triangulaire inférieure. Par ailleurs, pour qu'une méthode définie par (5.5) converge, il faut (et il suffit), en vertu du théorème 42, que la valeur du rayon spectral de sa matrice d'itération, donnée par $M^{-1}N$, soit strictement inférieure à 1 et il est même souhaitable, compte tenu du théorème 43, qu'elle soit la plus petite possible.

nom de la méthode	M	N
Jacobi	D	$E+F$
sur-relaxation simultanée	$\frac{1}{\omega} D$	$\frac{1-\omega}{\omega} D + E + F$
Gauss–Seidel	$D-E$	F
sur-relaxation successive	$\frac{1}{\omega} D - E$	$\frac{1-\omega}{\omega} D + F$
Richardson stationnaire	$\frac{1}{\alpha} I_n$	$\frac{1}{\alpha} I_n - A$

TABLE 5.1: Choix des matrices M et N apparaissant dans la relation de récurrence (5.5) pour les méthodes itératives linéaires stationnaires du premier degré présentées dans ce chapitre (les matrices D, E et F apparaissant dans les expressions sont celles de la décomposition (5.7) de A et les réels ω et α sont supposés non nuls).

Plusieurs résultats de convergence, propres aux méthodes de Jacobi et de Gauss–Seidel (ainsi que leurs variantes relaxées) ou à la méthode de Richardson stationnaire, sont donnés dans la section 5.5. De manière plus générale, le résultat ci-après fournit une condition nécessaire et suffisante de convergence d'une méthode itérative associée à un choix de décomposition (5.4) d'une matrice A hermitienne [4] définie positive.

Théorème 44 (« théorème de Householder–John »[5]) *Soit A une matrice hermitienne inversible, dont la décomposition sous la forme* (5.4)*, avec M une matrice inversible, est telle que la matrice hermitienne $M^* + N$ est définie positive. On a alors $\rho(M^{-1}N) < 1$ si et seulement si A est définie positive.*

4. Comme on l'a déjà mentionné, tous les résultats énoncés le sont pour une matrice à coefficients complexes, mais restent vrais dans le cas réel. Ici, il suffit en remplaçant le mot « hermitienne » par « réelle symétrique » et la transconjugaison par la transposition.

5. La preuve de suffisance de la condition est attribuée à John, celle de sa nécessité à Householder.

DÉMONSTRATION. La matrice A (supposée d'ordre n) étant hermitienne, la matrice $M^* + N$ est effectivement hermitienne puisque

$$M^* + N = M^* + M - A = M + M^* - A^* = M + N^*.$$

Supposons la matrice A définie positive. L'application $\|\cdot\|$ de $M_{n,1}(\mathbb{C})$ dans $\mathbb{R}$ définie par

$$\forall \boldsymbol{v} \in M_{n,1}(\mathbb{C}),\ \|\boldsymbol{v}\| = (\boldsymbol{v}^* A \boldsymbol{v})^{1/2},$$

est alors une norme vectorielle euclidienne. On désigne par $\|\cdot\|$ la norme matricielle qui lui est subordonnée.

Nous allons établir que $\|M^{-1}N\| < 1$. Par définition, on a

$$\|M^{-1}N\| = \|I_n - M^{-1}A\| = \sup_{\substack{\boldsymbol{v} \in M_{n,1}(\mathbb{C}) \\ \|\boldsymbol{v}\|=1}} \|\boldsymbol{v} - M^{-1}A\boldsymbol{v}\|.$$

D'autre part, pour tout vecteur $\boldsymbol{v}$ de $M_{n,1}(\mathbb{C})$ tel que $\|\boldsymbol{v}\| = 1$, on vérifie que

$$\begin{aligned}
\|\boldsymbol{v} - M^{-1}A\boldsymbol{v}\|^2 &= (\boldsymbol{v} - M^{-1}A\boldsymbol{v})^* A (\boldsymbol{v} - M^{-1}A\boldsymbol{v}) \\
&= \boldsymbol{v}^* A \boldsymbol{v} - \boldsymbol{v}^* A (M^{-1}A\boldsymbol{v}) - (M^{-1}A\boldsymbol{v})^* A \boldsymbol{v} + (M^{-1}A\boldsymbol{v})^* A (M^{-1}A\boldsymbol{v}) \\
&= \|\boldsymbol{v}\|^2 - (M^{-1}A\boldsymbol{v})^* M^* (M^{-1}A\boldsymbol{v}) - (M^{-1}A\boldsymbol{v})^* M (M^{-1}A\boldsymbol{v}) + (M^{-1}A\boldsymbol{v})^* A (M^{-1}A\boldsymbol{v}) \\
&= 1 - (M^{-1}A\boldsymbol{v})^* (M^* + N)(M^{-1}A\boldsymbol{v}) < 1,
\end{aligned}$$

puisque la matrice $M^* + N$ est définie positive par hypothèse. La fontion de $M_{n,1}(\mathbb{C})$ dans $\mathbb{R}$ qui à $\boldsymbol{v}$ associe $\|\boldsymbol{v} - M^{-1}A\boldsymbol{v}\|$ étant continue sur le compact $\{\boldsymbol{v} \in M_{n,1}(\mathbb{C}) \mid \|\boldsymbol{v}\| = 1\}$, elle y atteint sa borne supérieure. Ceci achève la première partie de la démonstration.

Supposons à présent que $\rho(M^{-1}N) < 1$. En vertu du théorème 42, la suite $(\boldsymbol{x}^{(k)})_{k\in\mathbb{N}}$, définie par

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = M^{-1}N\boldsymbol{x}^{(k)},$$

converge vers le vecteur nul pour toute initialisation $\boldsymbol{x}^{(0)}$.

Raisonnons par l'absurde et faisons l'hypothèse que la matrice A n'est pas définie positive. Il existe alors un vecteur $\boldsymbol{x}^{(0)}$ non nul tel que $(\boldsymbol{x}^{(0)})^* A \boldsymbol{x}^{(0)} \leq 0$. Le vecteur $M^{-1}A\boldsymbol{x}^{(0)}$ étant non nul, on déduit des calculs effectués plus haut et de l'hypothèse sur $M^* + N$ que

$$(\boldsymbol{x}^{(0)})^* (A - (M^{-1}N)^* A (M^{-1}N))\boldsymbol{x}^{(0)} = (M^{-1}A\boldsymbol{x}^{(0)})^* (M^* + N)(M^{-1}A\boldsymbol{x}^{(0)}) > 0.$$

La matrice $A - (M^{-1}N)^* A (M^{-1}N)$ étant définie positive (elle est en effet congruente à $M^* + N$ qui est définie positive), on a par ailleurs

$$\forall k \in \mathbb{N},\ 0 \leq (\boldsymbol{x}^{(k)})^* (A - (M^{-1}N)^* A (M^{-1}N))\boldsymbol{x}^{(k)} = (\boldsymbol{x}^{(k)})^* A \boldsymbol{x}^{(k)} - (\boldsymbol{x}^{(k+1)})^* A \boldsymbol{x}^{(k+1)},$$

l'inégalité étant stricte pour $k = 0$, d'où

$$\forall k \in \mathbb{N}^*,\ (\boldsymbol{x}^{(k+1)})^* A \boldsymbol{x}^{(k+1)} \leq (\boldsymbol{x}^{(k)})^* A \boldsymbol{x}^{(k)} \leq \cdots \leq (\boldsymbol{x}^{(1)})^* A \boldsymbol{x}^{(1)} < (\boldsymbol{x}^{(0)})^* A \boldsymbol{x}^{(0)} \leq 0.$$

Ceci contredit le fait que $\boldsymbol{x}^{(k)}$ tend vers $\mathbf{0}$ lorsque l'entier k tend vers l'infini ; la matrice A est donc définie positive. □

5.2 Méthode de Jacobi

Observons tout d'abord que, si les coefficients diagonaux de la matrice A sont non nuls, il est possible d'isoler la i^{e} inconnue dans la i^{e} équation du système linéaire (4.1), $1 \leq i \leq n$ et l'on obtient alors le système équivalent

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x_j \right),\ i = 1, \ldots, n.$$

La méthode de Jacobi se base sur ces relations pour construire, à partir d'un vecteur initial $\boldsymbol{x}^{(0)}$ donné, une suite $\left(\boldsymbol{x}^{(k)}\right)_{k\in\mathbb{N}}$ par récurrence

$$\forall k \in \mathbb{N},\ x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij} x_j^{(k)} \right),\ i = 1, \ldots, n, \tag{5.6}$$

ce qui implique que $M = D$ et $N = E+F$ dans la décomposition (5.4) de la matrice A, où D est la matrice diagonale contenant les coefficients diagonaux de A, $d_{ij} = a_{ij}\,\delta_{ij}$, $1 \le i \le j \le n$, E est la matrice triangulaire inférieure de coefficients $e_{ij} = -a_{ij}$ si $1 \le j < i \le n$ et 0 sinon, et F est la matrice triangulaire supérieure telle que $f_{ij} = -a_{ij}$ si $1 \le i < j \le n$ et 0 sinon. Autrement dit, on a

$$A = D - (E + F), \tag{5.7}$$

la matrice d'itération de la méthode de Jacobi étant donnée par

$$B_J = D^{-1}(E + F).$$

On remarquera que la matrice diagonale D doit ici être inversible pour que la méthode soit bien définie. Cette condition n'est cependant pas très restrictive dans la mesure où l'ordre des équations et des inconnues peut être modifié. On observe par ailleurs que le calcul des composantes de la nouvelle approximation de la solution à chaque itération peut se faire de manière concomitante (on parle de calculs pouvant être effectués *en parallèle*). Pour cette raison, la méthode est aussi connue sous le nom de *méthode des déplacements simultanés* (*method of simultaneous displacements* en anglais).

Une généralisation de la méthode de Jacobi est la *méthode de sur-relaxation simultanée* ou *de sur-relaxation de Jacobi* (*simultaneous over-relaxation* ou *Jacobi over-relaxation (JOR) method* en anglais), dans laquelle un paramètre de relaxation réel non nul, noté ω, est introduit. Dans ce cas, les relations de récurrence deviennent

$$\forall k \in \mathbb{N},\ x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}\, x_j^{(k)} \right) + (1-\omega)\, x_i^{(k)},\ i = 1, \ldots, n,$$

ce qui correspond aux choix

$$M = \frac{1}{\omega}\, D,\ N = \frac{1-\omega}{\omega}\, D + E + F$$

et à la matrice d'itération

$$B_{JOR}(\omega) = \omega\, D^{-1}(E + F) + (1-\omega)\, I_n. \tag{5.8}$$

Cette méthode est consistante pour toute valeur non nulle de ω et coïncide avec la méthode de Jacobi pour $\omega = 1$. L'idée de relaxer la méthode repose sur le fait que, si l'efficacité de la méthode se mesure par la valeur du rayon spectral de la matrice d'itération, alors, en utilisant le fait que la fonction $\omega \mapsto \rho(B_{JOR}(\omega))$ est continue, on peut trouver une valeur du paramètre ω pour laquelle ce rayon est le plus petit possible, donnant ainsi une méthode itérative potentiellement plus efficace que la méthode de Jacobi. Ce type de technique s'applique également à la méthode de Gauss–Seidel (voir la prochaine section), pour laquelle il est d'ailleurs bien plus couramment employé.

L'étude des méthodes de relaxation pour un type de matrice donné consiste en général à déterminer, s'ils existent, un intervalle I de $\mathbb{R}$ ne contenant pas l'origine, tel que la méthode converge pour toute valeur de ω choisie dans I, et une valeur optimale ω_o du paramètre de relaxation telle que (dans le cas présent)

$$\rho(B_{JOR}(\omega_o)) = \inf_{\omega \in I} \rho(B_{JOR}(\omega)).$$

5.3 Méthodes de Gauss–Seidel et de sur-relaxation successive

Remarquons à présent que, lors d'un calcul *séquentiel* des composantes du vecteur $\boldsymbol{x}^{(k+1)}$ par les formules de récurrence (5.6), les premières $i-1^{\text{es}}$ composantes sont connues au moment de la détermination de i^{e} composante, si l'entier i est compris entre 2 et n. La méthode de Gauss–Seidel, parfois appelée *méthode des déplacements successifs* (*method of successive displacements* en anglais), utilise ce fait, en se servant des composantes du vecteur $\boldsymbol{x}^{(k+1)}$ déjà obtenues pour le calcul des suivantes. Ceci conduit aux relations

$$\forall k \in \mathbb{N},\ x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right),\ i = 1, \ldots, n, \tag{5.9}$$

ce qui équivaut, en utilisant la décomposition (5.7), à poser $M = D - E$ et $N = F$ dans la relation (5.4), la matrice d'itération associée à la méthode étant alors

$$B_{GS} = (D-E)^{-1}F.$$

Pour que cette méthode soit bien définie, il faut que la matrice $D-E$ soit inversible, ce qui équivaut une nouvelle fois à ce que la matrice D soit inversible. Comme on l'a vu précédemment, une condition de ce type n'est pas très restrictive en pratique si A est inversible. Par rapport à la méthode de Jacobi, la méthode de Gauss–Seidel présente l'avantage de ne pas nécessiter le stockage simultané de deux approximations successives de la solution.

Comme pour la méthode de Jacobi, on peut utiliser une version relaxée de cette méthode. On parle alors de *méthode de sur-relaxation successive* (*successive over-relaxation (SOR) method* en anglais), définie par [6]

$$\forall k \in \mathbb{N},\ x_i^{(k+1)} = \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)}\right) + (1-\omega)\, x_i^{(k)},\ i = 1, \ldots, n,$$

et dont la matrice d'itération est

$$B_{SOR}(\omega) = (D - \omega E)^{-1}((1-\omega)D + \omega F) = (I_n - \omega D^{-1}E)^{-1}\left((1-\omega)I_n + \omega D^{-1}F\right),$$

ce qui revient à poser

$$M = \frac{1}{\omega}D - E \text{ et } N = \frac{1-\omega}{\omega}D + F.$$

Cette dernière méthode est consistante pour toute valeur de ω non nulle et coïncide avec la méthode de Gauss–Seidel pour $\omega = 1$. Si $\omega > 1$, on parle de *sur-relaxation* (*over-relaxation* en anglais) et de *sous-relaxation* (*under-relaxation* en anglais) si $\omega < 1$. En pratique, il s'est avéré que la valeur optimale ω_o du paramètre était généralement plus grande que 1, d'où le nom de la méthode.

5.4 Méthode de Richardson stationnaire

On peut remarquer que la relation de récurrence (5.5) définissant les méthodes itératives vues jusqu'à présent peut encore s'écrire sous la forme « corrective » suivante

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + M^{-1}\boldsymbol{r}^{(k)}, \tag{5.10}$$

faisant intervenir le vecteur $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ qui est le résidu à l'étape k de la méthode. Dans cette dernière, le choix $M^{-1} = \alpha I_n$ avec α un réel non nul, correspondant à la décomposition $M = \frac{1}{\alpha} I_n$ et $N = \frac{1}{\alpha} I_n - A$, conduit à la méthode de Richardson *stationnaire* [7]

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha\ (\boldsymbol{b} - A\boldsymbol{x}^{(k)}), \tag{5.11}$$

de matrice d'itération

$$B_R(\alpha) = I_n - \alpha A.$$

6. Du fait du caractère nécessairement séquentiel de la méthode de Gauss–Seidel, la relaxation doit s'interpréter dans le sens suivant : en introduisant le vecteur « auxiliaire » $\boldsymbol{x}^{(k+\frac{1}{2})}$ dont les composantes sont définies par

$$\forall k \in \mathbb{N},\ a_{ii}\, x_i^{(k+\frac{1}{2})} = b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)},\ i = 1, \ldots, n,$$

on voit que les composantes de l'approximation $\boldsymbol{x}^{(k+1)}$ de la méthode SOR sont données par les moyennes pondérées

$$\forall k \in \mathbb{N},\ x_i^{(k+1)} = \omega\, x_i^{(k+\frac{1}{2})} + (1-\omega)\, x_i^{(k)},\ i = 1, \ldots, n.$$

7. Dans l'article original de Richardson, il s'avère que la valeur du paramètre α dépend de l'itération, c'est-à-dire que l'on a la relation de récurrence

$$\forall k \in \mathbb{N},\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha^{(k)}\,(\boldsymbol{b} - A\boldsymbol{x}^{(k)}),$$

qui donne lieu à une méthode itérative *instationnaire* (*non-stationary iterative method* en anglais).

5.5 Résultats de convergence

Dans toute la suite, nous faisons l'hypothèse que les coefficients diagonaux de la matrice A sont non nuls. Avant de considérer la résolution de systèmes linéaires dont les matrices possèdent des propriétés ou une structure particulières, nous commençons par donner un résultat général pour la méthode de sur-relaxation successive.

Théorème 45 (condition nécessaire de convergence pour la méthode SOR) *Le rayon spectral de la matrice de la méthode de sur-relaxation successive vérifie toujours l'inégalité*

$$\forall \omega \in]0,+\infty[,\ \rho(B_{SOR}(\omega)) \geq |\omega - 1|\,.$$

Cette méthode ne peut donc converger que si ω appartient à l'intervalle $]0,2[$.

DÉMONSTRATION. On remarque que le déterminant de la matrice $B_{SOR}(\omega)$ vaut

$$\det(B_{SOR}(\omega)) = \frac{\det\left(\frac{1-\omega}{\omega}D + F\right)}{\det\left(\frac{D}{\omega} - E\right)} = (1-\omega)^n,$$

compte tenu des structures, respectivement diagonale et triangulaires, des matrices D, E et F de la décomposition (5.7). En notant $\lambda_1, \ldots, \lambda_n$ les valeurs propres de la matrice, on en déduit alors que

$$\rho(B_{SOR}(\omega))^n \geq \prod_{i=1}^{n} |\lambda_i| = |\det(B_{SOR}(\omega))| = |1-\omega|^n\,,$$

l'égalité n'ayant lieu que lorsque toutes les valeurs propres de $B_{SOR}(\omega)$ sont de module égal à $|1-\omega|$. □

On a le résultat général suivant concernant la méthode de Richardson stationnaire.

Théorème 46 (condition nécessaire et suffisante de convergence pour la méthode de Richardson stationnaire) *La méthode de Richardson stationnaire est convergente si et seulement si*

$$\forall i \in \{1,\ldots,n\},\ \frac{2\,\mathrm{Re}(\lambda_i)}{\alpha\,|\lambda_i|^2} > 1,$$

où les scalaires $\lambda_1, \ldots, \lambda_n$, sont les valeurs propres de la matrice A du système linéaire à résoudre.

DÉMONSTRATION. Les valeurs propres de la matrice d'itération $B_R(\alpha) = I_n - \alpha A$ étant données par $1 - \alpha\,\lambda_i$, $i = 1,\ldots,n$. La condition nécessaire et suffisante de convergence équivaut alors à satisfaire les inégalités

$$\forall i \in \{1,\ldots,n\},\ |1 - \alpha\,\lambda_i| < 1,$$

soit encore

$$\forall i \in \{1,\ldots,n\},\ (1 - \alpha\,\mathrm{Re}(\lambda_i))^2 + (\alpha\,\mathrm{Im}(\lambda_i))^2 < 1,$$

dont se déduit immédiatement la condition donnée dans l'énoncé. □

5.5.1 Cas des matrices à diagonale strictement dominante

Nous avons déjà abordé le cas particulier des matrices à diagonale strictement dominante dans le cadre de leur factorisation au chapitre précédent. Pour de telles matrices, on est en mesure d'affirmer *a priori* que les méthodes itératives que l'on a introduites convergent, comme initialement établi par von Mises et Geiringer et par Collatz.

Théorème 47 *Si A est une matrice à diagonale strictement dominante par lignes, alors les méthodes de Jacobi et de Gauss–Seidel sont convergentes.*

5.5.2 Cas des matrices hermitiennes définies positives

Lorsque la matrice du système à résoudre est hermitienne définie positive, on peut établir que la condition nécessaire de convergence de la méthode de sur-relaxation successive du théorème 45 est suffisante.

Théorème 48 (« théorème d'Ostrowski–Reich ») *Si la matrice A est hermitienne à coefficients diagonaux strictement positifs, alors la méthode de sur-relaxation successive converge pour tout ω appartenant à $]0,2[$ si et seulement si A est définie positive.*

DÉMONSTRATION. On peut voir ce résultat comme un corollaire du théorème de Householder–John. En effet, la matrice A étant hermitienne, on a, à partir de (5.7), $D-E-F=D^*-E^*-F^*$, et donc $D=D^*$ et $F=E^*$, compte tenu de la définition de ces matrices. Le paramètre ω étant un réel non nul, il vient alors

$$M^*+N=\frac{D^*}{\omega}-E^*+\frac{1-\omega}{\omega}D+F=\frac{2-\omega}{\omega}D.$$

La matrice D étant définie positive par hypothèse, la matrice M^*+N est définie positive pour ω appartenant à l'intervalle $]0,2[$ et il suffit alors d'appliquer le théorème 44 pour conclure. □

Le théorème 46, traitant de la convergence de la méthode de Richardson stationnaire, peut également être complété dans ce cas.

Théorème 49 *Si la matrice A est hermitienne définie positive, alors la méthode de Richardson stationnaire converge si et seulement si*

$$0<\alpha<\frac{2}{\rho(A)}.$$

De plus, la valeur du rayon spectral de la matrice d'itération de la méthode est minimale pour le choix

$$\alpha_o=\frac{2}{\lambda_n+\lambda_1}$$

les réels strictement positifs λ_1 et λ_n étant respectivement la plus petite et la plus grande des valeurs propres de la matrice A, et vaut

$$\rho(B_R(\alpha_o))=\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}=\frac{\mathrm{cond}_2(A)-1}{\mathrm{cond}_2(A)+1},$$

où le réel $\mathrm{cond}_2(A)$ est le conditionnement de la matrice A relativement à la norme spectrale.

DÉMONSTRATION. La matrice A étant hermitienne définie positive, ses valeurs propres sont des réels strictement positifs et l'on a $\rho(B_R(\alpha))=\max\{|1-\alpha\lambda_1|,|1-\alpha\lambda_n|\}$ (voir la figure 5.1). La condition nécessaire et suffisante du théorème 46 se résume alors à

$$1-\alpha\lambda_1<1 \text{ et } 1-\alpha\lambda_n>-1,$$

d'où

$$0<\alpha<\frac{2}{\lambda_n}.$$

On conclut alors en remarquant que $\rho(A)=\lambda_n$.

Pour minimiser la valeur du rayon spectral $\rho(B_R(\alpha))$, il suffit de voir, comme observé sur la figure 5.1, que la valeur optimale α_o du paramètre α est celle l'abscisse du point d'intersection des graphes des fonctions $\alpha\mapsto|1-\alpha\lambda_1|$ et $\alpha\mapsto|1-\alpha\lambda_n|$, ce qui signifie encore que

$$\alpha_o\lambda_n-1=1-\alpha_o\lambda_1.$$

On en déduit par conséquent que

$$\alpha_o=\frac{2}{\lambda_n+\lambda_1}$$

et alors

$$\rho(B_R(\alpha_o))=1-\frac{2}{\lambda_n+\lambda_1}\lambda_1=\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}.$$

Enfin, la dernière égalité de l'énoncé est une conséquence du théorème 60. □

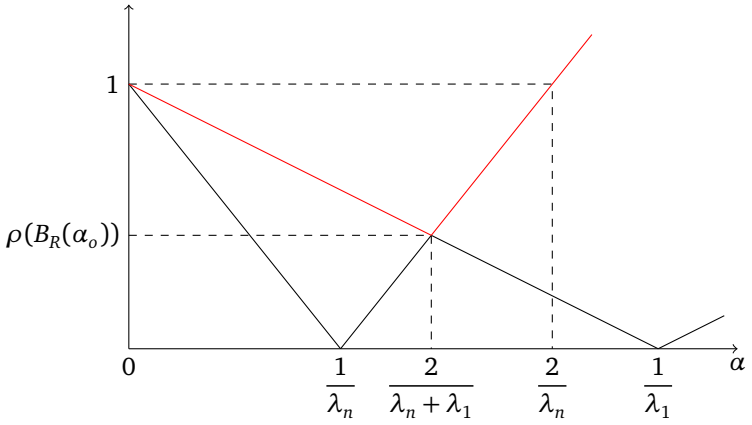

FIGURE 5.1: Graphes des fonctions $\alpha \mapsto |1-\alpha\lambda_1|$, $\alpha \mapsto |1-\alpha\lambda_n|$ et de la valeur du rayon spectral de la matrice d'itération $B_R(\alpha)$ en fonction de la valeur du paramètre α (le dernier étant tracé en rouge) dans le cas d'une matrice A hermitienne définie positive.

5.5.3 Cas des matrices tridiagonales

On en mesure comparer la convergence des méthodes de Jacobi, de Gauss–Seidel et de sur-relaxation successive dans le cas particulier des matrices tridiagonales.

Théorème 50 *Si la matrice A est tridiagonale, alors les rayons spectraux des matrices d'itération des méthodes de Jacobi et de Gauss–Seidel sont liés par la relation*

$$\rho(B_{GS}) = \rho(B_J)^2$$

de sorte que les deux méthodes convergent ou divergent simultanément. En cas de convergence, la méthode de Gauss–Seidel converge plus rapidement que celle de Jacobi.

Pour démontrer ce résultat, on a besoin d'un lemme technique.

Lemme 51 *Pour tout scalaire non nul μ, on définit la matrice tridiagonale $A(\mu)$ d'ordre n par*

$$A(\mu) = \begin{pmatrix} a_1 & \mu^{-1}c_1 & 0 & \dots & 0 \\ \mu\, b_2 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \mu^{-1}c_{n-1} \\ 0 & \dots & 0 & \mu\, b_n & a_n \end{pmatrix}. \tag{5.12}$$

Le déterminant de cette matrice ne dépend pas de μ. En particulier, on a $\det(A(\mu)) = \det(A(1))$.

DÉMONSTRATION. Les matrices $A(\mu)$ et $A(1)$ sont semblables, car si l'on introduit la matrice diagonale d'ordre n inversible (μ étant non nul)

$$Q(\mu) = \begin{pmatrix} \mu & & & \\ & \mu^2 & & \\ & & \ddots & \\ & & & \mu^n \end{pmatrix},$$

on a $A(\mu) = Q(\mu)A(1)Q(\mu)^{-1}$, d'où le résultat. □

DÉMONSTRATION DU THÉORÈME 50. Les valeurs propres de la matrice d'itération de la méthode de Jacobi $B_J = D^{-1}(E+F)$ sont les racines du polynôme caractéristique

$$\chi_{B_J}(\lambda) = \det(\lambda I_n - B_J) = \det(D^{-1})\det(\lambda D - E - F).$$

les matrices D, E et F étant celles de la décomposition (5.7). De même, les valeurs propres de la matrice d'itération de la méthode de Gauss–Seidel $B_{GS} = (D-E)^{-1}F$ sont les zéros du polynôme

$$\chi_{B_{GS}}(\lambda) = \det(\lambda I_n - B_{GS}) = \det((D-E)^{-1})\det(\lambda D - \lambda E - F).$$

Compte tenu de la structure tridiagonale de A, la matrice $A(\mu) = \lambda^2 D - \mu\lambda^2 E - \mu^{-1}F$ est bien de la forme (5.12) et l'application du lemme 51 avec le choix $\mu = \lambda^{-1}$ montre que

$$\det(\lambda^2 D - \lambda^2 E - F) = \det(\lambda^2 D - \lambda E - \lambda F) = \lambda^n \det(\lambda D - E - F),$$

d'où

$$\chi_{B_{GS}}(\lambda^2) = \frac{\det(D)}{\det(D-E)}\,\lambda^n \chi_J(\lambda) = \lambda^n \chi_J(\lambda).$$

De cette dernière relation, on déduit que, pour tout scalaire λ non nul,

$$\lambda^2 \in \sigma(B_{GS}) \iff \pm\lambda \in \sigma(B_J),$$

et donc $\rho(B_{GS}) = \rho(B_J)^2$. □

On remarque que, dans la démonstration ci-dessus, on a établi une bijection entre les valeurs propres non nulles de la matrice B_{GS} et les paires de valeurs propres opposées non nulles de la matrice B_J.

5.6 Remarques sur la mise en œuvre des méthodes

En pratique, une méthode itérative de résolution de système linéaire ne fournit qu'une *approximation* [8] de la solution du système, puisque le nombre d'itérations qu'il est possible d'effectuer est nécessairement fini. Idéalement, il conviendrait de mettre fin aux calculs à la première itération pour laquelle l'erreur est « suffisamment petite », c'est-à-dire pour le premier entier naturel k tel que

$$\|\boldsymbol{e}^{(k)}\| = \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| \leq \varepsilon,$$

où ε est une tolérance fixée et $\|\cdot\|$ est une norme donnée. Cependant, on ne sait généralement pas évaluer l'erreur, puisque la solution $\boldsymbol{x}$ n'est pas connue, et il faut donc avoir recours à un autre critère d'arrêt. Deux choix naturels s'imposent alors.

Tout d'abord, le résidu $\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}$ étant facile à calculer, on peut tester si $\|\boldsymbol{r}^{(k)}\| \leq \delta$, avec δ une tolérance fixée. Puisque l'on a

$$\|\boldsymbol{e}^{(k)}\| = \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| = \|\boldsymbol{x}^{(k)} - A^{-1}\boldsymbol{b}\| = \|A^{-1}\boldsymbol{r}^{(k)}\| \leq \|A^{-1}\|\,\|\boldsymbol{r}^{(k)}\|,$$

on voit que doit alors choisir δ tel que $\delta \leq \frac{\varepsilon}{\|A^{-1}\|}$. Ce critère peut être trompeur si la norme de A^{-1} est grande et qu'on ne dispose pas d'une bonne estimation de cette dernière. Il est en général plus judicieux de considérer dans le test d'arrêt un résidu *normalisé*,

$$\frac{\|\boldsymbol{r}^{(k)}\|}{\|\boldsymbol{r}^{(0)}\|} \leq \delta, \text{ ou encore } \frac{\|\boldsymbol{r}^{(k)}\|}{\|\boldsymbol{b}\|} \leq \delta,$$

la seconde possibilité correspondant au choix de l'initialisation $\boldsymbol{x}^{(0)} = \boldsymbol{0}$. Dans ce dernier cas, on obtient le contrôle suivant de l'erreur *relative*

$$\frac{\|\boldsymbol{e}^{(k)}\|}{\|\boldsymbol{x}\|} \leq \|A^{-1}\|\,\frac{\|\boldsymbol{r}^{(k)}\|}{\|\boldsymbol{x}\|} \leq \mathrm{cond}(A)\,\delta,$$

où $\mathrm{cond}(A)$ est le conditionnement de la matrice A relativement à la norme subordonnée à la norme $\|\cdot\|$ considérée dans le test.

Un autre critère parfois utilisé est basé sur l'*incrément* $\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$. La suite des erreurs d'une méthode itérative de la forme (5.2) vérifiant la relation de récurrence

$$\forall k \in \mathbb{N},\ \boldsymbol{e}^{(k+1)} = B\boldsymbol{e}^{(k)},$$

8. Ceci n'est pas nécessairement rédhibitoire : l'emploi d'une méthode directe (voir le chapitre 4) conduit aussi, sauf cas particulier, à une approximation de la solution dès lors que les calculs sont faits en précision finie.

on obtient, par utilisation de l'inégalité triangulaire,

$$\forall k \in \mathbb{N},\ \|\boldsymbol{e}^{(k+1)}\| \leq \|B\|\,\|\boldsymbol{e}^{(k)}\| \leq \|B\|\left(\|\boldsymbol{e}^{(k+1)}\| + \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|\right),$$

d'où

$$\forall k \in \mathbb{N},\ \|\boldsymbol{e}^{(k+1)}\| \leq \frac{\|B\|}{1-\|B\|}\,\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\|.$$

Parlons maintenant de la mise en œuvre proprement dite des méthodes présentées. Supposant qu'un test d'arrêt basé sur le résidu est employé, on considère chacune des méthodes décrites dans ce chapitre au travers de la relation de récurrence (5.5) écrite sous la forme « corrective » (5.10).

Pour l'initialisation de la méthode, on choisit habituellement, sauf si l'on dispose *a priori* d'informations sur la solution, le vecteur nul, c'est-à-dire $\boldsymbol{x}^{(0)} = \boldsymbol{0}$. Ensuite, à chaque étape de la boucle de l'algorithme, on devra réaliser les opérations suivantes :

- le calcul du résidu,
- la résolution du système linéaire ayant M pour matrice et le résidu comme second membre,
- la mise à jour de l'approximation de la solution,

jusqu'à [9] ce que la norme du résidu soit plus petite qu'une tolérance prescrite.

Le nombre d'opérations élémentaires requises à chaque itération pour un système linéaire d'ordre n se décompose en n^2 additions et soustractions et n^2 multiplications pour le calcul du résidu, n divisions (pour la méthode de Jacobi) ou $\frac{n(n-1)}{2}$ additions et soustractions, $\frac{n(n-1)}{2}$ multiplications et n divisions (pour la méthode de Gauss–Seidel) ou n multiplications (pour la méthode de Richardson stationnaire) pour la résolution du système linéaire associé à la matrice M, n additions pour la mise à jour de la solution approchée, $n-1$ additions, n multiplications et une extraction de racine carrée pour le calcul de la norme euclidienne du résidu servant au critère d'arrêt (on peut également réaliser le test directement sur la norme du résidu au carré, ce qui évite d'extraire une racine carrée). Ce compte d'opérations, de l'ordre de $\frac{1}{2}\,n^2$ additions et soustractions, $\frac{3}{2}\,n^2$ multiplications et n divisions, montre que l'utilisation d'une méthode itérative s'avère très favorable par rapport à celle d'une des méthodes directes du chapitre 4 si le nombre d'itérations à effectuer reste petit devant n.

Terminons en répétant que chaque composante de l'approximation courante de la solution peut être calculée indépendamment des autres dans la méthode de Jacobi (ou de sur-relaxation simultanée). Cette méthode est donc facilement parallélisable. Au contraire, pour la méthode de Gauss–Seidel (ou de sur-relaxation successive), ce calcul ne peut se faire que séquentiellement, mais sans qu'on ait toutefois besoin de conserver l'approximation de la solution à l'étape précédente au cours de celui-ci, ce qui constitue un (léger) gain en termes d'espace mémoire alloué à la méthode.

5.A Annexe du chapitre

5.A.1 Normes de matrices

Nous introduisons dans cette section des normes sur les espaces de matrices. En plus des propriétés habituelles d'une norme, on demande généralement qu'une norme de matrices satisfasse à une propriété de *sous-multiplicativité* qui la rend intéressante en pratique [10]. On parle dans ce cas de norme *matricielle*.

Dans toute la suite, on ne va considérer que des matrices à coefficients complexes, mais les résultats s'appliquent aussi bien à des matrices à coefficients réels, en remplaçant le cas échéant les mots « complexe », « hermitien » et « unitaire » par « réel », « réel symétrique » et « orthogonale », respectivement.

Définition 52 (norme consistante) *On dit qu'une norme $\|\cdot\|$, définie sur $M_{n,m}(\mathbb{C})$ pour toutes valeurs de m et n dans $\mathbb{N}^*$, est* ***consistante*** *si elle vérifie la propriété de* ***sous-multiplicativité (submultiplicativity** en anglais)*

$$\|AB\| \leq \|A\|\,\|B\| \tag{5.13}$$

dès que le produit de matrices AB a un sens.

9. En pratique, il est aussi nécessaire de limiter le nombre d'itérations, afin d'éliminer tout problème lié à l'absence de convergence d'une méthode.

10. Sur $M_n(\mathbb{K})$, une telle norme est alors une *norme d'algèbre*.

Définition 53 (norme matricielle) *Une **norme matricielle** (**matrix norm** en anglais) est une application de $M_{n,m}(\mathbb{C})$ dans $\mathbb{R}$, définie pour toutes valeurs de m et n dans $\mathbb{N}^*$, vérifiant les propriétés d'une norme et la propriété de sous-multiplicativité* (5.13).

Proposition 54 (norme de matrice subordonnée) *Soit m et n deux entiers naturels non nuls. Étant donné deux normes vectorielles $\|\cdot\|_\alpha$ et $\|\cdot\|_\beta$ sur $M_{m,1}(\mathbb{C})$ et $M_{n,1}(\mathbb{C})$ respectivement, l'application $\|\cdot\|_{\alpha,\beta}$ de $M_{n,m}(\mathbb{C})$ dans $\mathbb{R}$ définie par*

$$\|A\|_{\alpha,\beta} = \sup_{\substack{\boldsymbol{v}\in M_{m,1}(\mathbb{C})\\ \boldsymbol{v}\neq\mathbf{0}}} \frac{\|A\boldsymbol{v}\|_\beta}{\|\boldsymbol{v}\|_\alpha} = \sup_{\substack{\boldsymbol{v}\in M_{m,1}(\mathbb{C})\\ \|\boldsymbol{v}\|_\alpha\leq 1}} \|A\boldsymbol{v}\|_\beta = \sup_{\substack{\boldsymbol{v}\in M_{m,1}(\mathbb{C})\\ \|\boldsymbol{v}\|_\alpha=1}} \|A\boldsymbol{v}\|_\beta, \tag{5.14}$$

*est une norme de matrice dite **subordonnée** (**subordinate matrix norm** en anglais) aux normes $\|\cdot\|_\alpha$ et $\|\cdot\|_\beta$.*

DÉMONSTRATION. On remarque tout d'abord que la quantité $\|A\|_{\alpha,\beta}$ est bien définie pour tout matrice A de $M_{n,m}(\mathbb{C})$: ceci découle de la continuité de l'application de $M_{m,1}(\mathbb{C})$ dans $\mathbb{R}$ qui à un vecteur $\boldsymbol{v}$ associe $\|A\boldsymbol{v}\|_\beta$ sur la sphère unité, qui est compacte car l'espace est de dimension finie. La vérification des propriétés satisfaites par une norme est alors immédiate. □

Théorème 55 *Soit A une matrice carrée et $\|\cdot\|$ une norme matricielle. Alors, on a*

$$\rho(A) \leq \|A\|.$$

D'autre part, étant donné une matrice A et un nombre strictement positif ε, il existe au moins une norme matricielle subordonnée telle que

$$\|A\| \leq \rho(A) + \varepsilon.$$

DÉMONSTRATION. Si λ est une valeur propre de A, il existe un vecteur propre $\boldsymbol{v} \neq \mathbf{0}$ associé, tel que $A\boldsymbol{v} = \lambda\boldsymbol{v}$. Soit $\boldsymbol{w}$ un vecteur tel que la matrice $\boldsymbol{v}\boldsymbol{w}^*$ ne soit pas nulle. On a alors

$$|\lambda|\ \|\boldsymbol{v}\boldsymbol{w}^*\| = \|\lambda\,\boldsymbol{v}\boldsymbol{w}^*\| = \|A\boldsymbol{v}\boldsymbol{w}^*\| \leq \|A\|\,\|\boldsymbol{v}\boldsymbol{w}^*\|,$$

d'après la propriété de sous-multiplicativité d'une norme matricielle, et donc $|\lambda| \leq \|A\|$. Cette dernière inégalité étant vraie pour toute valeur propre de A, elle l'est en particulier quand $|\lambda|$ est égal au rayon spectral de la matrice et la première inégalité du théorème se trouve démontrée.

Supposons à présent que la matrice A est d'ordre n. Il existe une matrice unitaire U telle que $T = U^{-1}AU$ soit triangulaire (supérieure par exemple) et que les éléments diagonaux de T soient les valeurs propres de A. À tout réel $\delta > 0$, on définit la matrice diagonale D_δ telle que $d_{ii} = \delta^{i-1}$, $i = 1,\ldots,n$. Étant donné $\varepsilon > 0$, on peut choisir δ suffisamment petit pour que les éléments extradiagonaux de la matrice $(UD_\delta)^{-1}A(UD_\delta) = (D_\delta)^{-1}TD_\delta$ soient aussi petits, par exemple de façon à avoir

$$\sum_{j=i+1}^{n} \delta^{j-i}\left|t_{ij}\right| \leq \varepsilon,\ 1 \leq i \leq n-1.$$

On a alors

$$\|(UD_\delta)^{-1}A(UD_\delta)\|_\infty = \max_{1\leq i\leq n} \sum_{j=i}^{n} \delta^{j-i}\left|t_{ij}\right| \leq \rho(A) + \varepsilon.$$

Il reste à vérifier que l'application qui à une matrice B d'ordre n associe $\|(UD_\delta)^{-1}B(UD_\delta)\|_\infty$ est une norme matricielle (qui dépend de A et de ε), ce qui est immédiat puisque c'est la norme subordonnée à la norme vectorielle $\|(UD_\delta)^{-1}\cdot\|_\infty$. □

Théorème 56 *Soit A une matrice carrée. Les conditions suivantes sont équivalentes.*

i) $\lim_{k\to+\infty} A^k = 0$,

ii) $\lim_{k\to+\infty} A^k\boldsymbol{v} = \mathbf{0}$ *pour tout vecteur* $\boldsymbol{v}$,

iii) $\rho(A) < 1$,

iv) $\|A\| < 1$ *pour au moins une norme subordonnée* $\|\cdot\|$.

DÉMONSTRATION. Prouvons que i implique ii. Soit $\|\cdot\|$ une norme vectorielle et $\|\cdot\|$ la norme matricielle subordonnée lui correspondant. Pour tout vecteur $\boldsymbol{v}$, on a l'inégalité

$$\|A^k \boldsymbol{v}\| \leq \|A^k\| \, \|\boldsymbol{v}\|,$$

qui montre que $\lim_{k \to +\infty} A^k \boldsymbol{v} = \boldsymbol{0}$. Montrons ensuite que ii implique iii. Si $\rho(A) \geq 1$, alors il existe λ une valeur propre de A et $\boldsymbol{v} \neq \boldsymbol{0}$ un vecteur propre associé tels que

$$A\boldsymbol{v} = \lambda \boldsymbol{v} \text{ et } |\lambda| \leq 1.$$

La suite $(A^k \boldsymbol{v})_{k \in \mathbb{N}}$ ne peut donc converger vers $\boldsymbol{0}$, puisque $A^k \boldsymbol{v} = \lambda^k \boldsymbol{v}$. Le fait que iii implique iv est une conséquence immédiate du théorème 55. Il reste à montrer que iv implique i. Il suffit pour cela d'utiliser l'inégalité

$$\forall k \in \mathbb{N}, \ \|A^k\| \leq \|A\|^k,$$

vérifiée par la norme subordonnée de l'énoncé. □

On déduit de ce théorème un résultat sur la convergence d'une série géométrique remarquable de matrice, dite *série de Neumann*.

Corollaire 57 *Soit A une matrice carrée d'ordre n telle que $\lim_{k \to +\infty} A^k = 0$. Alors, la matrice $I_n - A$ est inversible et on a*

$$\sum_{i=1}^{+\infty} A^k = (I_n - A)^{-1}.$$

DÉMONSTRATION. On sait d'après le théorème 56 que $\rho(A) < 1$ si $\lim_{k \to +\infty} A^k = 0$, la matrice $I_n - A$ est donc inversible. En considérant l'identité

$$(I_n - A)(I_n + A + \cdots + A^k) = I_n + A^{k+1}$$

et en faisant tendre k vers l'infini, on obtient alors l'identité recherchée. □

Nous pouvons maintenant prouver le résultat suivant, qui précise un peu plus le lien existant entre le rayon spectral et la norme d'une matrice.

Théorème 58 (« formule de Gelfand ») *Soit A une matrice carrée et $\|\cdot\|$ une norme matricielle. On a*

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

DÉMONSTRATION. Puisque $\rho(A) \leq \|A\|$ d'après le théorème 55 et comme $\rho(A) = \left(\rho(A^k)\right)^{1/k}$, on sait déjà que

$$\rho(A) \leq \|A^k\|^{1/k}, \ \forall k \in \mathbb{N}.$$

Soit $\varepsilon > 0$ donné. La matrice

$$A_\varepsilon = \frac{A}{\rho(A) + \varepsilon}$$

vérifie $\rho(A_\varepsilon) < 1$ et on déduit du théorème 56 que $\lim_{k \to +\infty} A_\varepsilon^k = 0$. Par conséquent, il existe un entier l, dépendant de ε, tel que

$$k \geq l \implies \|A_\varepsilon^k\| = \frac{\|A^k\|}{(\rho(A) + \varepsilon)^k} \leq 1.$$

Ceci implique que

$$k \geq l \implies \|A_\varepsilon^k\|^{1/k} \leq \rho(A) + \varepsilon,$$

et démontre donc l'égalité cherchée. □

5.A.2 Conditionnement d'une matrice

La résolution d'un système linéaire par les méthodes numériques des chapitres 4 et 5 est sujette à des erreurs d'arrondis dont l'accumulation peut détériorer notablement la précision de la solution obtenue. Afin de mesurer la sensibilité de la solution $\boldsymbol{x}$ d'un système linéaire $A\boldsymbol{x} = \boldsymbol{b}$ par rapport à des perturbations des données A et $\boldsymbol{b}$, on utilise une quantité appelée *conditionnement*.

Définition 59 (conditionnement d'une matrice) *Soit $\|\cdot\|$ une norme matricielle. Pour toute matrice inversible A d'ordre n, on appelle* ***conditionnement de*** *A* ***relativement à la norme*** *$\|\cdot\|$ le nombre*

$$\mathrm{cond}(A) = \|A\|\,\|A^{-1}\|.$$

La valeur du conditionnement d'une matrice dépendant en général de la norme matricielle choisie, on a coutume de signaler celle-ci en ajoutant un indice dans la notation, par exemple $\mathrm{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$. On note que l'on a toujours $\mathrm{cond}(A) \geq 1$ pour une norme matricielle subordonnée (et $\mathrm{cond}_F(A) \geq \sqrt{n}$), puisque $\|A\|\,\|A^{-1}\| \geq \|AA^{-1}\| = \|I_n\| = 1$ (et $\|I_n\|_F = \sqrt{n}$). D'autres propriétés évidentes du conditionnement sont rassemblées dans le résultat suivant.

Théorème 60 *Soit A une matrice inversible d'ordre n.*

1. *On a $\mathrm{cond}(A) = \mathrm{cond}(A^{-1})$ et $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$ pour tout scalaire α non nul.*
2. *On a*
$$\mathrm{cond}_2(A) = \frac{\mu_n}{\mu_1},$$
où μ_1 et μ_n désignent respectivement la plus petite et la plus grande des valeurs singulières de A.
3. *Si A est une matrice normale, on a*
$$\mathrm{cond}_2(A) = \frac{\max\limits_{1\leq i\leq n} |\lambda_i|}{\min\limits_{1\leq i\leq n} |\lambda_i|} = \rho(A)\rho(A^{-1}),$$
où les scalaires λ_i, $1 \leq i \leq n$, sont les valeurs propres de A.
4. *Si A est une matrice unitaire ou orthogonale, son conditionnement $\mathrm{cond}_2(A)$ vaut 1.*
5. *Le conditionnement $\mathrm{cond}_2(A)$ est invariant par transformation unitaire (ou orthogonale),*
$$UU^* = I_n \implies \mathrm{cond}_2(A) = \mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(U^*AU).$$

DÉMONSTRATION.

1. Les égalités découlent de la définition du conditionnement et des propriétés de la norme.
2. On a $\|A\|_2 = \sqrt{\rho(A^*A)}$ et, d'après la définition des valeurs singulières de A, on a donc $\|A\|_2 = \mu_n$. Par ailleurs, on voit que
$$\|A^{-1}\|_2 = \sqrt{\rho((A^{-1})^*A^{-1})} = \sqrt{\rho(A^{-1}(A^{-1})^*)} = \sqrt{\rho((A^*A)^{-1})} = \frac{1}{\mu_1},$$
ce qui démontre le résultat.
3. La propriété résulte de l'égalité $\|A\|_2 = \rho(A)$ vérifiée par les matrices normales.
4. Le résultat découle de l'égalité $\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(I_n)} = 1$.
5. La propriété est une conséquence de l'invariance par transformation unitaire de la norme $\|\cdot\|_2$.

□

Chapitre 6

Calcul de valeurs et de vecteurs propres

Nous abordons dans ce chapitre le problème du calcul de valeurs propres et, éventuellement, de vecteurs propres d'une matrice d'ordre n diagonalisable. C'est un problème beaucoup plus difficile que celui de la résolution d'un système linéaire. En effet, les valeurs propres d'une matrice étant les racines de son polynôme caractéristique, on pourrait naïvement penser qu'il suffit de factoriser ce dernier pour les obtenir. On sait cependant (par le théorème d'Abel–Ruffini) qu'il n'est pas toujours possible d'exprimer les racines d'un polynôme de degré supérieur ou égal à cinq à partir des coefficients du polynôme et d'opérations élémentaires (addition, soustraction, multiplication, division et extraction de racines). Par conséquent, il ne peut exister de méthode directe, c'est-à-dire fournissant le résultat en un nombre fini d'opérations, de calcul de valeurs propres d'une matrice et on a recours à des méthodes itératives.

Parmi ces méthodes, il convient distinguer celles qui permettent le calcul d'*une* valeur propre (en général celle de plus grand ou de plus petit module, mais pas seulement) de celles qui conduisent à une approximation de l'*ensemble* du spectre d'une matrice. D'autre part, certaines méthodes permettent le calcul de vecteurs propres associés aux valeurs propres obtenues, alors que d'autres non. C'est le cas par exemple de la *méthode de la puissance*, qui fournit une approximation d'un couple particulier de valeur et vecteur propres.

6.1 Exemple d'application : PageRank

Lorsqu'un internaute formule une requête sur le site d'un moteur de recherche, ce dernier interroge la base de données dont il dispose pour y répondre. Cette base est construite par une indexation des ressources collectées par des robots qui explorent systématiquement la « Toile » (*World Wide Web* en anglais) en suivant récursivement tous les hyperliens qu'ils trouvent. Le traitement de la requête consiste alors en l'application d'un algorithme identifiant (via l'index) les documents qui correspondent le mieux aux mots apparaissant dans la requête, afin de présenter les résultats de recherche par ordre de pertinence supposée.

De nombreux procédés existent pour améliorer la qualité des réponses fournies par un moteur de recherche. Le plus connu d'entre eux est certainement la technique *PageRank*™ employée par Google, qui est une méthode de calcul d'un *indice de notoriété*, associé à chaque page de la « Toile » et servant à affiner le classement préalablement obtenu. Celle-ci est basée sur l'idée simple que plus une page d'un ensemble est la cible d'hyperliens, plus elle est susceptible de posséder un contenu pertinent. Le but de cette sous-section est de donner les grandes lignes de la modélisation et de la théorie mathématiques sur lesquelles s'appuie PageRank [1] et leur lien avec une des méthodes de calcul numérique de valeurs et vecteurs propres présentées dans ce chapitre.

On peut considérer la « Toile » comme un ensemble de n pages (avec n un entier naturel aujourd'hui extrêmement grand) reliées entre elles par des hyperliens, représentable par un graphe orienté dont les nœuds symbolisent les pages et les arcs les hyperliens. La figure 6.1 présente un minuscule échantillon d'un tel graphe.

À ce graphe orienté, on associe une matrice d'adjacence, définie comme la matrice C d'ordre n dont le coefficient c_{ij}, $1 \leq i, j \leq n$, est égal à un s'il existe un hyperlien sur la j^{e} page pointant vers la i^{e} page et vaut zéro sinon (on ignore les hyperliens « internes », c'est-à-dire ceux pointant d'une page vers elle-même, de sorte que l'on a

1. On notera que, en plus d'être formé des mots anglais *page*, qui signifie page, et *rank*, qui signifie rang, ce nom est aussi lié à celui de l'un des inventeurs de cette méthode, Larry Page.

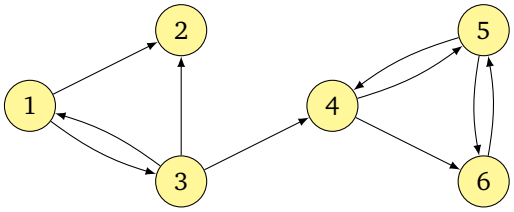

FIGURE 6.1: Graphe orienté représentant un échantillon de « toile » constitué de six pages.

$c_{ii} = 0$, $i = 1, \ldots, n$). Pour l'exemple de graphe de la figure 6.1, la matrice d'adjacence est

$$C = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Pour toute page qui en contient, on suppose que chacun des hyperliens possède la même [2] chance d'être suivi. En pondérant les coefficients de la matrice d'adjacence du graphe orienté pour tenir compte de ce fait, on obtient une matrice $\tilde{Q}$ d'ordre n qui, pour l'exemple introduit plus haut est égale à

$$\tilde{Q} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Cette modification de la matrice d'adjacence fait apparaître des similitudes avec la matrice de transition d'une *chaîne* [3] *de Markov* à temps discret et n états, qui est une *matrice stochastique* [4]. Cependant, la structure de la matrice construite à partir des seuls liens réellement existant laisse entrevoir un obstacle à une telle identification. En effet, lorsqu'une page ne possède aucun hyperlien (en anglais, on nomme *dangling node* le nœud du graphe correspondant), la ligne qui lui est associée dans la matrice $\tilde{Q}$ est nulle et cette dernière ne peut dans ce cas être stochastique. Un remède [5] est de remplacer toutes les lignes en question par $\frac{1}{n}\boldsymbol{e}^T$, avec $\boldsymbol{e}$ la matrice colonne de $M_{n,1}(\mathbb{R})$ dont les coefficients sont tous égaux à un, conduisant à la matrice (stochastique) Q. Pour l'exemple de la figure 6.1, on obtient ainsi

$$Q = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

2. On peut en fait utiliser n'importe quelle autre distribution de probabilité que la distribution uniforme, obtenue par exemple en se basant sur des statistiques disponibles pour les pages concernées.

3. Une chaîne de Markov est un processus stochastique vérifiant une propriété caractérisant une « absence de mémoire ». De manière simplifiée, la meilleure prévision que l'on puisse faire, connaissant le passé et le présent, du futur d'un tel processus est identique à la meilleure prévision qu'on puisse faire de ce futur connaissant uniquement le présent.

4. Une matrice stochastique est une matrice carrée dont tous les éléments sont des réels positifs compris entre 0 et 1 et dont la somme des éléments d'une même ligne (ou d'une même colonne) est égale à 1. Le spectre d'une matrice stochastique est contenu dans le disque unité.

5. Là encore, on peut utiliser n'importe quelle distribution de probabilité autre que la distribution uniforme. Une des premières modifications suggérées du modèle de base a été l'utilisation d'un vecteur « personnalisé » $\boldsymbol{v}$, tel que $v_i \geq 0$, $\forall i \in \{1, \ldots, n\}$, et $\sum_{i=1}^{n} v_i = 1$, différenciant les internautes en fonction de leurs habitudes ou de leurs goûts, pour éventuellement proposer des classements adaptés à l'utilisateur qui formule la requête.

On définit alors l'indice de notoriété d'une page comme la fraction de temps passé sur cette page lorsque le temps tend vers l'infini, correspondant encore à la probabilité fournie par la *distribution stationnaire* (ou *invariante*) de la chaîne de Markov considérée, donnée par un vecteur propre *à gauche*, dont les composantes sont positives et de somme égale à un, associé à la valeur propre (dominante) égale à 1 de la matrice de transition de cette chaîne. Si le *théorème de Perron–Frobenius* assure qu'un tel vecteur existe toujours pour une matrice stochastique, il faut pour garantir son unicité (ce qui revient à demander que la valeur propre 1 soit de multiplicité simple) que la chaîne de Markov en question soit *irréductible* [6]. Pour que ce soit le cas, on suppose qu'un internaute peut passer de la page sur laquelle il se trouve à

- l'une des pages vers lesquelles mènent les hyperliens qu'elle contient avec une probabilité de valeur $\alpha > 0$, que l'on appelle le *facteur d'amortissement* (*damping factor* en anglais),
- toute page de la « Toile » avec une probabilité uniforme, valant $\frac{1-\alpha}{n}$.

Cette dernière hypothèse conduit à considérer [7] la matrice stochastique suivante

$$A = \alpha Q + \frac{1-\alpha}{n} \boldsymbol{e}\boldsymbol{e}^T,$$

pour laquelle il existe un unique vecteur $\boldsymbol{\pi}$ de $M_{n,1}(\mathbb{R})$ tel que

$$\boldsymbol{\pi}^T A = \boldsymbol{\pi}^T,\ \forall i \in \{1,\dots,n\},\ \pi_i \geq 0,\ \text{et}\ \sum_{i=1}^{n} \pi_i = 1.$$

L'indice de notoriété de la i^{e} page est ainsi donné par la valeur de la composante π_i. Le vecteur $\boldsymbol{\pi}$ étant également un vecteur propre (à droite) associé à la valeur propre dominante de la matrice A^T, Google a recours à la méthode de la puissance (voir la section 6.2), qui ne nécessite que d'effectuer des produits entre la matrice A^T et des vecteurs donnés. Cependant, la taille de la « Toile » est si gigantesque qu'il est impossible de stocker la matrice A^T ou d'effectuer de manière conventionnelle son produit avec un vecteur. Pour rendre la méthode applicable, il faut observer que la matrice Q est très creuse [8] et tirer parti de cette structure particulière en ne stockant que ses éléments non nuls, tout en adaptant les algorithmes de multiplication matricielle. Même en réalisant ces optimisations, la puissance de calcul requise reste considérable et nécessite des serveurs informatiques adaptés à cette tâche, que Google effectuerait apparemment chaque mois.

Parlons pour finir de la valeur du facteur d'amortissement α. On peut montrer que la valeur propre sous-dominante de la matrice A est égale à α si la matrice Q n'est pas la matrice de transition d'une chaîne de Markov irréductible, strictement inférieure à α si c'est le cas [9]. La vitesse de convergence de la méthode de la puissance se trouve par conséquent directement affectée par le choix de cette valeur. Il y a alors un compromis délicat à trouver entre une convergence rapide de la méthode (c'est-à-dire α choisi proche de 0) et un modèle rendant assez fidèlement compte de la structure de la « Toile » et du comportement des internautes (c'est-à-dire α choisi proche de 1). La valeur utilisée par les fondateurs de Google est $\alpha = 0,85$.

6.2 Méthode de la puissance

La méthode de la puissance (*power (iteration) method* en anglais) est certainement la méthode la plus simple fournissant une approximation de la valeur propre de plus grand module d'une matrice et d'un vecteur propre associé. Après l'avoir présentée et avoir analysé sa convergence, nous verrons comment, par des modifications adéquates, elle peut être utilisée pour calculer quelques autres couples de valeur et vecteur propres de la même matrice. Dans toute la suite, on considère une matrice A de $M_n(\mathbb{C})$ diagonalisable et on note λ_j, $j = 1,\dots,n$, ses valeurs propres (comptées avec leurs multiplicités algébriques respectives) et $\boldsymbol{v}_j$, $j = 1,\dots,n$, des vecteurs propres associés. On suppose de plus que les valeurs propres de A sont ordonnées de la manière suivante

$$|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_n|. \tag{6.1}$$

6. Une chaîne de Markov est dite irréductible si tout état est accessible à partir de n'importe quel autre état.
7. Dans le cas, évoqué plus haut, d'une personnalisation de la recherche par l'utilisation d'un vecteur $\boldsymbol{v}$, on aura $A = \alpha Q + (1-\alpha)\boldsymbol{e}\boldsymbol{v}^T$.
8. Typiquement, on a seulement de trois à dix termes non nuls sur chacune des lignes de cette matrice.
9. Ce résultat reste vrai en cas d'utilisation d'un vecteur de personnalisation $\boldsymbol{v}$.

6.2.1 Approximation de la valeur propre de plus grand module

Faisons l'hypothèse que la valeur propre λ_n est de multiplicité algébrique égale à 1 et que la dernière des inégalités de (6.1) est une inégalité stricte. On dit alors que λ_n est la valeur propre *dominante* de A. Pour l'approcher, on peut considérer une méthode itérative, appelée méthode de la puissance. Celle-ci consiste, à partir d'un vecteur initial unitaire $\boldsymbol{q}^{(0)}$ arbitraire, en le calcul des termes des suites définies par

$$\forall k \in \mathbb{N}^*,\ \boldsymbol{z}^{(k)} = A\boldsymbol{q}^{(k-1)},\ \boldsymbol{q}^{(k)} = \frac{\boldsymbol{z}^{(k)}}{\|\boldsymbol{z}^{(k)}\|_2} \text{ et } \nu^{(k)} = (\boldsymbol{q}^{(k)})^* A\boldsymbol{q}^{(k)}. \tag{6.2}$$

Analysons les propriétés de la suite $(\boldsymbol{q}^{(k)})_{k\in\mathbb{N}}$. Par une simple récurrence sur l'indice k, on vérifie que

$$\forall k \in \mathbb{N},\ \boldsymbol{q}^{(k)} = \frac{A^k\boldsymbol{q}^{(0)}}{\|A^k\boldsymbol{q}^{(0)}\|_2}, \tag{6.3}$$

et l'on voit alors plus clairement le rôle joué par les puissances de la matrice A, qui donnent son nom à la méthode. En effet, l'ensemble $\{\boldsymbol{v}_j\}_{j=1,\ldots,n}$ des vecteurs propres de A formant une base de l'espace, le vecteur $\boldsymbol{q}^{(0)}$ peut s'écrire comme la combinaison linéaire suivante

$$\boldsymbol{q}^{(0)} = \sum_{j=1}^{n} \alpha_j \boldsymbol{v}_j,$$

et l'on a alors, en supposant le coefficient α_n non nul,

$$\forall k \in \mathbb{N},\ A^k\boldsymbol{q}^{(0)} = \sum_{j=1}^{n} \alpha_j (A^k\boldsymbol{v}_j) = \sum_{j=1}^{n} \alpha_j {\lambda_j}^k \boldsymbol{v}_j = \alpha_n {\lambda_n}^k \left(\boldsymbol{v}_n + \sum_{j=1}^{n-1} \frac{\alpha_j}{\alpha_n} \left(\frac{\lambda_j}{\lambda_n}\right)^k \boldsymbol{v}_j \right). \tag{6.4}$$

Comme $\left|\frac{\lambda_j}{\lambda_n}\right| < 1$ pour tout entier j dans $\{1,\ldots,n-1\}$, la composante du vecteur $\boldsymbol{q}^{(k)}$ le long de $\boldsymbol{v}_n$ augmente en module avec l'entier k relativement aux composantes le long des autres directions $\boldsymbol{v}_j$, $j = 1,\ldots,n-1$. En supposant que les vecteurs de la base $\{\boldsymbol{v}_j\}_{j=1,\ldots,n}$ sont de norme euclidienne unitaire, il vient alors

$$\left\| \sum_{j=1}^{n-1} \frac{\alpha_j}{\alpha_n} \left(\frac{\lambda_j}{\lambda_n}\right)^k \boldsymbol{v}_j \right\|_2 \leq \sum_{j=1}^{n-1} \left|\frac{\alpha_j}{\alpha_n}\right| \left|\frac{\lambda_j}{\lambda_n}\right|^k \leq \left|\frac{\lambda_{n-1}}{\lambda_n}\right|^k \sum_{j=1}^{n-1} \left|\frac{\alpha_j}{\alpha_n}\right| = C \left|\frac{\lambda_{n-1}}{\lambda_n}\right|^k,$$

et l'on déduit de (6.3), de (6.4) et de cette dernière inégalité que

$$\forall k \in \mathbb{N},\ \boldsymbol{q}^{(k)} = \frac{\alpha_n {\lambda_n}^k (\boldsymbol{v}_n + \boldsymbol{w}^{(k)})}{\|\alpha_n {\lambda_n}^k (\boldsymbol{v}_n + \boldsymbol{w}^{(k)})\|_2},$$

où la suite de vecteurs $(\boldsymbol{w}^{(k)})_{k\in\mathbb{N}^*}$ a pour limite le vecteur nul. Le vecteur $\boldsymbol{q}^{(k)}$ devient donc peu à peu colinéaire avec le vecteur propre $\boldsymbol{v}_n$ associé à la valeur propre dominante λ_n quand k tend vers l'infini et ce d'autant plus rapidement que le rapport $\left|\frac{\lambda_{n-1}}{\lambda_n}\right|$ est petit, ce qui correspond à des valeurs propres dominante et sous-dominante bien séparées. La suite des *quotients de Rayleigh*

$$\forall k \in \mathbb{N},\ (\boldsymbol{q}^{(k)})^* A\boldsymbol{q}^{(k)} = \nu^{(k)},$$

converge donc vers la valeur propre λ_n, et on a démontré le résultat suivant.

Théorème 61 *Soit n un entier naturel supérieur ou égal à 2 et A une matrice diagonalisable d'ordre n, dont les valeurs propres satisfont*

$$|\lambda_1| \leq \cdots \leq |\lambda_{n-1}| < |\lambda_n|.$$

On suppose que le vecteur initial $\boldsymbol{q}^{(0)}$ de la méthode de la puissance (6.2) n'est pas strictement contenu dans le sous-espace vectoriel engendré par les vecteurs propres associés aux valeurs propres autres que la valeur dominante λ_n. Alors, la méthode converge [10] et sa vitesse de convergence est d'autant plus rapide que le module du rapport entre λ_{n-1} et λ_n est petit.

10. Pour une valeur propre dominante λ_n *réelle*, la convergence a lieu au sens où l'on a

$$\lim_{k\to+\infty} \nu^{(k)} = \lambda_n \text{ et } \lim_{k\to+\infty} \boldsymbol{q}^{(k)} = \boldsymbol{v}_n \text{ si } \lambda_n > 0 \text{ ou } \lim_{k\to+\infty} (-1)^k \boldsymbol{q}^{(k)} = \boldsymbol{v}_n \text{ si } \lambda_n < 0,$$

le vecteur $\boldsymbol{v}_n$ étant un vecteur propre unitaire associé à λ_n.

Bien qu'elle soit difficile à vérifier quand on ne dispose d'aucune information *a priori* sur le sous-espace propre associé à λ_n, on remarquera que l'hypothèse faite sur le vecteur initial $\boldsymbol{q}^{(0)}$ n'est pas très contraignante en pratique, car, même si celui-ci est bel et bien strictement contenu dans $\mathrm{Vect}\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{n-1}\}$, il est probable que, du fait des erreurs d'arrondi, l'un des vecteurs $\boldsymbol{q}^{(k)}$, avec k un entier naturel non nul, aura une « petite » composante dans la direction de $\boldsymbol{v}_n$, ce qui entraînera alors la convergence de la méthode. Habituellement, on observe ce phénomène après quelques itérations. Par ailleurs, on voit, au moyen de l'expression (6.4), que la suite de vecteurs $(A^k\boldsymbol{q}^{(0)})_{k\in\mathbb{N}}$ n'est, en général, pas convergente. C'est la raison pour laquelle on choisit de travailler avec la suite de vecteurs unitaires $(\boldsymbol{q}^{(k)})_{k\in\mathbb{N}}$.

Si la valeur propre λ_n n'est pas de multiplicité simple tout en étant néanmoins la seule valeur propre de plus grand module de A, on a encore convergence de la suite $(\nu^{(k)})_{k\in\mathbb{N}}$ vers λ_n, alors que la suite $(\boldsymbol{q}^{(k)})_{k\in\mathbb{N}}$ converge vers un élément du sous-espace propre associé. En revanche, s'il existe plusieurs valeurs propres de plus grand module, la méthode de la puissance ne converge généralement pas. Dans le cas de deux valeurs propres dominantes complexes conjuguées, *i.e.*, $\lambda_{n-1}=\overline{\lambda_n}$, la suite $(\boldsymbol{q}^{(k)})_{k\in\mathbb{N}}$ présente notamment un comportement oscillatoire non amorti. Indiquons que supposer l'existence d'une unique valeur propre dominante n'est en principe pas une hypothèse restrictive, puisque l'on peut montrer qu'une matrice possède des valeurs propres de modules distincts de manière générique [11].

Enfin, si la matrice A n'est pas diagonalisable, la convergence de la méthode n'est pas assurée.

Indiquons que la mise en œuvre de la méthode de la puissance ne nécessite que de calculer des produits scalaires entre différents vecteurs, ainsi que des produits entre la matrice A et les vecteurs de la suites $(\boldsymbol{q}^{(k)})_{k\in\mathbb{N}}$. Il n'est en particulier pas obligatoire de stocker la matrice A sous la forme d'un tableau, ce qui peut être particulièrement intéressant lorsque celle-ci est creuse et de grande taille (voir la sous-section 6.1 pour un exemple). Dans le cas d'une matrice d'ordre n quelconque, le coût d'une itération de la méthode sera de n^2 multiplications et $n(n-1)$ additions pour effectuer le produit matrice-vecteur, n^2+n multiplications, $n-1$ additions, une division et une extraction de racine carrée pour la normalisation du vecteur courant, n^2 multiplications et $n-1$ additions pour le calcul du quotient de Rayleigh.

6.2.2 Déflation

Pour approcher d'autres valeurs propres que celle de plus grand module, on peut utiliser une technique, nommée *déflation*, consistant en la construction, à partir d'une matrice donnée, d'une matrice possédant le même spectre, à l'exception d'une valeur propre choisie qui se trouve remplacée par 0. Il devient alors possible, en appliquant la méthode de la puissance à la matrice résultant de la déflation par la valeur propre dominante λ_n dèjà trouvée, d'obtenir une suite convergeant vers la valeur propre ayant le *deuxième* plus grand module, et ainsi de suite...

Le procédé de déflation le plus simple, dit *de Hotelling*, demande, dans le cas d'une matrice A générale, de connaître une valeur propre λ_j, l'entier j appartenant à $\{1,\ldots,n\}$, et un vecteur $\boldsymbol{v}_j$ associé, issu d'une base $\{\boldsymbol{v}_i\}_{i=1,\ldots,n}$ formée de vecteurs propres, ainsi que le vecteur $\boldsymbol{u}_j$ correspondant de la *base duale* de $\{\boldsymbol{v}_i\}_{i=1,\ldots,n}$, c'est-à-dire le vecteur vérifiant [12]

$$\forall i\in\{1,\ldots,n\},\ (\boldsymbol{u}_j)^*\boldsymbol{v}_i=\delta_{ij}.$$

En considérant la perturbation de rang un suivante de la matrice A,

$$A_j=A-\lambda_j\,\boldsymbol{v}_j(\boldsymbol{u}_j)^*,$$

il vient

$$\forall i\in\{1,\ldots,n\},\ A_j\boldsymbol{v}_i=A\boldsymbol{v}_i-\lambda_j\,\boldsymbol{v}_j(\boldsymbol{u}_j)^*\boldsymbol{v}_i=\lambda_i\,\boldsymbol{v}_i-\lambda_j\,\boldsymbol{v}_j\delta_{ij},$$

ce qui correspond bien à la modification annoncée du spectre, les vecteurs propres associés restant inchangés. Lorsque la matrice A est symétrique ou hermitienne et que ses valeurs propres sont deux à deux distinctes, on

11. En topologie, une propriété est dite *générique* si elle est vraie sur un ensemble ouvert dense ou, plus généralement, sur un ensemble *résiduel* (c'est-à-dire un ensemble contenant une intersection dénombrable d'ouverts denses).

12. Le spectre de la matrice adjointe de A étant constitué des conjugués des valeurs propres de A, on voit que le vecteur $\boldsymbol{u}_j$ est un vecteur propre de A^*, associé à la valeur propre $\overline{\lambda_j}$ et normalisé de manière convenable. On a en effet

$$\forall i\in\{1,\ldots,n\},\ \lambda_j(\boldsymbol{u}_j)^*\boldsymbol{v}_i=(\overline{\lambda_j}\,\boldsymbol{u}_j)^*\boldsymbol{v}_i(A^*\boldsymbol{u}_j)^*\boldsymbol{v}_i=(\boldsymbol{u}_j)^*A\boldsymbol{v}_i=(\boldsymbol{u}_j)^*(A\boldsymbol{v}_i)=\lambda_i(\boldsymbol{u}_j)^*\boldsymbol{v}_i.$$

notera que l'on a seulement besoin que de connaître la valeur propre λ_j et un vecteur propre $\boldsymbol{v}_j$ associé, puisque l'on peut dans ce cas définir la matrice A_j par

$$A_j = A - \lambda_j \frac{\boldsymbol{v}_j (\boldsymbol{v}_j)^*}{(\boldsymbol{v}_j)^* \boldsymbol{v}_j}$$

en vertu de la propriété d'orthogonalité des vecteurs propres.

Utilisée en conjonction avec la méthode de la puissance, cette technique permette en théorie d'approcher l'*ensemble* du spectre d'une matrice donnée. Cependant, les valeurs et vecteurs obtenus successivement n'étant en pratique que des approximations des véritables valeurs et vecteurs propres, l'accumulation des erreurs et la mauvaise stabilité numérique de la déflation la rendent difficilement utilisable pour le calcul de plus de deux ou trois valeurs propres.

6.2.3 Méthode de la puissance inverse

On peut facilement adapter la méthode de la puissance pour le calcul de la valeur propre de plus *petit* module, que l'on a noté λ_1, d'une matrice A *inversible* : il suffit de l'appliquer à l'inverse de A, dont la plus *grande* valeur propre en module est ${\lambda_1}^{-1}$. On parle dans ce cas de *méthode de la puissance inverse* (*inverse power method* en anglais).

De manière plus générale, cette variante permet d'approcher la valeur propre de la matrice A la *plus proche* d'un nombre μ donné n'appartenant pas à son spectre. Considérons en effet la matrice $(A-\mu I_n)^{-1}$ dont les valeurs propres sont $(\lambda_i - \mu)^{-1}$, $i = 1, \ldots, n$, et supposons qu'il existe un entier m tel que

$$\forall i \in \{1, \ldots, n\} \setminus \{m\},\ |\lambda_m - \mu| < |\lambda_i - \mu|\,,$$

ce qui revient à supposer que la valeur propre λ_m qui est la plus proche de μ a une multiplicité algébrique égale à 1 (en particulier, si $\mu = 0$, λ_m sera la valeur propre de A ayant le plus petit module). L'application de la méthode de la puissance inverse pour le calcul de λ_m se résume alors à la construction, étant donné un vecteur initial unitaire $\boldsymbol{q}^{(0)}$ arbitraire, des suites définies par les relations de récurrence suivantes

$$\forall k \in \mathbb{N}^*,\ (A-\mu I_n)\boldsymbol{z}^{(k)} = \boldsymbol{q}^{(k-1)},\ \boldsymbol{q}^{(k)} = \frac{\boldsymbol{z}^{(k)}}{\|\boldsymbol{z}^{(k)}\|_2} \text{ et } \nu^{(k)} = (\boldsymbol{q}^{(k)})^* A \boldsymbol{q}^{(k)}. \tag{6.5}$$

Les vecteurs propres de la matrice $A - \mu I_n$ étant ceux de la matrice A, le quotient de Rayleigh ci-dessus fait simplement intervenir A et non $A - \mu I_n$. Le nombre μ peut être vu comme un paramètre permettant de « décaler » (on parle en anglais de *shift*) le spectre de la matrice A de manière à pouvoir approcher toute valeur propre de A dont on possède une estimation *a priori*. De ce point de vue, la méthode de la puissance inverse se prête donc particulièrement bien au raffinement d'une approximation grossière d'une valeur propre. Par rapport à la méthode de la puissance (6.2), il faut cependant résoudre, à chaque itération de la méthode, un système linéaire (ayant pour matrice $A - \mu I_n$) pour obtenir les vecteurs de la suite $(\boldsymbol{z}^{(k)})_{k\in\mathbb{N}}$. En pratique, on réalise une fois pour toutes la factorisation LU (voir la section 4.3) de cette matrice au début du calcul de manière à n'effectuer par la suite que la résolution de deux systèmes linéaires triangulaires, pour un coût de l'ordre de n^2 opérations, à chaque étape.

S'il est souhaitable que la valeur du paramètre μ soit aussi voisine que possible de la valeur propre λ_m pour que la convergence soit rapide, il faut néanmoins qu'il n'en soit pas proche au point de rendre la matrice $A - \mu I_n$ *numériquement singulière* (cette dernière notion, liée à la présence d'erreurs d'arrondi, étant essentiellement empirique).