Chapitre III

Méthodes itératives

Dans les deux chapitres qui précèdent, nous avons considéré divers problèmes : optimisation linéaire, résolution de systèmes linéaires, décomposition triangulaire d'une matrice inversible (ou d'une matrice qui s'en déduit par une permutation de lignes), détermination des coefficients du polynôme caractéristique d'une matrice, détermination du nombre de racines réelles d'un polynôme à coefficients réels contenues dans un intervalle donné, ... Toutes les méthodes étudiées alors aboutissent à la solution du problème considéré après un nombre fini d'opérations, et la solution ainsi déterminée est exacte (aux erreurs d'arrondi près, dues à la précision nécessairement finie des opérations arithmétiques portant sur des nombres réels). De telles méthodes sont dites *directes*.

Nous allons maintenant étudier des méthodes d'une nature différente, appelées *méthodes itératives*, qui consistent à déterminer successivement les termes d'une suite de solutions approchées du problème considéré. Lorsque l'application de la méthode est justifiée, la suite ainsi formée converge vers la solution exacte du problème. Le premier terme de la suite est choisi de manière plus ou moins arbitraire, puis chacun des termes suivants est calculé à partir du terme précédent (ou, parfois, de plusieurs termes précédents). L'opération qui consiste à calculer un terme de plus de la suite de solutions approchées est appelée *itération*. En pratique, on effectue seulement un nombre fini d'itérations; on s'arrête lorsqu'on estime que la dernière solution approchée obtenue est suffisamment proche de la solution exacte pour l'usage qu'on veut en faire.

Par leur principe même, les méthodes itératives font intervenir la notion de *convergence* d'une suite, qui est une notion topologique. C'est pourquoi nous allons rappeler brièvement quelques notions de topologie, relatives aux espaces vectoriels normés.

1. Espaces métriques, espaces vectoriels normés

1.1. Définition. — *On appelle distance sur un ensemble non vide E une application $d : E \times E \to \mathbb{R}^+$ qui vérifie les propriétés suivantes :*

(i) *Propriété de symétrie : pour tous x et $y \in E$,*
$$d(x, y) = d(y, x) .$$

(ii) *Inégalité du triangle : pour tous x, y et $z \in E$,*
$$d(x, z) \leq d(x, y) + d(y, z) .$$

(iii) *Deux éléments x et y de E vérifient $d(x, y) = 0$ si et seulement si $x = y$.*

L'ensemble E muni de la distance d est appelé espace métrique et noté (E, d).

1.2. La deuxième inégalité du triangle. — Soit (E, d) un espace métrique, x, y et z trois points de E. D'après l'inégalité du triangle, nous avons

$$d(x,y) \leq d(x,z) + d(z,y) \qquad \text{et} \qquad d(z,y) \leq d(z,x) + d(x,y)\,.$$

Compte tenu de la symétrie de d, nous en déduisons

$$d(x,y) - d(z,y) \leq d(x,z) \qquad \text{et} \qquad d(z,y) - d(x,y) \leq d(x,z)\,,$$

d'où nous déduisons l'inégalité très souvent utile, appelée *deuxième inégalité du triangle*,

$$d(x,z) \geq \big|d(x,y) - d(z,y)\big|\,.$$

1.3. Définitions. — *Soit (E, d) un espace métrique.*

1. *Soit x un point de E et r un réel strictement positif. On appelle boule ouverte de centre x et de rayon r, et on note $B(x, r)$, l'ensemble des points y de E qui vérifient $d(x, y) < r$.*
2. *On appelle partie ouverte (ou ouvert) de E, la réunion d'une famille quelconque de boules ouvertes de E. En particulier, la partie vide (réunion d'une famille vide) est un ouvert de E.*
3. *Soit x un point de E. On appelle voisinage du point x toute partie V de E, telle qu'il existe un ouvert U de E vérifiant $x \in U$ et $U \subset V$.*
4. *Soit $(x_n\,,\ n \in \mathbb{N})$ une suite d'éléments de E. On dit que cette suite converge vers un élément l de E, ou que l'élément l de E est limite de la suite $(x_n\,,\ n \in \mathbb{N})$, si pour tout voisinage V de l, il existe un entier N tel que pour tout $n \geq N$, on ait $x_n \in V$.*
5. *Soit $(x_n\,,\ n \in \mathbb{N})$ une suite d'éléments de E. On dit que cette suite est de Cauchy si pour tout $\varepsilon > 0$, il existe un entier N tel que, pour tous n et m vérifiant $n \geq N$ et $m \geq N$, ou ait $d(x_n, x_m) \leq \varepsilon$.*

1.4. Commentaires. — Les hypothèses sont celles des définition précédentes.

a) Il est facile de vérifier qu'une partie U de E est ouverte si et seulement si, pour tout point x de U, il existe $\varepsilon > 0$ tel que $B(x, \varepsilon) \subset U$. Il est facile aussi de vérifier qu'étant donné un point x de E, une partie V de E est voisinage de x si et seulement s'il existe $\varepsilon > 0$ tel que $B(x, \varepsilon) \subset V$. On peut aisément en déduire qu'une partie U de E est ouverte si et seulement si elle est voisinage de chacun de ses points.

b) Soit $(x_n\,,\ n \in \mathbb{N})$ une suite d'éléments de E. On vérifie aisément que cette suite converge vers un élément l de E si et seulement si, pour tout $\varepsilon > 0$, il existe un entier n tel que, pour tout $n \geq N$, on ait $d(x_n, l) < \varepsilon$. On vérifie également que lorsqu'une suite converge, l'élément l de E vers lequel elle converge est unique.

c) On montre qu'une suite $(x_n\,,\ n \in \mathbb{N})$ d'éléments de E qui converge vers un élément l de E est de Cauchy. La propriété réciproque n'est en général pas vraie; les espaces métriques pour lesquels cette réciproque est vraie (définis formellement ci-dessous) sont dits *complets*.

1.5. Définition. — *Un espace métrique (E, d) est dit complet si dans cet espace, toute suite de Cauchy converge (vers un élément de cet espace).*

1.6. Remarque. — La notion d'espace métrique complet est très importante car, dans un tel espace, on peut prouver qu'une suite converge sans avoir besoin de connaître d'avance l'élément limite vers lequel elle converge : il suffit de vérifier que cette suite est de Cauchy, ce qu'on peut faire en étudiant les distances mutuelles des termes de cette suite.

1.7. Définition. — *Soit E un espace vectoriel sur le corps $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Une norme sur E est une application de E dans $\mathbb{R}^+$, notée $x \mapsto \|x\|$, qui vérifie les conditions suivantes :*

(i) *pour tous $\lambda \in \mathbb{K}$, $x \in E$, on a $\|\lambda x\| = |\lambda| \|x\|$;*
(ii) *pour tous x et $y \in E$, on a $\|x + y\| \leq \|x\| + \|y\|$;*
(iii) *un élément $x \in E$ vérifie $\|x\| = 0$ si et seulement si $x = 0$.*

Un espace vectoriel muni d'une norme est appelé espace vectoriel normé.

1.8. Distance associée à une norme. — La donnée d'une norme $x \mapsto \|x\|$ sur un espace vectoriel E détermine, sur cet espace, une distance d, donnée par la formule (dans laquelle x et y sont deux éléments de E),

$$d(x, y) = \|x - y\| .$$

On dit que cette distance est *associée* à la norme considérée. Un espace vectoriel normé est donc un espace métrique, car on convient toujours de le munir de la distance associée à sa norme.

1.9. Définition. — *Un espace de Banach est un espace vectoriel normé complet (pour la distance associée à sa norme).*

1.10. Définition. — *Deux normes $x \mapsto \|x\|_1$ et $x \mapsto \|x\|_2$ sur un espace vetoriel E sont dites équivalentes s'il existe deux réels $k > 0$ et $k' > 0$ tels que l'on ait, pour tout $x \in E$,*

$$k' \|x\|_1 \leq \|x\|_2 \leq k \|x\|_1 .$$

1.11. Commentaires

a) *Relation d'équivalence.* — Soit E un espace vectoriel. La propriété, pour deux normes sur cet espace, d'être équivalentes, est une *relation d'équivalence* sur l'ensemble des normes sur E. Cela signifie que cette relation vérifie les propriétés suivantes :

(i) une norme est équivalente à elle-même;
(ii) (symétrie); une norme $x \mapsto \|x\|_1$ est équivalente à une norme $x \mapsto \|x\|_2$ si et seulement si la norme $x \mapsto \|x\|_2$ est équivalente à la norme $x \mapsto \|x\|_1$;
(iii) (transitivité); si une norme $x \mapsto \|x\|_1$ est équivalente à une norme $x \mapsto \|x\|_2$ et si la norme $x \mapsto \|x\|_2$ est équivalente à une troisième norme $x \mapsto \|x\|_3$, alors la norme $x \mapsto \|x\|_1$ est équivalente à la norme $x \mapsto \|x\|_3$.

b) *Conséquences de l'équivalence de deux normes.* — Soit E un espace vectoriel muni d'une norme $x \mapsto \|x\|_1$, donc aussi d'une distance $(x, y) \mapsto d_1(x, y) = \|x - y\|_1$. On peut donc définir, sur cet espace, les ouverts, les suites convergentes, les suites de Cauchy. Remplaçons la norme $x \mapsto \|x\|_1$ par une autre norme, $x \mapsto \|x\|_2$. Il est facile de vérifier que la famille des ouverts de E pour la distance associée à la nouvelle norme est la même que la famille des ouverts de E pour la distance associée à l'ancienne norme si et seulement si les deux normes sont équivalentes. Lorsque c'est le cas, toute suite qui converge pour la distance associée à l'une de ces normes converge aussi pour la distance associée à l'autre norme; de même, toute suite qui est de Cauchy pour la distance associée à une de ces normes est aussi de Cauchy pour la distance associée à l'autre norme; par suite, si l'espace est de Banach pour une des normes, il l'est aussi pour l'autre.

Le commentaire qui précède et le très important théorème qui suit montrent que l'apparent arbitraire lié au choix d'une norme particulière est sans conséquence lorsque l'espace vectoriel considéré est de dimension finie.

1.12. Théorème. — *Sur un espace vectoriel de dimension finie, toutes les normes sont équivalentes.*

Preuve : Soit E un espace vectoriel sur le corps $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, de dimension finie n. Soit $(e_1, \ldots, e_n)$ une base de E. Tout élément x de E s'exprime, de manière unique, sous la forme $x = \sum_{i=1}^{n} x^i e_i$. Les éléments x^i de $\mathbb{K}$, $1 \leq i \leq n$, sont les *composantes* de x dans la base $(e_1, \ldots, e_n)$. Posons

$$\|x\| = \sum_{i=1}^{n} |x^i| .$$

Il est facile de vérifier que l'application $x \mapsto \|x\|$ ainsi définie est une norme sur E. Muni de la topologie associée à cette norme, l'espace E est homéomorphe à $\mathbb{K}^n$ muni de sa topologie usuelle. Par suite, toute partie de E fermée et bornée, pour la topologie et la distance associées à la norme $x \mapsto \|x\|$, est compacte.
Soit $p : E \to \mathbb{R}^+$ une autre norme sur E. Compte tenu des propriétés d'une norme on a, pour tout $x \in E$,

$$p(x) = p\left(\sum_{i=1}^{n} x^i e_i\right) \leq \sum_{i=1}^{n} |x^i| p(e_i) \leq \left(\sup_{1 \leq j \leq n} p(e_j)\right) \|x\| .$$

Pour tout couple de points x et y de E, on a (deuxième inégalité du triangle)

$$|p(x) - p(y)| \leq p(x - y) \leq \left(\sup_{1 \leq j \leq n} p(e_j)\right) \|x - y\| ,$$

ce qui prouve que l'application $p : E \to \mathbb{R}^+$ est continue (et même lipschitzienne) lorsqu'on munit E de la topologie asociée à la norme $x \mapsto \|x\|$.
Soit S la sphère de rayon 1 centrée à l'origine, pour la norme $x \mapsto \|x\|$:

$$S = \left\{ x \in E \ ; \ \ \|x\| = 1 \right\} .$$

La partie S de E est fermée et bornée pour la norme $x \mapsto \|x\|$ et la topologie qui lui est associée. D'après ce qui précède, elle est compacte. La restriction à S de la fonction continue p est donc bornée et atteint ses bornes. Par suite, sa borne inférieure est strictement positive, puisque p est une norme. Posons donc

$$m = \inf_{z \in S} p(z)\,.$$

On a, pour tout $x \in E$, $x \neq 0$,

$$p(x) = \|x\|\, p\left(\frac{1}{\|x\|}\, x\right) \geq \|x\|\, m\,.$$

En résumé on a, pour tout $x \in E$,

$$m\|x\| \leq p(x) \leq \left(\sup_{1 \leq j \leq n} p(e_j)\right) \|x\|\,,$$

ce qui prouve que la norme p est équivalente à la norme $x \mapsto \|x\|$. L'équivalence des normes étant une relation d'équivalence, on conclut que toutes les normes sur E sont équivalentes. □

1.13. Conséquence. — Tout espace vectoriel de dimension finie, muni d'une norme quelconque, est complet. En effet d'après le théorème précédent, toutes les normes sur un tel espace sont équivalentes, et un espace vectoriel de dimension finie n sur le corps $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$ s'identifie à $\mathbb{R}^n$ ou à $\mathbb{C}^n$, qui sont complets.

Tout sous-espace vectoriel d'un espace vectoriel de dimension finie est de dimension finie, donc complet, donc fermé.

2. Applications linéaires continues

2.1. Proposition. — *Soient E et F deux espaces vectoriels normés sur le corps $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, et $f : E \to F$ une application linéaire. Les propriétés suivantes sont équivalentes :*

(i) *l'application f est continue en un point x_0 de E ;*

(ii) *l'application f est continue (en tout point de E) ;*

(iii) *il existe $M \geq 0$ tel que, pour tout $x \in E$,*

$$\|f(x)\| \leq M\|x\| \ ;$$

(iv) *l'application f est uniformément continue.*

Lorsque ces propriétés sont satisfaites, on pose

$$\|f\| = \sup_{x \in E,\ \|x\| \leq 1} \|f(x)\|\,.$$

On a alors aussi

$$\begin{aligned}\|f\| &= \sup_{x \in E,\ x \neq 0} \frac{\|f(x)\|}{\|x\|} \\ &= \sup_{x \in E,\ \|x\| = 1} \|f(x)\| \\ &= \inf\big\{\, M \ ;\ \ M \in \mathbb{R}^+\,,\ \forall x \in E\,,\ \|f(x)\| \leq M\|x\| \,\big\}\,.\end{aligned}$$

On a, pour tout $x \in E$,

$$\|f(x)\| \leq \|f\| \, \|x\| \, .$$

Preuve : Supposons f continue en $x_0 \in E$. Pour tout $\varepsilon > 0$, il existe donc $\eta > 0$ tel que pour tout $z \in E$, $\|z - x_0\| \leq \eta$ implique $\|f(z) - f(x_0)\| \leq \varepsilon$. Montrons que f est continue en un autre point x de E. Pour tout $y \in E$, on a, puisque f est linéaire,

$$f(y) - f(x) = f(y + x_0 - x) - f(x_0) \, .$$

D'autre part $(y + x_0 - x) - x_0 = y - x$. Par suite, $y \in E$, $\|y - x\| \leq \eta$, implique $\|f(y) - f(x)\| \leq \varepsilon$, ce qui prouve que f est continue en x.
Supposons f continue, donc en particulier continue à l'origine. Il existe donc $\eta > 0$ tel que pour tout $z \in E$ vérifiant $\|z\| \leq \eta$, $\|f(z)\| \leq 1$. Pour tout $x \in E$ vérifiant $x \neq 0$, on peut écrire

$$x = \frac{\|x\|}{\eta} z \, , \quad \text{avec } z = \frac{\eta}{\|x\|} \, x \, , \quad \|z\| = \eta \, .$$

Par suite,

$$f(x) = \frac{\|x\|}{\eta} \, f(z) \, ,$$

d'où

$$\|f(x)\| = \frac{\|x\|}{\eta} \, \|f(z)\| \leq \frac{1}{\eta} \, \|x\| \, .$$

Cette dernière inégalité est trivialement vérifiée lorsque $x = 0$. L'application f vérifie donc la propriété (iii) avec $M = \eta^{-1}$.
Supposons la propriété (iii) vérifiée. Puisque f est linéaire on a, pour tous x et $y \in E$,

$$\|f(x) - f(y)\| = \|f(x - y)\| \leq M \|x - y\| \, .$$

Cela exprime que f est M-lipschitzienne, donc uniformément continue.
Si f est uniformément continue, elle est évidemment continue en tout point, donc la propriété (i) est satisfaite. On a donc bien prouvé l'équivalence des propriétés (i) à (iv).
Enfin, l'équivalence des diverses expressions de $\|f\|$, et l'inégalité indiquée à la fin de l'énoncé, se vérifient sans difficulté. □

2.2. Proposition. — *Soient E et F deux espaces vectoriels normés sur le corps $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. L'ensemble, noté $\mathcal{L}(E, F)$, des applications linéaires continues de E dans F, est un espace vectoriel sur $\mathbb{K}$, et l'application $f \mapsto \|f\|$ définie dans la proposition 2.1 est une norme sur cet espace. Cette norme sur $\mathcal{L}(E, F)$ est dite associée aux normes dont les espaces E et F sont munis. Si l'espace F est complet, $\mathcal{L}(E, F)$ l'est aussi.*

Preuve : Les premières assertions de l'énoncé ($\mathcal{L}(E, F)$ est un espace vectoriel, et $f \mapsto \|f\|$ est une norme sur cet espace) se vérifient immédiatement. Montrons que si F

est complet, $\mathcal{L}(E,F)$ est complet. Soit $(f_n\,,\ n\in\mathbb{N})$ une suite de Cauchy dans $\mathcal{L}(E,F)$. Pour tout $x\in E$, l'inégalité

$$\|f_m(x)-f_n(x)\|\leq\|f_m-f_n\|\|x\|$$

montre que la suite $\big(f_n(x)\,,\ n\in\mathbb{N}\big)$ est de Cauchy. Puisque F est complet, cette suite converge; soit $f(x)$ sa limite. On a ainsi défini une application $f:E\to F$. On a pour tous $n\in\mathbb{N}$, x et $y\in E$, $\lambda\in\mathbb{K}$,

$$f_n(x+y)=f_n(x)+f_n(y)\ ;\qquad f_n(\lambda x)=\lambda f_n(x)\,.$$

En faisant tendre n vers $+\infty$ dans ces égalités, on voit que f est linéaire.
Puisqu'elle est de Cauchy, la suite (f_n) est bornée; il existe donc un réel $M\geq 0$ tel que, pour tous $n\in\mathbb{N}$ et $x\in E$,

$$\|f_n(x)\|\leq M\|x\|\,.$$

En faisant tendre n vers $+\infty$ dans cette inégalité, on voit que f est continue et que $\|f\|\leq M$.
Soit $\varepsilon>0$. Puisque la suite (f_n) est de Cauchy, il existe $N\in\mathbb{N}$ tel que pour tous n et $m\in\mathbb{N}$ vérifiant $n\geq N$, $m\geq N$, et tout $x\in E$, on ait

$$\|f_n(x)-f_m(x)\|\leq\varepsilon\|x\|\,.$$

Faisons tendre m vers $+\infty$. On voit que pour tout $n\geq N$ et tout $x\in E$,

$$\|f_n(x)-f(x)\|\leq\varepsilon\|x\|\,,$$

donc

$$\|f_n-f\|\leq\varepsilon\,,$$

ce qui montre que la suite (f_n) converge vers f, pour la norme qu'on a définie sur $\mathcal{L}(E,F)$.
□

2.3. Proposition. — *Soient E, F et G trois espaces vectoriels normés. Soient $f\in\mathcal{L}(E,F)$ et $g\in\mathcal{L}(F,G)$. Alors $g\circ f\in\mathcal{L}(E,G)$, et on a*

$$\|g\circ f\|\leq\|g\|\,\|f\|\,.$$

Preuve : On a, pour tout $x\in E$,

$$\big\|g\big(f(x)\big)\big\|\leq\|g\|\,\big\|f(x)\big\|\leq\|g\|\,\|f\|\,\|x\|\,,$$

d'où le résultat. □

2.4. Théorème. — *Toute application linéaire d'un espace vectoriel de dimension finie dans un espace vectoriel normé quelconque est continue.*

Preuve : Soit f une application linéaire d'un espace vectoriel E de dimension finie n dans un espace vectoriel normé F. Soit $(e_1,\dots,e_n)$ une base de E. Puisque toutes les normes

sur E sont équivalentes, on peut choisir, par exemple, la norme qui, à un point x de E, de composantes $x^1, \dots, x^n$ dans la base $(e_1, \dots, e_n)$, associe

$$\|x\| = \sup_{1 \le i \le n} |x^i| .$$

On peut alors écrire, pour tout $x \in E$,

$$f(x) = \sum_{i=1}^{n} x^i f(e_i) ,$$

d'où

$$\|f(x)\| \le \left(\sum_{j=1}^{n} \|f(e_j)\| \right) \sup_{1 \le i \le n} |x^i| = \left(\sum_{j=1}^{n} \|f(e_j)\| \right) \|x\| ,$$

ce qui montre que f est continue. □

2.5. Remarque. — Il est facile de donner des exemples d'applications linéaires non continues d'un espace vectoriel normé E de dimension infinie dans un autre espace vectoriel normé F. Prenons pour E l'espace vectoriel des polynômes à coefficients réels, muni de la norme

$$\|P\| = \sup_{x \in [0,1]} \big|P(x)\big| .$$

En remarquant qu'un polynôme non identiquement nul n'a qu'un nombre fini de racines, le lecteur vérifiera aisément que cette expression définit bien une norme sur E. Soit $a \in \mathbb{R}$, et $f_a : E \to \mathbb{R}$ l'application définie par

$$f_a(P) = P(a) .$$

On montre aisément que f_a est continue si et seulement si a est élément de l'intervalle $[0, 1]$.

3. Normes et rayon spectral sur un espace de matrices

3.1. Normes sur un espace de matrices carrées. — On considère l'espace vectoriel $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ des matrices $n \times n$ à coefficients dans le corps $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Cet espace, qui s'identifie à $\mathbb{K}^{n^2}$, peut être muni de diverses normes, toutes équivalentes puisqu'il est de dimension finie n^2 sur le corps $\mathbb{K}$.

En particulier, on peut munir $\mathbb{K}^n$ d'une norme $x \mapsto \|x\|$, puis munir $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ de la norme associée, au sens de la proposition 2.2. Rappelons que cette norme, notée $A \mapsto \|A\|$, vérifie, pour tout $x \in \mathbb{K}^n$,

$$\|Ax\| \le \|A\| \, \|x\| .$$

Parmi les normes couramment utilisées sur $\mathbb{K}^n$, citons celles définies par les formules, dans lesquelles $x = (x^1, \dots, x^n) \in \mathbb{K}^n$,

$$\|x\|_1 = \sum_{i=1}^{n} |x^i| , \quad \|x\|_2 = \left(\sum_{i=1}^{n} |x^i|^2 \right)^{1/2} , \quad \|x\|_\infty = \sup_{1 \le i \le n} |x^i| ,$$

ou encore par la formule, dans laquelle p est un réel vérifiant $p \geq 1$,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x^i|^p \right)^{1/p} .$$

Ces normes sont bien sûr toutes équivalentes. La norme $x \mapsto \|x\|_2$ est souvent appelée *norme euclidienne*. À chacune de ces normes est associée une norme sur l'espace de matrices $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$.
Mais il existe bien d'autres normes sur $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ que celles associées à une norme sur $\mathbb{K}^n$.
Remarquons que si une norme $A \mapsto \|A\|$ sur l'espace de matrices $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ est associée à une norme $x \mapsto \|x\|$ sur $\mathbb{K}^n$, la norme de la matrice unité I_n est $\|I_n\| = 1$.

3.2. Définition. — *Une norme $A \mapsto \|A\|$ sur $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ est dite multiplicative, ou matricielle, si pour tous A et $B \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$, on a*

$$\|AB\| \leq \|A\| \, \|B\| .$$

3.3. Exemples. — Pour toute norme $x \mapsto \|x\|$ sur $\mathbb{K}^n$, la norme associée sur $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ est multiplicative. Cela résulte en effet immédiatement de la proposition 2.3.
Mais il existe des normes sur $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ qui sont multiplicatives sans pour autant être associées à une norme sur $\mathbb{K}^n$. C'est, par exemple, le cas de la *norme de Frobenius*,

$$\|A\|_F = \left(\mathrm{Trace}({}^t\overline{A}\, A) \right)^{1/2} = \left(\sum_{(i,j)} |a_{i\,j}|^2 \right)^{1/2} .$$

On a noté $\overline{A} = (\overline{a_{i\,j}})$ la matrice complexe conjuguée de A, si $\mathbb{K} = \mathbb{C}$, et $\overline{A} = A$ si $\mathbb{K} = \mathbb{R}$.
On rappelle que le signe ^{t}, placé à gauche d'une matrice, désigne la transposition.
Pour prouver ce résultat, remarquons d'abord que la norme de Frobenius, définie par la formule ci-dessus, est bien une norme sur $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$, puisque c'est la norme euclidienne sur cet espace lorsqu'on convient de l'identifier à $\mathbb{K}^{n^2}$.
Soient $A = (a_{i\,j})$ et $B = (b_{i,j})$ deux éléments de $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$. Nous avons, compte tenu de l'inégalité de Schwarz,

$$\begin{aligned}
\|AB\|_F^2 &= \sum_{(i,j)} \left| \sum_{k=1}^{n} a_{i\,k} b_{k\,j} \right|^2 \\
&\leq \sum_{(i,j)} \left(\sum_{k=1}^{n} |a_{i\,k}|^2 \right) \left(\sum_{l=1}^{n} |b_{l\,j}|^2 \right) \\
&\leq \left(\sum_{(i,k)} |a_{i\,k}|^2 \right) \left(\sum_{(l,j)} |b_{l\,j}|^2 \right) \\
&\leq \|A\|_F^2 \|B\|_F^2 .
\end{aligned}$$

Enfin, la norme de Frobenius sur $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ n'est pas associée à une norme sur $\mathbb{K}^n$, puisque la norme de Frobenius de la matrice unité I_n est $\sqrt{n}$ et non 1.

3.4. Définition. — *Soit $A \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ une matrice $n \times n$ à coefficients dans le corps $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. On appelle rayon spectral de A, et on note $\rho(A)$, le plus grand des modules des valeurs propres (réelles ou complexes) de A.*

3.5. Proposition. — *On munit $\mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)$ d'une norme multiplicative quelconque. Pour tout $A \in \mathcal{L}(\mathbb{C}^n, \mathbb{C}^n)$, on a*

$$\rho(A) \leq \|A\| .$$

Preuve : Soit λ une valeur propre de A telle que $|\lambda| = \rho(A)$, et soit $x \in \mathbb{C}^n$, $x \neq 0$, un vecteur propre associé à la valeur propre λ. On a

$$Ax = \lambda x , \quad \text{donc} \quad A\, x\, {}^t\overline{x} = \lambda\, x\, {}^t\overline{x} .$$

La norme sur $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ étant multiplicative, on en déduit

$$\|A\, x\, {}^t\overline{x}\| = |\lambda|\, \|x\, {}^t\overline{x}\| = \rho(A) \|x\, {}^t\overline{x}\| \leq \|A\|\, \|x\, {}^t\overline{x}\| ,$$

d'où puisque $\|x\, {}^t\overline{x}\| \neq 0$,

$$\rho(A) \leq \|A\| .$$

□

On admettra le résultat suivant, que l'on démontre en en utilisant la décomposition canonique d'un endomorphisme linéaire.

3.6. Proposition. — *Soit $A \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Pour tout $\varepsilon > 0$, il existe une norme sur $\mathbb{K}^n$ telle que, lorsqu'on munit $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ de la norme associée,*

$$\|A\| \leq \rho(A) + \varepsilon .$$

3.7. Proposition. — *Soit $A \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Les deux propriétés suivantes sont équivalentes :*

(i) *pour tout $x \in \mathbb{K}^n$, la suite $(A^k x,\ k \in \mathbb{N})$ converge vers 0 ;*
(ii) *le rayon spectral $\rho(A)$ de A vérifie $\rho(A) < 1$.*

Preuve : On suppose $\rho(A) < 1$. Soit $\varepsilon > 0$ tel que $\rho(A) + \varepsilon < 1$. On munit $\mathbb{K}^n$ d'une norme telle que, pour la norme associée, $\|A\| \leq \rho(A) + \varepsilon < 1$. Alors, pour tout $x \in \mathbb{K}^n$, $\|A^k x\| \leq \|A\|^k \|x\|$, ce qui prouve que $(A^k x)$ comverge vers 0 lorsque $k \to +\infty$.
On suppose $\rho(A) \geq 1$. On doit considérer successivement deux cas :
Si $\mathbb{K} = \mathbb{C}$, ou bien si $\mathbb{K} = \mathbb{R}$ et s'il existe une valeur propre λ réelle de A de module égal à $\rho(A)$, on prend pour $x \in \mathbb{K}^n$ un vecteur propre associé à la valeur propre λ de module $\rho(A)$, et on voit que $(A^k x)$ ne converge pas vers 0 lorsque $k \to +\infty$.

Si $\mathbb{K} = \mathbb{R}$ et si toutes les valeurs propres de A de module $\rho(A)$ sont complexes, il existe deux valeurs propres de A, $\rho(A)e^{i\theta}$ et $\rho(A)e^{-i\theta}$, complexes conjuguées l'une de l'autre, et deux vecteurs non nuls x et y de $\mathbb{R}^n$, tels que $x+iy$ et $x-iy$ soient vecteurs propres du complexifié de A associés, respectivement, aux valeurs propres $\rho(A)e^{i\theta}$ et $\rho(A)e^{-i\theta}$. On a alors

$$A^k(x+iy) = \big(\rho(A)\big)^k e^{ik\theta}(x+iy)\,, \quad A^k(x-iy) = \big(\rho(A)\big)^k e^{-ik\theta}(x-iy)\,,$$

d'où

$$A^k x = \big(\rho(A)\big)^k \big(x\cos(k\theta) - y\sin(k\theta)\big)\,, \quad A^k y = \big(\rho(A)\big)^k \big(x\sin(k\theta) + y\cos(k\theta)\big)\,,$$

ce qui prouve que $(A^k x)$ et $(A^k y)$ ne convergent pas vers 0 lorsque $k \to +\infty$. □

4. Le théorème du point fixe

4.1. Définition. — *Une application f d'un espace métrique (E,d) dans un autre espace métrique (F,δ) est dite lipschitzienne s'il existe un réel $k \geq 0$ tel que, pour tous x et $y \in E$,*

$$\delta\big(f(x), f(y)\big) \leq k\, d(x,y)\,.$$

Le réel k est appelé rapport de l'application lipschitzienne f. L'application f est dite contractante si elle est lipschitzienne de rapport $k < 1$.

4.2. Définition. — *Un point fixe d'une application f d'un ensemble E dans lui-même est un point $x \in E$ tel que $f(x) = x$.*

4.3. Théorème. — *Une application contractante d'un espace métrique complet dans lui-même possède un point fixe unique.*

Preuve : Soit x_0 un point de E. On pose $x_1 = f(x_0)$, puis, pour tout $n \geq 1$, $x_n = f(x_{n-1})$. On a, pour tout $n \geq 1$,

$$d(x_{n+1}, x_n) \leq k d(x_n, x_{n-1}) \leq \cdots \leq k^n d(x_1, x_0)\,,$$

d'où on déduit, pour tous m et n vérifiant $1 \leq n \leq m$,

$$\begin{aligned} d(x_m, x_n) &\leq d(x_m, x_{m-1}) + d(x_{m-1}, x_{m-2}) + \cdots + d(x_{n+1}, x_n) \\ &\leq (k^{m-1} + k^{m-2} + \cdots + k^n) d(x_1, x_0) \\ &\leq k^n \,\frac{1-k^{m-n}}{1-k}\, d(x_1, x_0)\,. \end{aligned}$$

La dernière inégalité montre que la suite $(x_n\,,\ n \in \mathbb{N})$ est de Cauchy. Puisque E est complet, cette suite converge; soit a sa limite. En faisant tendre n vers $+\infty$ dans l'égalité $f(x_n) = x_{n+1}$, on obtient $f(a) = a$, ce qui prouve que a est un point fixe de f. Si b est un autre point fixe de f, on peut écrire

$$d(a,b) = d\big(f(a), f(b)\big) \leq k\, d(a,b)\,,$$

ce qui implique

$$(1-k)d(a,b) \leq 0\,,$$

d'ou, puisque $1-k>0$, $d(a,b)=0$, c'est-à-dire $a=b$. Le point fixe de f est donc unique. □

4.4. Exemple. — Considérons l'application f de $\mathbb{C}$ dans $\mathbb{C}$:

$$f(z)=z^2\,.$$

Soit r un réel vérifiant $0<r<1/2$, et D le disque fermé de centre l'origine et de rayon r :

$$D=\{\,z\in\mathbb{C}\;;\quad |z|\leq r\,\}\,.$$

L'application f applique le disque D dans lui-même, car pour tout $z\in D$, $|z^2|=|z|^2\leq r^2\leq r$. De plus, la restriction de f au disque D est lipschitzienne de rapport $2r<1$. En effet, pour tous y et $z\in D$,

$$\big|f(y)-f(z)\big|=|y^2-z^2|=|y+z|\,|y-z|\leq\big(|y|+|z|\big)|y-z|\leq 2r|y-z|\,.$$

Le disque D étant compact, la restriction à ce disque de l'application f possède un point fixe unique (qui est bien sûr l'origine).

4.5. Remarques

a) Remarquons que pour tout réel r vérifiant $0\leq r<1$, la restriction de l'application f au disque de centre l'origine et de rayon r applique ce disque dans lui-même, et admet dans ce disque un point fixe unique, l'origine. Cependant, si $1/\sqrt{2}<r<1$, la restriction de f à ce disque n'est pas contractante.

b) La démonstration du théorème du point fixe donne un exemple très simple de méthode itérative convergente. En effet, le point fixe de l'application f est obtenu comme limite de la suite $(x_n\,,\ n\in\mathbb{N})$, avec $x_n=f^n(x_0)$, le point initial x_0 étant un point quelconque de l'espace E.

5. Résolution itérative de systèmes linéaires : généralités

5.1. Le problème étudié. — On considère le système linéaire

$$Ax=b\,,$$

où $A\in\mathcal{L}(\mathbb{K}^m\,,\mathbb{K}^n\,)$ et $b\in\mathbb{K}^n$ sont des données et où $x\in\mathbb{K}^m$ est l'inconnue, avec $\mathbb{K}=\mathbb{R}$ ou $\mathbb{C}$. Le lecteur remarquera que si A n'est pas inversible, et en particulier, si $m\neq n$, la solution de ce système n'existe pas toujours, et lorsqu'elle existe, n'est pas nécessairement unique. On supposera dans la suite qu'il existe une solution $\omega\in\mathbb{K}^m$ de ce système; on a donc $A\omega=b$. Bien entendu, si A est inversible (ce qui suppose $m=n$) la solution ω du système est unique, elle est donnée par

$$\omega=A^{-1}b\,.$$

Les méthodes itératives de résolution de ce système consistent à construire, par divers procédés, une suite $(x_n\,,\ n\in\mathbb{N})$ d'éléments de $\mathbb{K}^n$ telle que la suite $(Ax_n\,,\ n\in\mathbb{N})$ converge vers b.

La proposition suivante montre que lorsque A est inversible, la suite $(x_n\,,\ n \in \mathbb{N})$ converge vers la solution ω du système.

5.2. Proposition. — *Soit $A \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$, et $b \in \mathbb{K}^n$, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Soit $(x_n\,,\ n \in \mathbb{N})$ une suite dans $\mathbb{K}^n$ telle que la suite $(Ax_n\,,\ n \in \mathbb{N})$ converge vers b. Si A est inversible, la suite $(x_n\,,\ n \in \mathbb{N})$ converge vers $\omega = A^{-1}b$.*

Preuve : Il suffit de remarquer que A^{-1} est continue et que pour tout $n \in \mathbb{N}$, $x_n = A^{-1}(Ax_n)$. La suite $(Ax_n\,,\ n \in \mathbb{N})$ convergeant vers b, son image $(x_n\,,\ n \in \mathbb{N})$ par l'application continue A^{-1} converge vers $A^{-1}b$. □

5.3. Remarque. — Lorsque A n'est pas inversible, la convergence de la suite $(Ax_n\,,\ n \in \mathbb{N})$ n'implique pas celle de la suite $(x_n\,,\ n \in \mathbb{N})$. Mais si cette suite converge ou, plus généralement, si une suite $(x_{\sigma(n)}\,,\ n \in \mathbb{N})$, extraite de la suite $(x_n\,,\ n \in \mathbb{N})$, converge vers une limite ω, alors cette limite vérifie $A\omega = b$. On a noté σ une application strictement croissante de $\mathbb{N}$ dans $\mathbb{N}$. Ce résultat est vrai même lorsque $A \in \mathcal{L}(\mathbb{K}^m, \mathbb{K}^n)$, avec $m \neq n$.

6. Itérations linéaires

6.1. Principe général des itérations linéaires. — On considère le système linéaire

$$Ax = b\,, \tag{1}$$

où $A \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$ et $b \in \mathbb{K}^n$ sont des données et où $x \in \mathbb{K}^n$ est l'inconnue, avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. On suppose A inversible, et on note

$$\omega = A^{-1}b$$

l'unique solution de ce système.

Les méthodes de résolution du système (1) par itérations linéaires consistent toutes à exprimer A sous la forme

$$A = M - N\,, \qquad M \text{ et } N \in \mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)\,,$$

où M est inversible et choisie de telle sorte que les systèmes linéaires de la forme

$$My = c\,,$$

avec $c \in \mathbb{K}^n$ donné, $y \in \mathbb{K}^n$ étant l'inconnue, soient faciles à résoudre. Par exemple, M pourra être une matrice diagonale, ou une matrice triangulaire (inférieure ou supérieure), à coeffficients diagonaux tous non nuls. Le système (1) équivaut alors à

$$Mx = Nx + b\,.$$

Cela suggère la construction d'une suite $(x^{(0)},\ x^{(1)},\ \ldots,\ x^{(k)},\ \ldots)$ d'éléments de $\mathbb{K}^n$, en résolvant successivement les systèmes

$$\begin{aligned} Mx^{(0)} &= b\,,\\ Mx^{(1)} &= Nx^{(0)} + b\,,\\ &\ldots\\ Mx^{(k)} &= Nx^{(k-1)} + b\,,\\ &\ldots \end{aligned} \tag{2}$$

La matrice M étant supposée inversible, on a

$$x^{(0)} = M^{-1}b\,, \tag{3}$$

et, pour tout $k \geq 1$,

$$x^{(k)} = M^{-1}(Nx^{(k-1)} + b)\,. \tag{4}$$

On peut alors affirmer :

6.2. Proposition. — *Si la suite $(x^k,\ k \in \mathbb{N})$ construite par application des formules (3) et (4) ci-dessus converge, sa limite $x^{(\infty)}$ est la solution ω du système linéaire (1).*

Preuve : On a, pour tout $k \geq 1$,

$$Mx^{(k)} = Nx^{(k-1)} + b\,.$$

En faisant tendre k vers $+\infty$, on obtient

$$Mx^{(\infty)} = Nx^{(\infty)} + b\,, \quad \text{ou} \quad Ax^{(\infty)} = b\,.$$

□

Les deux propositions suivantes donnent, la première, une condition suffisante de convergence, et la seconde, une condition nécessaire et suffisante de convergence.

6.3. Proposition. — *On suppose qu'il existe sur $\mathbb{K}^n$ une norme $x \mapsto \|x\|$, telle que, pour la norme associée sur $\mathcal{L}(\mathbb{K}^n, \mathbb{K}^n)$, on ait*

$$\|M^{-1}N\| < 1\,.$$

Alors quel que soit $b \in \mathbb{K}^n$, la suite $(x^{(k)},\ k \in \mathbb{N})$ définie par (3) et (4) converge vers la solution du système (1).

Preuve : Soit ω la solution de (1). On a

$$\omega = M^{-1}(N\omega + b)$$

et, pour tout entier $k \geq 1$,

$$x^{(k)} = M^{-1}(Nx^{(k-1)} + b)\,,$$

d'où

$$x^{(k)} - \omega = M^{-1}N(x^{(k-1)} - \omega)\,.$$

Par suite,

$$\|x^{(k)} - \omega\| \leq \|M^{-1}N\|\,\|x^{(k-1)} - \omega\|\,.$$

On en déduit immédiatement

$$\|x^{(k)} - \omega\| \leq \left(\|M^{-1}N\|\right)^k \|x^0 - \omega\|\,.$$

Si $\|M^{-1}N\| < 1$, on conclut que $\lim_{k\to+\infty} \|x^{(k)} - \omega\| = 0$. □

6.4. Proposition. — *La suite $(x^{(k)},\ k \in \mathbb{N})$ définie par (3) et (4) converge vers la solution du système (1) quel que soit l'élément b de $\mathbb{K}^n$ si et seulement si le rayon spectral de $M^{-1}N$ est strictement inférieur à 1.*

Preuve : On a vu, lors de la démonstration de la proposition précédente, que

$$x^{(k)} - \omega = M^{-1}N(x^{(k-1)} - \omega)\,.$$

Le résultat est donc une conséquence immédiate de la proposition 3.7. □

6.5. La méthode de Jacobi. — Cette méthode consiste à choisir pour matrice $M = (m_{i\,j})$ la partie diagonale de A :

$$m_{i\,j} = \begin{cases} a_{i\,j} & \text{si } i = j, \\ 0 & \text{si } i \neq j. \end{cases}$$

Cette méthode suppose donc tous les coefficients diagonaux de A non nuls. Les formules qui permettent de calculer $x^{(k)}$ une fois $x^{(k-1)}$ déterminé s'écrivent

$$\begin{aligned} a_{1\,1}x_1^{(k)} &= -\sum_{2\leq j\leq n} a_{1\,j}x_j^{(k-1)} + b_1\,, \\ &\dots \\ a_{i\,i}x_i^{(k)} &= -\sum_{1\leq j\leq n\,,\ j\neq i} a_{i\,j}x_j^{(k-1)} + b_i\,, \\ &\dots \\ a_{n\,n}x_n^{(k)} &= -\sum_{1\leq j\leq n-1} a_{n\,j}x_j^{(k-1)} + b_n\,. \end{aligned}$$

6.6. La méthode de Gauss-Seidel. — Cette méthode consiste à choisir pour matrice $M = (m_{i\,j})$ la partie triangulaire inférieure de A, diagonale comprise :

$$m_{i\,j} = \begin{cases} a_{i\,j} & \text{si } j \leq i, \\ 0 & \text{si } j > i. \end{cases}$$

Cette méthode, comme celle de Jacobi, suppose que tous les coefficients diagonaux de A sont non nuls. Les formules qui permettent de calculer $x^{(k)}$ une fois $x^{(k-1)}$ déterminé s'écrivent

$$\begin{aligned} a_{1\,1}x_1^{(k)} &= -\sum_{2\leq j\leq n} a_{1\,j}x_j^{(k-1)} + b_1\,, \\ a_{2\,2}x_2^{(k)} &= -a_{2\,1}x_1^{(k)} - \sum_{3\leq j\leq n} a_{2\,j}x_j^{(k-1)} + b_2\,, \\ &\dots \\ a_{i\,i}x_i^{(k)} &= -\sum_{1\leq j\leq i-1} a_{i\,j}x_j^{(k)} - \sum_{i+1\leq j\leq n} a_{i\,j}x_j^{(k-1)} + b_i\,, \\ &\dots \\ a_{n\,n}x_n^{(k)} &= -\sum_{1\leq j\leq n-1} a_{n\,j}x_j^{(k)} + b_n\,. \end{aligned}$$

On remarque qu'avec cette méthode, le calcul de $x^{(k)}$ n'est pas plus compliqué qu'avec la méthode de Jacobi : le nombre d'opérations à effectuer est le même dans les deux

méthodes. L'application de la méthode de Jacobi impose de garder en mémoire toutes les composantes de $x^{(k-1)}$ pendant tout le calcul de $x^{(k)}$, tandis que dans la méthode de Gauss-Seidel, on peut "oublier" $x_i^{(k-1)}$ dès qu'on a calculé $x_{i-1}^{(k)}$. La méthode de Gauss-Seidel nécessite donc moins de place disponible en mémoire que celle de Jacobi.

6.7. Les méthodes de relaxation. — Ce sont des généralisations de la méthode de Gauss-Seidel, qui consistent à prendre pour matrice $M = (m_{i\,j})$ une combinaison linéaire de la partie diagonale de A et de la partie triangulaire inférieure de A, diagonale non comprise. On pose

$$m_{i\,j} = \begin{cases} \big(1 + (\lambda^{-1} - 1)\delta_{i\,j}\big)a_{i\,j} & \text{si } j \leq i, \\ 0 & \text{si } j > i. \end{cases}$$

Dans ces formules, λ est un paramètre réel non nul, que l'on choisit afin d'assurer la convergence de la méthode, ou de l'accélérer. Pour $\lambda = 1$, on retrouve la méthode de Gauss-Seidel. Pour $\lambda < 1$, la méthode est dite "méthode de sous-relaxation", et pour $\lambda > 1$, "méthode de sur-relaxation". Comme les méthodes de Jacobi et de Gauss-Seidel, les méthodes de relaxation supposent les coefficients diagonaux de A tous non nuls. Les formules qui permettent de calculer $x^{(k)}$ une fois $x^{(k-1)}$ déterminé s'écrivent

$$\lambda^{-1}a_{1\,1}x_1^{(k)} = -(1-\lambda^{-1})a_{1\,1}x_1^{(k-1)} - \sum_{2\leq j\leq n} a_{1\,j}x_j^{(k-1)} + b_1\,,$$

$$\lambda^{-1}a_{2\,2}x_2^{(k)} = -a_{2\,1}x_1^{(k)} - (1-\lambda^{-1})a_{2\,2}x_2^{(k-1)} - \sum_{3\leq j\leq n} a_{2\,j}x_j^{(k-1)} + b_2\,,$$

$$\dots$$

$$\lambda^{-1}a_{i\,i}x_i^{(k)} = -\sum_{1\leq j\leq i-1} a_{i\,j}x_j^{(k)} - (1-\lambda^{-1})a_{i\,i}x_i^{(k-1)} - \sum_{i+1\leq j\leq n} a_{i\,j}x_j^{(k-1)} + b_i\,,$$

$$\dots$$

$$\lambda^{-1}a_{n\,n}x_n^{(k)} = -\sum_{1\leq j\leq n-1} a_{n\,j}x_j^{(k)} - (1-\lambda^{-1})a_{n\,n}x_n^{(k-1)} + b_n\,.$$

7. Valeurs propres d'une matrice symétrique : méthode de Jacobi

7.1. Rappel sur les matrices symétriques et les matrices orthogonales. — Dans tout ce paragraphe, on note $(e_1, \dots, e_n)$ la base canonique de $\mathbb{R}^n$. Les éléments de $\mathbb{R}^n$ sont considérés comme des matrices à 1 colonne et n lignes. Le vecteur de base e_i est le vecteur-colonne dont la i-ème composante est 1, les autres étant nulles. On munit $\mathbb{R}^n$ du produit scalaire usuel

$$(x, y) \mapsto (x|y) = {}^t x\, y = \sum_{i=1}^{n} x_i y_i\,,$$

où x et y sont deux éléments de $\mathbb{R}^n$, ${}^t x$ désignant le vecteur-ligne transposé de x.
Une matrice $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ est dite *symétrique* si

$${}^t A = A\,,$$

où ${}^tA$ désigne la matrice transposée de A. On sait que A est symétrique si et seulement si, pour tous x et $y \in \mathbf{R}^n$,

$$(Ax|y) = (x|Ay)\,.$$

Une matrice $R \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ est dite *orthogonale* si, pour tous x et $y \in \mathbb{R}^n$,

$$(Rx|Ry) = (x|y)\,. \qquad (*)$$

On rappelle que la transposée ${}^tM$ d'une matrice $M \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ vérifie, pour tous x et $y \in \mathbb{R}^n$,

$$({}^tMx|y) = (x|My)\,.$$

En utilisant $(*)$ et cette propriété, on voit immédiatement qu'une matrice R est orthogonale si et seulement si elle est inversible et a pour inverse

$$R^{-1} = {}^tR\,.$$

Lorsque c'est le cas, $R^{-1} = {}^tR$ est elle aussi orthogonale. On vérifie aussi que si R et S sont deux matrices orthogonales, RS est orthogonale. L'ensemble des matrices orthogonales est donc un groupe, appelé *groupe orthogonal* et noté $\mathbf{O}(n)$.
Soit $R = (r_{ij}) \in \mathbf{O}(n)$ une matrice orthogonale. Compte tenue de $(*)$, on voit que les vecteurs $Re_1, \ldots, Re_n$ sont deux à deux orthogonaux et tous de norme 1. Ils forment donc une base orthonormée de $\mathbb{R}^n$. Cette propriété s'exprime, au moyen des coefficients r_{ij} de R, par

$$\sum_{j=1}^{n} r_{ji} r_{jk} = \delta_{ik}\,. \qquad (**)$$

Comme la transposée ${}^tR = R^{-1}$ de R est orthogonale, on a aussi

$$\sum_{j=1}^{n} r_{ij} r_{kj} = \delta_{ik}\,. \qquad (***)$$

Les formules $(**)$ et $(***)$ expriment simplement que ${}^tR\,R = R\,{}^tR = I_n$, matrice unité. Une matrice R dont les coefficients vérifient une de ces formules est donc orthogonale.
Ces formules montrent aussi que $\mathbf{O}(n)$ est une partie fermée et bornée, donc compacte, de $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$.

7.2. Lemme. — *Soit A une matrice $n \times n$ réelle symétrique, $R \in \mathbf{O}(n)$ une matrice orthogonale. La matrice*

$$B = R^{-1}AR = {}^tRAR$$

est symétrique, et ses coefficients b_{ij} vérifient

$$\sum_{(i,j)} (b_{ij})^2 = \sum_{(i,j)} (a_{ij})^2\,.$$

Preuve : On a

$${}^tB = {}^tR\,{}^tA\,R = {}^tR\,A\,R = B\,,$$

donc B est symétrique. D'autre part

$$ {}^tB\,B = {}^tR({}^tA\,A)\,R\,, $$

donc

$$ \mathrm{Trace}({}^tB\,B) = \mathrm{Trace}({}^tA\,A)\,, $$

c'est-à-dire

$$ \sum_{(i,j)} (b_{i\,j})^2 = \sum_{(i,j)} (a_{i\,j})^2\,. $$

□

7.3. Rotations planes. — Soient e_p et e_q, $p \neq q$, deux vecteurs distincts de la base canonique de $\mathbb{R}^n$, et $\theta \in \mathbf{S}^1 = \mathbb{R}/2\pi\mathbb{Z}$ un angle, c'est-à-dire un réel modulo 2π. La matrice $R \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ définie par

$$ \begin{aligned} Re_p &= \cos\theta e_p + \sin\theta e_q\,, \\ Re_q &= -\sin\theta e_p + \cos\theta e_q\,, \\ Re_k &= e_k \quad \text{pour } 1 \leq k \leq n\,,\ k \neq p\ \ k \neq q\,, \end{aligned} $$

représente une rotation d'angle θ dans le plan $E_{p,q}$ engendré par les vecteurs e_p et e_q, qui laisse fixe chaque point du sous-espace, orthogonal à ce plan, engendré par les vecteurs e_k, $k \neq p$, $k \neq q$. Cette rotation est un élément de $\mathbf{O}(n)$ qui a pour coefficients

$$ \begin{aligned} &r_{p\,p} = \cos\theta\,, \quad r_{p\,q} = -\sin\theta\,, \quad rp\,k = 0 \quad \text{pour } k \neq p\,,\ k \neq q\,, \\ &r_{q\,p} = \sin\theta\,, \quad r_{q\,q} = \cos\theta\,, \quad rq\,k = 0 \quad \text{pour } k \neq p\,,\ k \neq q\,, \\ &r_{k\,l} = \delta_{k\,l} \quad \text{pour } k \neq p\,,\ k \neq q\,,\ l \neq p\,,\ l \neq q\,. \end{aligned} $$

7.4. Principe de la méthode de Jacobi. — Soit A une matrice $n \times n$ réelle symétrique. On sait que pour toute matrice inversible R, $R^{-1}\,A\,R$ a le même polynôme caractéristique que A, donc les mêmes valeurs propres avec les mêmes multiplicités. C'est, en particulier, le cas lorsque R est orthogonale; on a alors $R^{-1} = {}^tR$.

La méthode de Jacobi, pour la détermination des valeurs propres de A, consiste à construire une suite de matrices orthogonales $R^{(0)}$, $R^{(1)}$, …, $R^{(k)}$, …, et à poser

$$ A^{(0)} = A\,, \quad \text{et pour tout } k \geq 0\,, \quad A^{(k+1)} = {}^tR^{(k)}\,A^{(k)}\,R^{(k)}\,. $$

Les matrices $A^{(k)}$ ont toutes les mêmes valeurs propres, avec les mêmes multiplicités. On verra qu'on peut choisir les matrices orthogonales $R^{(k)}$ de manière telle que tous les coefficients non diagonaux $a_{i\,i}^{(k)}$ de $A^{(k)}$, $i \neq j$, tendent vers 0 lorsque $k \to +\infty$, et que ses coefficients diagonaux $a_{i\,j}^{(k)}$ tendent vers des limites λ_i. En d'autres termes, la suite de matrices $(A^{(k)})$ converge, lorsque $k \to +\infty$, vers une matrice diagonale Λ. On montrera que les coefficients diagonaux λ_i de λ sont les valeurs propres de A. On aura ainsi retrouvé un résultat bien connu d'algèbre linéaire : les valeurs propres d'une matrice $n \times n$ réelle symétrique sont toutes réelles.

En pratique, on choisit pour les matrices orthogonales $R^{(k)}$ des rotations planes, du type étudié au paragraphe 7.3.

On remarque d'autre part que, pour tout $k \geq 0$,

$$A^{(k+1)} = {}^t S^{(k)}\, A\, S^{(k)}\,, \quad \text{avec} \quad S^{(k)} = R^{(0)}\, R^{(1)} \cdots R^{(k)}\,.$$

Les $S^{(k)}$ sont des matrices orthogonales. On montrera que si la matrice A a n valeurs propres deux à deux distinctes, la suite de matrices $(S^{(k)})$ converge vers une matrice orthogonale S, qui vérifie

$$\Lambda = {}^t S\, A\, S\,. \qquad (*)$$

Lorsque la matrice A admet des valeurs propres multiples, on ne peut pas affirmer que la suite de matrices $(S^{(k)})$ converge. Cependant, la compacité de $\mathbf{O}(n)$ nous permettra d'affirmer qu'il existe une suite, extraite de la suite $(S^{(k)})$, qui converge vers une matrice orthogonale S qui vérifie encore l'égalité $(*)$. On aura ainsi retrouvé un résultat bien connu d'algèbre linéaire : il existe un changement de base orthonormée de $\mathbb{R}^n$ qui transforme la matrice symétrique A en la matrice diagonale Λ.

7.5. Choix de la matrice $R^{(k)}$. — Soit $k \geq 0$ un entier fixé. On suppose les $A^{(l)}$ déterminés pour tous les l vérifiant $0 \leq l \leq k$. On va déterminer $R^{(k)}$, puis $A^{(k+1)} = {}^t R^{(k)}\, A^{(k)}\, R^{(k)}$. Afin d'alléger l'écriture, on écrira dans le présent paragraphe

- A au lieu de $A^{(k)}$,
- R au lieu de $R^{(k)}$,
- $B = {}^t R\, A\, R$ au lieu de $A^{(k+1)}$.

Si tous les coefficients non diagonaux de A sont nuls, le résultat recherché est atteint. Sinon, on choisit un coefficient non diagonal $a_{p\,q}$ de A, le plus grand possible en module :

$$|a_{p\,q}| = \max_{(i,j)\,,\; i \neq j} |a_{i\,j}|\,.$$

Il peut exister plusieurs termes non diagonaux de A vérifiant cette propriété; on devra fixer une règle permettant d'en choisir un, par exemple celui rencontré en premier lorsqu'on parcourt la partie triangulaire inférieure de A ligne par ligne, de haut en bas et de gauche à droite.

Soit $\theta \in \mathbf{S}^1$ un angle, et R la rotation plane d'angle θ dans le plan engendré par e_p et e_q. On pose

$$B = {}^t R\, A\, R\,.$$

On peut énoncer :

7.6. Lemme. — *On choisit l'angle θ de manière telle que*

$$-\frac{\pi}{4} < \theta < \frac{\pi}{4} \quad \text{et} \quad \tan(2\theta) = \frac{2a_{p\,q}}{a_{p\,p} - a_{q\,q}} \quad \text{si } a_{p\,p} - a_{q\,q} \neq 0\,,$$

$$\theta = \frac{\pi}{4} \quad \text{si } a_{p\,p} - a_{q\,q} = 0\,.$$

Les coefficients $b_{i\,j}$ de B vérifient alors

$$b_{p\,q} = 0\,,$$

$$b_{i\,i} = a_{i\,i} \quad \text{si } i \neq p\,,\ i \neq q\,,$$

$$b_{p\,p} = a_{p\,p} + \tan\theta a_{p\,q}\,,$$

$$b_{q\,q} = a_{q\,q} - \tan\theta a_{p\,q}\,.$$

De plus,

$$\sum_{(i,j)\,,i\neq j} (b_{i\,j})^2 - \sum_{(i,j)\,,i\neq j} (a_{i\,j})^2 = -2(a_{p\,q})^2\,.$$

Preuve : En utilisant les expressions des $r_{i\,j}$ indiquées au paragraphe 7.3, on obtient après quelques calculs

$$b_{p\,q} = \cos(2\theta)a_{p\,q} - \frac{\sin(2\theta)}{2}(a_{p\,p} - a_{q\,q})\,.$$

En choisissant θ comme indiqué dans l'énoncé, on voit que

$$b_{p\,q} = 0\,.$$

De la même manière, on obtient immédiatement

$$b_{i\,i} = a_{i\,i} \quad \text{si } i \neq p\,,\ i \neq q\,,$$

ainsi que

$$b_{p\,p} = \cos^2\theta\, a_{p\,p} + \sin^2\theta\, a_{q\,q} + 2\cos\theta\sin\theta\, a_{p\,q}\,,$$

$$b_{q\,q} = \cos^2\theta\, a_{q\,q} + \sin^2\theta\, a_{p\,p} - 2\cos\theta\sin\theta\, a_{p\,q}\,.$$

Mais compte tenu du choix de θ, on en déduit

$$b_{p\,p} = a_{p\,p} + \tan\theta\, a_{p\,q}\,,$$

$$b_{q\,q} = a_{q\,q} - \tan\theta\, a_{p\,q}\,.$$

D'après le lemme 11.2, on a

$$\sum_{(i,j)} (b_{i\,j})^2 = \sum_{(i,j)} (a_{i\,j})^2\,.$$

On en déduit, en séparant les termes diagonaux des termes non diagonaux,

$$\sum_{(i,j)\,,\ i\neq j} (b_{i\,j})^2 - \sum_{(i,j)\,,\ i\neq j} (a_{i\,j})^2 = \sum_{i=1}^{n} (a_{i\,i})^2 - \sum_{i=1}^{n} (b_{i\,i})^2\,.$$

Mais d'après les valeurs obtenues pour les $b_{i\,i}$, $b_{p\,p}$ et $b_{q\,q}$, on trouve

$$\sum_{(i,j)\,,\ i\neq j} (b_{i\,j})^2 - \sum_{(i,j)\,,\ i\neq j} (a_{i\,j})^2 = -2a_{p\,q}\big(a_{p\,q}\tan^2\theta + (a_{p\,p} - a_{q\,q})\tan\theta\big)\,.$$

Compte tenu du choix de θ, on obtient finalement

$$\sum_{(i,j)\,,\ i\neq j} (b_{i\,j})^2 - \sum_{(i,j)\,,\ i\neq j} (a_{i\,j})^2 = -2(a_{p\,q})^2\,.$$

□

On peut maintenant énoncer :

7.7. Théorème. — *La suite de matrices symétriques $(A^{(k)})$, construite au moyen de l'algorithme de Jacobi, converge lorsque $k \to +\infty$ vers une matrice diagonale Λ, dont les termes diagonaux sont les valeurs propres de la matrice A.*

Preuve : La preuve donnée ici est due à Daniel Pecker. On suppose qu'à toutes les étapes de l'algorithme, la matrice $A^{(k)}$ n'est pas diagonale, car dans le cas contraire le résultat cherché est obtenu après un nombre fini d'étapes; cela suppose bien sûr $n > 2$, puisque pour $n = 1$, $A(0) = A$ est trivialement diagonale, et que pour $n = 2$, $A^{(1)}$ est diagonale. Pour alléger l'écriture on pose, pour toute matrice M,

$$\Delta(M) = \sum_{(i,j)\,,\, i\neq j} (m_{i\,j})^2 \,.$$

On remarque que $\Delta(M)$ est le carré de la distance euclidienne qui sépare M de sa diagonale. Le lemme 7.6 montre que

$$\Delta(A^{(k+1)}) - \Delta(A^{(k)}) = -2(a^{(k)}_{p\,q})^2 \,.$$

Mais puisque $a^{(k)}_{p\,q}$ est un coefficient non diagonal de la matrice $A^{(k)}$ de module maximal, et que le nombre de coefficients non diagonaux de cette matrice est $(n-1)n$, on a

$$\Delta(A^{(k)}) \leq (n-1)n(a^{(k)}_{p\,q})^2 \,.$$

On en déduit

$$\Delta(A^{(k+1)} \leq \rho^2 \Delta(A^{(k)}) \,, \qquad \text{avec} \qquad \rho^2 = 1 - \frac{2}{(n-1)n} < 1 \,. \qquad (*)$$

Ceci prouve que la suite $\big(\Delta(A^{(k)})\big)$ converge vers 0, lorsque $k \to +\infty$, au moins aussi vite que la suite géométrique de raison $\rho^2 < 1$. D'autre part, compte tenu du fait que l'angle θ des rotations planes $R^{(k)}$ vérifie toujours $|\tan\theta| \leq 1$, le lemme 7.6 montre que, pour tout i, $1 \leq i \leq n$,

$$|a^{(k+1)}_{i\,i} - a^{(k)}_{i\,i}| \leq |a^{(k)}_{p\,q}| \leq \big(\Delta(A^{(k)})\big)^{1/2} \leq C\rho^k \,,$$

où C est une constante. La série de terme général $a^{(k+1)}_{i\,i} - a^{(k)}_{i\,i}$, dont chaque terme est majoré en module par le terme correspondant d'une série géométrique de raison $\rho < 1$, est absolument convergente, donc convergente. La somme de ses k premiers termes n'est autre que $a^{(k+1)}_{i\,i}$, qui admet donc une limite lorsque $k \to +\infty$.

On a prouvé que pour tous i et j, $1 \leq i\,,\ j \leq n$, la suite $(a^{(k+1)}_{i\,i})$ a une limite, nulle si $i \neq j$, c'est-à-dire que la suite de matrices $(A^{(k)})$ converge vers une matrice diagonale Λ. Ces matrices ayant toutes le même polynôme caractéristique, et les coefficients du polynôme caractéristique d'une matrice étant des fonctions continues des coefficients de cette matrice, le polynôme caractéristique de Λ est le même que celui de A, ce qui termine la démonstration. □

7.8. Théorème. — *Si la matrice A a n valeurs propres deux à deux distinctes, la suite de matrices orthogonales $(S^{(k)} = \prod_{0\leq l\leq k} R^{(l)})$ converge, lorsque $k \to +\infty$, vers une matrice orthogonale S, qui vérifie*

$$\Lambda = {}^t S\, A\, S \,.$$

Preuve : La preuve donnée ici, comme celle du théorème précédent, est due à Daniel Pecker. Les seuls coefficients éventuellement non nuls de $R^{(k)} - I_n$, où I_n désigne la matrice unité, ont pour valeurs $\cos(\theta_k) - 1$ et $\pm \sin\theta_k$, en notant θ_k l'angle de la rotation plane $R^{(k)}$. Ces coefficients sont donc tous majorés par $|\theta_k|$, lui-même majoré par $|\tan(2\theta_k)|/2$. Lorsque $k \to +\infty$, les coefficients diagonaux de $A^{(k)}$ convergent vers les valeurs propres λ_i de A qui, par hypothèse, sont deux à deux distinctes. On est donc assuré que pour k assez grand, on a toujours $a_{p\,p}^{(k)} - a_{q\,q}^{(k)} \neq 0$. D'après le lemme 7.6,

$$\left| \frac{\tan(2\theta_k)}{2} \right| = \frac{|a_{p\,q}^{(k)}|}{|a_{p\,p}^{(k)} - a_{q\,q}^{(k)}|} \,.$$

Posons

$$K = 2 \sup_{(i,j)\,,\ i \neq j} \left(\frac{1}{|\lambda_i - \lambda_j|} \right) \,.$$

Pour k assez grand, on a

$$\frac{1}{|a_{p\,p}^{(k)} - a_{q\,q}^{(k)}|} \leq K \,,$$

donc, compte tenu de l'inégalité $(*)$ dans la démonstration de 7.7,

$$\|R^{(k)} - I_n\|_\infty \leq K |a_{p\,q}^{(k)}| \leq K \big(\Delta(A^{(k)})\big)^{1/2} \leq KC\rho^k \,,$$

où C est une constante. On a posé, pour toute matrice $M \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$,

$$\|M\|_\infty = \sup_{(i,j)} |m_{i\,j}| \,.$$

On sait que c'est une norme sur l'espace $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$. On a

$$\|S^{(k+1)} - S^{(k)}\|_\infty = \|(R^{(k)} - I_n)\, S^{(k)}\|_\infty \,.$$

Mais, d'après l'inégalité $(**)$ du paragraphe 11.1, la matrice $S^{(k)}$, comme toute matrice orthogonale, vérifie

$$\|S^{(k)}\|_\infty \leq 1 \,.$$

D'autre part, pour tout couple $(M\ ,\ N)$ de matrices $n \times n$, on a

$$(M\,N)_{i\,j} = \sum_{k=1}^{n} M_{i\,k} N_{k\,j} \,, \quad \text{donc} \quad \|M\,N\|_\infty \leq n \|M\|_\infty \|N\|_\infty \,.$$

On en déduit

$$\|S^{(k+1)} - S^{(k)}\|_\infty \leq n \|(R^{(k)} - I_n)\|_\infty \|S^{(k)}\|_\infty \leq n \|(R^{(k)} - I_n)\|_\infty \leq nKC\rho^k \,.$$

La série de terme général $S^{(k+1)} - S^{(k)}$, majorée terme à terme en module par une série géométrique de raison $\rho < 1$, est normalement convergente (pour la norme $\|\ \|_\infty$), donc convergente dans l'espace $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$. La somme de ses k premiers termes étant $S^{(k+1)}$, on voit que la suite de matrices $(S^{(k)})$ converge, lorsque $k \to +\infty$, vers une matrice S. En faisant tendre k vers $+\infty$ dans les égalités

$${}^t S^{(k)}\, S^{(k)} = I_n \,, \qquad A^{(k+1)} = {}^t S^{(k)}\, A\, S^{(k)} \,,$$

on voit que S est orthogonale et vérifie $\Lambda = {}^t S\, A\, S$. □

7.9. Proposition. — *Dans tous les cas, même lorsque la matrice symétrique A admet des valeurs propres multiples, il existe une suite extraite de la suite $(S^{(k)})$ qui converge, lorsque $k \to +\infty$, vers une matrice orthogonale S qui vérifie $\Lambda = {}^tS\,A\,S$.*

Preuve : La suite $(S^{(k)})$ étant contenue dans le compact $\mathbf{O}(n)$, il existe une suite $(S^{(\sigma(k))})$, extraite de cette suite, qui converge vers une matrice $S \in \mathbf{O}(n)$. En faisant tendre k vers $+\infty$ dans l'égalité $A^{(\sigma(k)+1)} = {}^tS^{(\sigma(k))}\,A\,S^{\left(\sigma(k)\right)}$, on obtient $\Lambda = {}^tS\,A\,S$. □

8. La méthode de plus profonde descente

8.1. Le problème étudié. — On munit $\mathbb{R}^n$ du produit scalaire euclidien usuel : si $x = (x^1, \dots, x^n)$ et $y = (y^1, \dots, y^n)$ sont deux éléments de $\mathbb{R}^n$, leur produit scalaire est

$$(x|y) = \sum_{i=1}^{n} x^i y^i \,.$$

On rappelle que lorsqu'on considère x et y comme vecteurs-colonnes, c'est-à-dire comme des matrices à une colonne et n lignes, onpeut écrire

$$(x|y) = {}^txy = {}^tyx \,,$$

où on a noté ${}^tx$ et ${}^ty$ les vecteurs-lignes transposés de x et de y, respectivement.
Soit $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$. On suppose A symétrique, c'est-à-dire vérifiant ${}^tA = A$ (en notant ${}^tA$ la transposée de A) et définie positive, c'est-à-dire vérifiant

$$(Ax|x) = {}^txAx > 0 \quad \text{pour tout} \quad x \in \mathbb{R}^n \,,\ x \neq 0 \,.$$

Soit $b \in \mathbb{R}^n$. On étudie le système linéaire

$$Ax = b \,,$$

dans lequel A et b sont des données et $x \in \mathbb{R}^n$ l'inconnue.
On rappelle que A, étant symétrique définie positive, est inversible; le système linéaire étudié a donc une solution unique $\omega = A^{-1}b$.
La méthode de plus profonde descente, étudiée dans ce paragraphe, et la méthode du gradient conjugué étudiée dans le suivant, sont basées sur une propriété remarquable de la solution ω de ce système : c'est l'unique point de $\mathbb{R}^n$ en lequel une certaine fonction F, que nous allons définir, atteint son minimum. Cette propriété est due au caractère défini positif de la matrice A. Elle est formalisée dans la proposition suivante.

8.2. Proposition. — *Soit $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ une matrice symétrique définie positive et $b \in \mathbb{R}^n$. Soit $F : \mathbb{R}^n \to \mathbb{R}$ la fonction*

$$F(x) = \frac{1}{2}\,(Ax - 2b|x) = \frac{1}{2}{}^tx\|(Ax - 2b) \,.$$

La fonction F atteint son minimum en un point unique ω de $\mathbb{R}^n$, qui est l'unique solution du système linéaire $Ax = b$.

Preuve : Soit $\omega = A^{-1}b$ la solution du système $Ax = b$. Pour tout $x \in \mathbb{R}^n$, on a

$$\begin{aligned}\big(A(x-\omega) \;\big|\; x-\omega\big) &= (Ax|x) - (A\omega|x) - (Ax|\omega) + (A\omega|\omega)\\ &= (Ax|x) - 2(A\omega|x) + (A\omega|\omega)\\ &= (Ax|x) - 2(b|x) + (b|\omega)\\ &= 2F(x) + (b|\omega)\,.\end{aligned}$$

En remarquant que $(b|\omega) = -2F(\omega)$, on peut écrire

$$F(x) = F(\omega) + \frac{1}{2}\big(A(x-\omega) \;\big|\; x-\omega\big)\,.$$

Comme on a supposé A symétrique définie positive, $\big(A(x-\omega) \;\big|\; x-\omega\big)$ est strictement positif pour tout $x \in \mathbb{R}^n$, $x \neq \omega$. On voit donc que F atteint son minimum, dont la valeur est $F(\omega) = -(1/2)(b|\omega)$, au seul point $\omega = A^{-1}b$ de $\mathbb{R}^n$. □

8.3. Rappel sur les notions de différentielle et de gradient. — On indique ici quelques notions de calcul différentiel qui sont approfondies dans une autre unité de valeur de la licence de mathématiques.

Soit $G : \mathbb{R}^n \to \mathbb{R}^p$ une application différentiable de classe C^1, c'est-à-dire une application dont toutes les composantes G^i, $1 \leq i \leq p$, sont des fonctions des composantes x^j, $1 \leq j \leq n$, de x, admettant en tout point de $\mathbb{R}^n$, des dérivées partielles $\dfrac{\partial G^i}{\partial x^j}$, qui sont des fonctions continues de x.

On appelle *différentielle* de G en un point $a \in \mathbb{R}^n$, et on note $DG(a)$, l'application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^p$ ayant pour valeur, pour tout $u = (u^1, \ldots, u^n) \in \mathbb{R}^n$,

$$\big(DG(a)(u)\big)^i = \sum_{j=1}^{n} \left(\frac{\partial G^i(x)}{\partial x^j}\;\Big|_{x=a}\right) u^j\,, \qquad 1 \leq i \leq p\,.$$

En d'autres termes, $DG(a)$ est l'application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}^p$ ayant pour matrice $\dfrac{\partial G^i(x)}{\partial x^j}\;\Big|_{x=a}$.

Considérons maintenant le cas où $p = 1$; l'application G est alors une fonction différentiable de classe C^1, définie sur $\mathbb{R}^n$ et à valeurs réelles. Sa différentielle $DG(a)$ en un point $a \in \mathbb{R}^n$ est une application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}$, c'est-à-dire un élément du dual de $\mathbb{R}^n$. Avec les notations matricielles, on l'identifie au vecteur-ligne

$$DG(a) = \left(\frac{\partial G(x)}{\partial x^1}\;\Big|_{x=a}, \ldots, \frac{\partial G(x)}{\partial x^n}\;\Big|_{x=a}\right)\,.$$

On appelle *gradient* de G au point a, et on note $\overrightarrow{\operatorname{grad}}\, G(a)$, l'élément de $\mathbb{R}^n$ transposé de $DG(a)$, c'est-à-dire le vecteur-colonne ayant pour j-ème composante

$$\big(\overrightarrow{\operatorname{grad}}\, G(a)\big)^j = \frac{\partial G(x)}{\partial x^j}\;\Big|_{x=a}\,, \qquad 1 \leq j \leq n\,.$$

En utilisant le produit scalaire euclidien usuel sur $\mathbb{R}^n$, on peut écrire, pour tout $u \in \mathbb{R}^n$,

$$DG(a)(u) = \big(\overrightarrow{\operatorname{grad}}\, G(a) \;\big|\; u\big) = \sum_{j=1}^{n} \left(\frac{\partial G(x)}{\partial x^j}\;\Big|_{x=a}\right) u^j\,.$$

En pratique, ainsi qu'on l'a vu au paragraphe 7 du chapitre II, on identifie souvent $\mathbb{R}^n$ et son dual au moyen du produit scalaire euclidien usuel, chaque élément y de $\mathbb{R}^n$ étant identifié à l'application linéaire de $\mathbb{R}^n$ dans $\mathbb{R}$: $x \mapsto (y|x)$. Cette convention conduit à identifier la différentielle de G au point a, $DG(a)$, avec la valeur en a du gradient de G, $\overrightarrow{\operatorname{grad}}\, G(a)$. Il faut n'utiliser cette identification qu'avec prudence afin de toujours bien distinguer, au moins du point de vue conceptuel, les éléments de $\mathbb{R}^n$ et ceux de son dual.
Revenons à la fonction

$$F(x) = \frac{1}{2}\,(Ax - 2b|x)\,.$$

8.4. Lemme. — *Le gradient de la fonction F en un point $x \in \mathbb{R}^n$ est*

$$\overrightarrow{\operatorname{grad}}\, F(x) = Ax - b\,.$$

Preuve : En explicitant les composantes x^i et b^i des vecteurs x et b et les composantes $a_{i\,j}$ de la matrice A, on peut écrire

$$F(x) = \frac{1}{2}\sum_{(i,j)}(a_{i\,j}x^j - 2b^i)x^i\,,$$

donc

$$\frac{\partial F(x)}{\partial x^k} = \frac{1}{2}\sum_{j=1}^{n} a_{k\,j}x^j - b^k + \sum_{i=1}^{n} a_{i\,k}x^i\,.$$

Compte tenu de la symétrie de A ($a_{i\,k} = a_{k\,i}$), on peut écrire

$$\frac{\partial F(x)}{\partial x^k} = \sum_{j=1}^{n} a_{k\,j}x^k - b^k\,.$$

La différentielle de F au point x est donc le vecteur-ligne de k-ième composante $\sum_{j=1}^{n} a_{k\,j}x^j - b^k$, et le gradient de F est le vecteur-colonne transposé du précédent, c'est-à-dire le vecteur $Ax - b$. □

8.5. Principe de la méthode de plus profonde descente. — On va résoudre le système linéaire

$$Ax = b\,,$$

où $b \in \mathbb{R}^n$, $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ matrice symétrique définie positive, en construisant une suite $(x_n\,,\ n \in \mathbb{N})$ d'éléments de $\mathbb{R}^n$ telle que la suite $(Ax_n\,,\ n \in \mathbb{N})$ converge vers b. La proposition 5.2 nous permettra alors d'affirmer que la suite $(x_n\,,\ n \in \mathbb{N})$ converge vers la solution $\omega = A^{-1}b$ du système.
Supposons $x_0,\ x_1, \ldots,\ x_k$ déjà déterminés. Posons

$$r_k = \overrightarrow{\operatorname{grad}}\, F(x_k) = Ax_k - b\,.$$

Si $r_k = 0$, $Ax_k = b$ donc $x_k = \omega$ est la solution du système. Dans le cas contraire, cherchons x_{k+1} en lui imposant d'être de la forme

$$x_{k+1} = x_k + s_k r_k\,,$$

avec $s_k \in \mathbb{R}$. Nous allons déterminer s_k de manière telle que $F(x_{k+1})$ soit le plus petit possible. Calculons donc $F(x_{k+1})$:

$$\begin{aligned} F(x_{k+1}) &= \frac{1}{2}\left(A(x_k + s_k r_k) \mid x_k + s_k r_k\right) - (b|x_k + s_k r_k) \\ &= \frac{1}{2}(Ax_k - 2b|x_k) + (Ax_k - b|r_k)s_k + \frac{1}{2}(Ar_k|r_k)s_k^2 \\ &= F(x_k) + (r_k|r_k)s_k + \frac{1}{2}(Ar_k|r_k)s_k^2 \,. \end{aligned}$$

Dérivons par rapport à s_k :

$$\frac{d}{ds_k} F\big(x_{k+1}(s_k)\big) = (Ar_k|r_k)s_k + (r_k|r_k) \,.$$

Comme $(Ar_k|r_k)$ et $(r_k|r_k)$ sont strictement positifs, nous voyons que $F(x_{k+1})$ est le plus petit possible lorsqu'on donne à s_k la valeur

$$s_k = -\frac{(r_k|r_k)}{(Ar_k|r_k)} \,.$$

8.6. L'algorithme de la plus profonde descente. — Il consiste à choisir pour point de départ un point $x_0 \in \mathbb{R}^n$ quelconque, puis à déterminer successivement x_1, x_2, ..., en posant, pour chaque $k \in \mathbb{N}$,

$$x_{k+1} = x_k + s_k r_k \,, \quad \text{avec} \quad r_k = ax_k - b \,, \quad s_k = -\frac{(r_k|r_k)}{(Ar_k|r_k)} \,.$$

L'algorithme s'arrête lorsque, après avoir déterminé x_k pour une certaine valeur de $k \in \mathbb{N}$, on trouve en calculant $r_k = Ax_k - b$, un résultat nul. On sait alors que x_k est la solution cherchée du système. Bien sûr, cela ne se produit qu'exceptionnellement : en général, la suite $(x_n \,,\ n \in \mathbb{N})$ construite par application de l'algorithme de la plus profonde descente est une suite infinie.

8.7. Théorème. — *La matrice A étant supposée symétrique définie positive, quel que soit le point initial x_0 de $\mathbb{R}^n$ choisi, la suite $(x_n \,,\ n \in \mathbb{N})$ construite au moyen de l'algorithme de la plus profonde descente converge vers la solution $\omega = A^{-1}b$ du système linéaire $Ax = b$.*

Preuve : Soit $\omega = A^{-1}b$. Posons, pour tout $x \in \mathbb{R}^n$,

$$\varphi(x) = \begin{cases} \omega & \text{si } x = \omega, \\ x - \dfrac{(Ax - b|Ax - b)}{\big(A(Ax - b) \mid Ax - b\big)}(Ax - b) & \text{si } x \neq \omega. \end{cases}$$

Il est facile de vérifier que la fonction φ ainsi définie est continue sur $\mathbb{R}^n$ (y compris au point ω). Il est facile aussi de vérifier que la suite $(x_n \,,\ n \in \mathbb{N})$ est obtenue par itération de l'application φ : on a en effet $x_1 = \varphi(x_0)$, $x_2 = \varphi(x_1)$, ..., $x_{k+1} = \varphi(x_k)$.
La suite $\big(F(x_n) \,,\ n \in \mathbb{N}\big)$ est décroissante et minorée par $F(\omega)$. Elle converge donc vers une limite $l \in \mathbb{R}$. La suite $(x_n \,,\ n \in \mathbb{N})$ est contenue dans $\big\{\, x \in \mathbb{R}^n \ ;\ \ F(x) \leq F(x_0) \,\big\}$,

qui est une partie compacte de $\mathbb{R}^n$. On peut donc en extraire une suite convergente $(y_n = x_{\sigma(n)}\,,\ n \in \mathbb{N})$, où σ désigne une application strictement croissante de $\mathbb{N}$ dans $\mathbb{N}$. Soit $y \in \mathbb{R}^n$ la limite de cette suite. Puisque F est continue, nous avons $F(y) = l$. Nous allons prouver par l'absurde que $y = \omega$. Supposons donc $y \neq \omega$. Nous avons alors $l > F(\omega)$ et $F\big(\varphi(y)\big) < l$. La fonction F étant continue, l'ensemble U des $z \in \mathbb{R}^n$ tels que $F(z) < (1/2)\Big(F\big(\varphi(y)\big) + l\Big)$ est un voisinage de $\varphi(y)$. L'application φ étant continue, l'ensemble $V = \varphi^{-1}(U)$ des $x \in \mathbb{R}^n$ tels que $\varphi(x) \in U$ est un voisinage de y. Puisque la suite $(y_n = x_{\sigma(n)}\,,\ n \in \mathbb{N})$ converge vers y, on peut affirmer que pour k assez grand, y_k est élément de V. Mais alors $\varphi(y_k) = x_{\sigma(k)+1}$ est élément de U, donc, d'après la définition de U, $F(x_{\sigma(k)+1}) < (1/2)\Big(F\big(\varphi(y)\big) + l\Big) < l$. Mais ce résultat est en contradiction avec le fait que la suite $\big(F(x_n)\,,\ n \in \mathbb{N}\big)$ est décroissante et a pour limite l. Nous avons donc bien prouvé que la limite y de la suite $(y_n = x_{\sigma(n)}\,,\ n \in \mathbb{N})$ est égale à ω. Nous avons aussi prouvé que toute suite convergente extraite de la suite $(x_n\,,\ n \in \mathbb{N})$ a nécessairement pour limite ω. La suite $(x_n\,,\ n \in \mathbb{N})$ étant contenue dans un compact, un résultat classique de Topologie permet d'affirmer que cette suite converge vers ω. □

Nous allons, dans la suite de ce paragraphe, donner une interprétation géométrique de la méthode de la plus profonde descente. Cette interprétation nous permettra de mieux faire comprendre, dans le paragraphe suivant, la méthode du gradient conjugué.

8.8. Les variétés de niveau de la fonction F. — Rappelons que la fonction $F : \mathbb{R}^n \to \mathbb{R}$ est définie par

$$F(x) = \frac{1}{2}\,(Ax - 2b|x)\,.$$

Rappelons aussi (proposition 8.2) que F atteint son minimum en un point unique $\omega = A^{-1}(b)$, et que $2F(\omega) = -(b|\omega)$. Posons, pour tout $\lambda \in \mathbb{R}$,

$$Q_\lambda = \big\{\, x \in \mathbb{R}^n \ ;\ \ F(x) = \lambda \,\big\}\,.$$

Les Q_λ sont appelées *variétés de niveau* de la fonction F. Bien évidemment, pour $\lambda < F(\omega)$, Q_λ est vide, et $Q_{F(\omega)} = \{\,\omega\,\}$. Pour $\lambda > F(\omega)$, Q_λ est une quadrique (surface définie par une équation polynomiale de degré 2) compacte de $\mathbb{R}^n$. Pour $n = 2$, c'est une ellipse, et pour $n = 3$ c'est un ellipsoïde.

8.9. Lemme. — *Le point $\omega = A^{-1}b$ est l'unique point de $\mathbb{R}^n$ centre de symétrie de toutes les quadriques Q_λ, $\lambda \geq F(\omega)$. De plus, les quadriques Q_λ sont toutes homothétiques ; plus précisément, l'homothétie de centre ω et de rapport $\mu > 0$ est, pour chaque réel $\lambda \geq F(\omega)$, une bijection de Q_λ sur $Q_{\lambda'}$, avec $\lambda' = F(\omega) + \mu^2\big(\lambda - F(\omega)\big)$.*

Preuve : Soient c et z deux points de $\mathbb{R}^n$. Un calcul facile donne

$$F(c+z) - F(c-z) = 2(Ac - b|z)\,.$$

Le point c est centre de symétrie d'une quadrique Q_λ si et seulement si pour tout $z \in \mathbb{R}^n$ tel que $F(c+z) = \lambda$, on a aussi $F(c-z) = \lambda$. L'expression ci-dessus de $F(c+z) - F(c-z)$ montre immédiatement que c est centre de symétrie de Q_λ si et seulement si $Ac = b$, c'est-à-dire si et seulement si $c = \omega$.
On a, pour tout $z \in \mathbb{R}^n$,

$$F(\omega + z) = F(\omega) + \frac{1}{2}\,(Az|z)\,, \quad F(\omega + \mu z) = F(\omega) + \frac{\mu^2}{2}\,(Az|z)\,,$$

d'où

$$F(\omega + \mu z) = \mu^2 F(\omega + z) + (1 - \mu^2)F(\omega)\,.$$

Par suite, $\omega + z$ appartient à Q_λ, c'est-à-dire $F(\omega + z) = \lambda$, si et seulement si $F(\omega + \mu z) = F(\omega) + \mu^2\big(\lambda - F(\omega)\big)$. □

8.10. Interprétation géométrique de l'algorithme. — Supposons que par application de l'algorithme de la plus profonde descente on ait obtenu, à la k-ième étape, le point $x_k \in \mathbb{R}^n$. Bien sûr, si $k = 0$, le point x_0 est un point quelconque de $\mathbb{R}^n$. Supposons $x_k \neq \omega$. Nous avons alors $r_k = \overrightarrow{\operatorname{grad}}\, F(x_k) = Ax_k - b \neq 0$. Soit $\lambda = F(x_k)$. Considérons la quadrique Q_λ; elle passe évidemment par le point x_k. Puisque Q_λ est l'ensemble des points de $\mathbb{R}^n$ où la fonction F prend la valeur λ, un vecteur $u \in \mathbb{R}^n$, considéré comme vecteur d'origine x_k, est tangent à la quadrique Q_λ au point x_k si et seulement si $DF(x_k)(u) = 0$, c'est-à-dire si et seulement si $\big(\overrightarrow{\operatorname{grad}}\, F(x_k) \;\big|\; u\big) = 0$. Nous avons donc prouvé que le vecteur $r_k = \overrightarrow{\operatorname{grad}}\, F(x_k)$ est normal à la quadrique Q_λ au point x_k. Par suite, le point x_{k+1}, auquel on impose d'être de la forme $x_k + s_k r_k$, est sur la droite D_k, qui passe par le point x_k et qui est normale, en ce point, à la quadrique Q_λ. De plus, on choisit s_k de manière que $F(x_{k+1})$ soit le plus petit possible. Cela prouve que la droite D_k est tangente en x_{k+1} à la quadrique Q_μ qui passe par ce point, avec bien sûr $\mu = F(x_{k+1})$.
La figure ci-dessous illustre le raisonnement.

Figure 1.

L'interprétation géométrique donnée ci-dessus nous permet aisément de prouver la proposition suivante.

8.11. Proposition. — *Lors de l'application de l'algorithme de plus profonde descente, pour tout $k \in \mathbb{N}$, les vecteurs r_k et r_{k+1} sont orthogonaux, c'est-à-dire vérifient $(r_k|r_{k+1}) = 0$.*

Preuve : D'après l'interprétation géométrique qui précède, r_{k+1} est normal en x_{k+1} à la quadrique Q_μ qui passe par x_{k+1}, tandis que la droite D_k, parallèle au vecteur r_k, est tangente en x_{k+1} à cette quadrique. Les vecteurs r_k et r_{k+1} sont donc orthogonaux. On va d'ailleurs vérifier ce résultat par un calcul direct. On a

$$(r_k|r_{k+1}) = \big(r_k \mid A(x_k + s_k r_k) - b\big) = (r_k|r_k) + s_k(r_k|Ar_k)$$
$$= (r_k|r_k) - \frac{(r_k|r_k)}{(Ar_k|r_k)}\,(r_k|Ar_k) = 0\,,$$

compte tenu de la définition même de s_k. □

9. La méthode du gradient conjugué

9.1. Principe de la méthode. — Dans la méthode du gradient conjugué, partant du point x_k, on cherche le point x_{k+1} sur la droite D_k, passant par x_k, et normale en ce point à la variété de niveau de F qui passe par x_k. C'est sur cette droite que la fonction F décroît initialement (tout près du point x_k) le plus vite possible, d'où le nom de la méthode. Mais lorsqu'on parcourt la droite D_k en partant de x_k et en allant vers x_{k+1}, la décroissance de F, qui est la plus rapide possible au début, se ralentit; puis, si on dépasse le point x_{k+1}, F se met à croître. On peut donc penser que choisir le point x_{k+1} sur la droite D_k n'est peut-être pas le meilleur choix possible : en le prenant sur une droite sur laquelle la décroissance initiale de F n'est pas la plus rapide possible au début, on pourra peut-être faire décroître F davantage en passant de x_k à x_{k+1}. La méthode du gradient conjugué est basée sur cette idée. Nous allons voir qu'elle possède une remarquable propriété, qui peut paraître assez inattendue lorsqu'on a oublié les propriétés géométriques élémentaires des ellipses et des ellipsoïdes : cette méthode aboutit à la solution exacte du problème après un nombre fini d'itérations (théorème de Stiefel). Croyant développer une méthode itérative, nous aurons donc l'heureuse surprise d'avoir développé une méthode directe!
Introduisons tout d'abord une définition.

9.2. Définition. — *Soit $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ une matrice symétrique. On dit que deux vecteurs u et $v \in \mathbb{R}^n$ sont conjugués par rapport à A s'ils vérifient*

$$(Au|v) = 0\,.$$

9.3. Commentaires

a) *Symétrie.* — En raison de la symétrie de A, on a

$$(Au|v) = (Av|u)\,.$$

La conjugaison est donc une relation symétrique.

b) *Directions conjuguées.* — Une *direction* dans $\mathbb{R}^n$ est une classe d'équivalence d'éléments de $\mathbb{R}^n \setminus \{0\}$, pour la relation d'équivalence suivante : deux points x et x', éléments de $\mathbb{R}^n \setminus \{0\}$, sont équivalentes s'il existe $\lambda \in \mathbb{R} \setminus \{0\}$ tel que $x' = \lambda x$.
On voit immédiatement que la conjugaison est en fait une relation d'équivalence entre directions dans $\mathbb{R}^n$: deux directions δ_1 et δ_2 dans $\mathbb{R}^n$ seront dites *conjuguées* si, x_1 et $x_2 \in \mathbb{R}^n \setminus \{0\}$ étant des représentants, respectivement, des directions δ_1 et δ_2, on a $(Ax_1|x_2) = 0$. On voit en effet que le résultat ne dépend pas du choix des représentants x_1 et x_2 des directions δ_1 et δ_2.

c) *Conjugaison par rapport à une quadrique.* — Soit $b \in \mathbb{R}^n$, et $\lambda \in \mathbb{R}$. On considère la quadrique (supposée non vide et non réduite à un point)

$$Q_\lambda = \left\{ x \in \mathbb{R}^n \ ; \ (Ax - 2b|x) = 2\lambda \right\}.$$

Deux directions δ_1 et δ_2 seront dites *conjuguées* par rapport à cette quadrique si elles sont conjuguées par rapport à A.
Lorsque A est inversible, cette définition a une interprétation géométrique simple. Soit $\omega = A^{-1}(b)$. Nous avons vu que c'est le centre de symétrie des quadriques Q_λ. D'après un calcul fait précédemment,

$$Q_\lambda = \left\{ x \in \mathbb{R}^n \ ; \ x = \omega + z\,,\ (Az|z) = 2\lambda - (1/2)(\omega|\omega) + (b|\omega) \right\}.$$

Soient alors deux directions δ_1 et δ_2 de $\mathbb{R}^n$, u_1 et $u_2 \in \mathbb{R}^n \setminus \{0\}$ des représentants, respectivement de δ_1 et de δ_2. Supposons que la droite passant par ω et parallèle à δ_1 rencontre la quadrique Q_λ en deux points, x_1 et x_1', évidemment symétriques l'un de l'autre par rapport à ω. Les hyperplans tangents en x_1 et en x_1' à Q_λ sont symétriques l'un de l'autre par rapport à ω, donc parallèles. On voit alors que les directions δ_1 et δ_2 sont conjuguées par rapport à Q_λ si et seulement si la direction δ_2 est parallèle aux hyperplans tangents à Q_λ aux points x_1 et x_1'.
Lorsque la droite passant par ω et parallèle à δ_1 ne rencontre pas Q_λ, on peut permuter les rôles de δ_1 et δ_2, ou encore raisonner dans $\mathbb{C}^n$ au lieu de $\mathbb{R}^n$, ou mieux dans l'espace projectif $\mathbb{PC}(n)$.

9.4. Principe de la méthode du gradient conjugué. — On revient au cas où la matrice A est définie positive. Comme celle de la plus profonde descente, la méthode du gradient conjugué consiste à partir d'un point arbitraire x_0 de $\mathbb{R}^n$, et à construire successivement des points $x_1, x_2, \ldots, x_k, \ldots$, de manière telle que la suite $\big(F(x_n)\,,\ n \in \mathbb{N}\big)$ soit décroissante. Ces points sont construits comme suit.
À l'étape p, on a déterminé un point x_p de $\mathbb{R}^n$, un vecteur non nul u_{p-1} de $\mathbb{R}^n$ tangent en x_p à la quadrique $Q_{F(x_p)}$ qui passe par ce point. On pose alors

$$r_p = \overrightarrow{\operatorname{grad}}\, F(x_p) = Ax_p - b\,.$$

Ainsi qu'on l'a vu, le vecteur r_p est normal en x_p à la quadrique $Q_{F(x_p)}$.
Dans la méthode de plus profonde descente, on cherchait le point suivant x_{p+1} sur la droite passant par x_p et parallèle à r_p. Dans celle du gradient conjugué, on cherche d'abord un

vecteur non nul u_p de $\mathbb{R}^n$, de la forme

$$u_p = -r_p + s_p u_{p-1}\,,$$

avec $s_p \in \mathbb{R}$. Un vecteur de cette forme est élément du plan déterminé par les deux vecteurs r_p et u_{p-1}. De plus, il est non nul et dirigé vers l'intérieur de la quadrique $Q_{F(x_p)}$, puisque r_p est non nul,normal à cette quadrique et dirigé vers l'extérieur de cette quadrique et que u_{p-1} est tangent à cette quadrique, donc orthogonal à r_p. On détermine s_p en imposant à u_{p-1} et u_p d'être conjugués par rapport à la quadrique $Q_{F(x_p)}$, ce qui s'exprime par

$$(Au_p|u_{p-1}) = 0\,,$$

ce qui donne

$$s_p = \frac{(Ar_p|u_{p-1})}{(Au_{p-1}|u_{p-1})}\,.$$

On cherche alors le point suivant x_{p+1} sur la droite passant par x_p et parallèle à u_p :

$$x_{p+1} = x_p + t_p u_p\,, \qquad \text{avec} \quad t_p \in \mathbb{R}\,.$$

On détermine le réel t_p en imposant à F d'atteindre en x_{p+1} son minimum sur la droite passant par x_p et parallèle à u_p. On écrit pour cela

$$F(x_{p+1}) = \frac{1}{2}\,(Ax_{p+1} - 2b|x_{p+1}) = F(x_p) + (r_p|u_p)t_p + \frac{1}{2}\,(Au_p|u_p)t_p^2\,.$$

On écrit que la dérivée de cette expression par rapport à t_p est nulle, ce qui conduit à

$$t_p = -\,\frac{(r_p|u_p)}{(Au_p|u_p)} = +\,\frac{(r_p|r_p)}{(Au_p|u_p)}\,,$$

compte tenu des égalités

$$(r_p|u_p) = (r_p|\,-r_p + s_p u_{p-1}) = -(r_p|r_p)\,,$$

provenant du fait que les vecteurs u_{p-1} et r_p sont orthogonaux, le premier étant tangent et le second normal à la quadrique $Q_{F(x_p)}$ au point x_p.

9.5. L'algorithme du gradient conjugué. — À l'étape $p \geq 1$, on dispose du couple $(x_p, u_{p-1}$, constitué d'un point x_p de $\mathbb{R}^n$ et d'un vecteur non nul u_p de cet espace. On détermine alors le couple (x_{p+1}, u_p) grâce aux formules

$$r_p = Ax_p - b\,, \qquad s_p = \frac{(Ar_p|u_{p-1})}{(Au_{p-1}|u_{p-1})}\,,$$

$$u_p = -r_p + s_p u_{p-1}\,, \qquad t_p = \frac{(r_p|r_p)}{(Au_p|u_p)}\,,$$

$$x_{p+1} = x_p + t_p u_p\,.$$

L'algorithme s'arrête lorsqu'on trouve, à une certaine étape, un vecteur r_p nul. En effet, par hypothèse, à l'étape précédente, u_{p-1} est non nul, u_{p-1} et r_p sont orthogonaux, donc si r_p est non nul, $u_p = -r_p + s_p u_{p-1}$ est non nul, on peut donc faire une étape de plus et déterminer x_{p+1}.

Pour initialiser l'algorithme, on prend un point quelconque x_0 de $\mathbb{R}^n$. On détermine alors successivement

$$r_0 = Ax_0 - B\,,$$
$$u_0 = -r_0\,, \qquad t_0 = \frac{(r_0|r_0)}{(Au_0|u_0)}\,,$$
$$x_1 = x_0 + t_0 u_0\,.$$

La première étape est donc la même que dans l'algorithme de la plus profonde descente, mais à partir de la deuxième étape, les deux algorithmes sont différents.

Le lemme ci-dessous va nous permettre de prouver que l'algorithme du gradient conjugué conduit à la solution exacte du système linéaire $Ax = b$ en un nombre fini d'étapes.

9.6. Lemme. — *Les suites de vecteurs u_{-1}, r_0, u_0, r_1, u_1, r_2, ... (avec, par convention, $u_{-1} = 0$), construites par application de l'algorithme du gradient conjugué, vérifient, pour tout couple d'entiers (p, q) vérifiant $0 \leq q < p$,*

$$(r_p|r_q) = 0\,, \tag{1}$$
$$(Ar_p|u_{q-1}) = 0\,, \tag{2}$$
$$(Au_p|u_q) = 0\,, \tag{3}$$
$$(Au_p|r_q) = 0\,. \tag{4}$$

Preuve : Rappelons d'abord que les formules qui font passer de (u_{p-1}, r_p) à (u_p, r_{p+1}) sont

$$s_p = \begin{cases} \dfrac{(Ar_p|u_{p-1})}{(Au_{p-1}|u_{p-1})} & \text{pour } p \geq 1, \\ 0 & \text{pour } p = 0, \end{cases} \qquad u_p = -r_p + s_p u_{p-1}\,,$$
$$t_p = \frac{(r_p|r_p)}{(Au_p|u_p)}\,, \qquad r_{p+1} = r_p + t_p Au_p\,.$$

Cette dernière formule résulte en effet de

$$r_{p+1} = Ax_{p+1} - b = A(x_p + t_p u_p) - b = Ax_p - b + t_p Au_p = r_p + t_p Au_p\,.$$

Faisons $p = 1$, $q = 0$. Nous avons

$$\begin{aligned}(r_1|r_0) &= (r_0 + t_0 Au_0|r_0) = (r_0|r_0) + t_0(Au_0|r_0) \\ &= (r_0|r_0) - t_0(Au_0|u_0) = (r_0|r_0) - \frac{(r_0|r_0)}{(Au_0|u_0)}\,(Au_0|u_0) \\ &= 0\,.\end{aligned}$$

La formule (1) est donc vérifiée. De même,

$$(Ar_1|u_{-1}) = 0\,,$$

car par convention $u_{-1} = 0$. La formule (2) est donc vérifiée. Nous avons aussi

$$(Au_1|u_0) = \big(A(-r_1 + s_1 u_0) \;\big|\; u_0\big) = -(Ar_1|u_0) + \frac{(Ar_1|u_0)}{(Au_0|u_0)}\,(Au_0|u_0) = 0\,.$$

La formule (3) est donc vérifiée. Enfin,

$$(Au_1|r_0) = (Au_1| - u_0) = 0\,,$$

cas $s_0 = 0$, donc $r_0 = -u_0$. La formule (4) est donc vérifiée.
Supposons maintenant les formules (1) à (4) vérifiées pour tous p, q tels que $0 \leq q < p \leq h$. Montrons qu'elles sont encore vérifiées pour tous p, q tels que $0 \leq q < p \leq h+1$. Nous avons

$$(r_{h+1}|r_q) = (r_h + t_h Au_h|r_q) = (r_h|r_q) + t_h(Au_h|r_q)\,.$$

Si $0 \leq q < h$, les deux termes du second membre sont nuls car d'après les hypothèses de récurrence, on peut appliquer les formules (1) et (4). Si $q = h$ on a, d'après l'expression de t_h,

$$(r_{h+1}|r_h) = (r_h|r_h) + \frac{(r_h|r_h)}{(Au_h|u_h)}\,(Au_h|r_h)\,.$$

Mais

$$(Au_h|u_h) = (Au_h| - r_h + s_h u_{h-1}) = -(Au_h|r_h) + s_h(Au_h|u_{h-1}) = -(Au_h|r_h)\,,$$

car d'après l'hypothèse de récurrence, $(Au_h|u_{h-1}) = 0$. D'où

$$(r_{h+1}|r_h) = 0\,.$$

On a ainsi prouvé que la formule (1) est vérifiée pour tous p et q tels que $0 \leq q < p \leq h+1$. De même, pour la formule (2),

$$(Ar_{h+1}|u_{q-1}) = (r_{h+1}|Au_{q-1})\,,$$

puisque A est symétrique. Si $q = 0$, u_{q-1} est nul par convention, donc $(Ar_{h+1}|u_{q-1})$ est nul. Si $q \geq 1$, $r_{q-1} \neq 0$, donc $t_{q-1} \neq 0$ et nous avons

$$Au_{q-1} = \frac{1}{t_{q-1}}\,(r_q - r_{q-1})\,.$$

On en déduit

$$(Ar_{h+1}|u_{q-1}) = \frac{1}{t_{q-1}}\,(r_{h+1}|r_q - r_{q-1}) = 0\,,$$

car d'après le résultat qui précède, la formule (1) est vérifiée pour tous p et q tels que $0 \leq q < p \leq h+1$. Donc la formule (2) est aussi vérifiée pour $0 \leq q < p \leq h+1$.
De même, pour la formule (3),

$$\begin{aligned}(Au_{h+1}|u_q) &= \big(A(-r_{h+1} + s_{h+1}u_h) \;\big|\; u_q\big)\\ &= -(Ar_{h+1}|u_q) + s_{h+1}(Au_h|u_q)\,.\end{aligned}$$

Pour $0 \leq q < h$, le premier terme du dernier membre est nul car nous venons de prouver que la formule (2) est vraie pour $0 \leq q < p \leq h+1$. Le second terme étant nul aussi d'après l'hypothèse de récurrence, nous voyons que la formule (3) est vérifiée pour $0 \leq q < p \leq h+1$ et $q < h$. Pour $q = h$ et $p = h+1$ on peut écrire, d'après l'expression de s_{h+1},

$$(Au_{h+1}|u_h) = -(Ar_{h+1}|u_h) + \frac{(Ar_{h+1}|u_h)}{(Au_h|u_h)}\,(Au_h|u_h) = 0\,.$$

C'est, en effet, la relation qui exprime que u_{h+1} et u_h sont conjugués. Nous voyons donc que la formule (3) est vérifiée pour $0 \leq q < p \leq h+1$.
Enfin, pour la formule (4),

$$(Au_{h+1}|r_q) = (Au_{h+1}| - u_q + s_q u_{q-1}) = -(Au_{h+1}|u_q) + s_q(Au_{h+1}|u_{q-1})\,.$$

Le premier terme du dernier membre est nul car on vient de prouver que la formule (3) est vraie pour $0 \leq q < p \leq h+1$. Si $q \geq 0$, le second terme est nul pour la même raison et, si $q = 0$, il est nul aussi car par convention $u_{-1} = 0$. Nous avons donc prouvé que la formule (4) est vraie pour $0 \leq q < p \leq h+1$, ce qui achève la démonstration du lemme par récurrence. □

9.7. Théorème de Stiefel. — *L'algorithme du gradient conjugué aboutit à la solution exacte du système $Ax = b$ en au plus n étapes.*

Preuve : On a vu que l'algorithme du gradient conjugué ne pouvait s'arrêter à une certaine étape p que lorsqu'on obtient, à cette étape, $r_p = 0$. Supposons que l'algorithme ne soit pas encore arrêté à l'étape k. On a donc une suite finie r_0, r_1, …, r_k d'éléments de $\mathbb{R}^n$, tous non nuls et deux à deux orthogonaux. Ces vecteurs forment une famille libre (c'est une conséquence immédiate du fait qu'ils sont tous non nuls et deux à deux orthogonaux), dans l'espace $\mathbb{R}^n$, de dimension n. Par suite, $k \leq n-1$. On atteint donc nécessairement une étape p (avec $p \leq n$) à laquelle on obtient $r_p = 0$. Mais par définition $r_p = Ax_p - b$, donc x_p est solution du système linéaire $Ax = b$. □

9.8. Remarque. — Dans la pratique, en raison de la précision nécessairement finie des calculs numériques en nombres réels, on n'arrive que très exceptionnellement à obtenir, à une certaine étape, un $x_p \in \mathbb{R}^n$ vérifiant exactement $Ax_p = b$. On doit cependant savoir que si, après plus de n étapes, on obtient toujours des r_p non négligeables, c'est que les erreurs d'arrondi dans l'application de la méthode la rendent inopérante à partir de la valeur arbitrairement choisie comme point de départ x_0. Il peut arriver qu'en poussant plus loin le calcul, c'est-à-dire en poursuivant le calcul pendant plus de n étapes, on aboutisse finalement à un résultat satisfaisant; mais dans ce cas seules les n dernières étapes de l'algorithme sont vraiment significatives et, pour étudier la validité de la solution trouvée, on devra examiner attentivement la précision des calculs effectués lors de ces n dernières étapes.

10. La méthode de Newton

10.1. Description de la méthode. — Soit U un ouvert de $\mathbb{R}$ et $f : U \to \mathbb{R}$ une fonction continue sur U, à valeurs réelles. On cherche les solutions de l'équation

$$f(x) = 0\,. \qquad (*)$$

La *méthode de Newton* permet de déterminer, par approximations successives, une solution $\omega \in U$ de cette équation, dont on suppose l'existence. Elle est applicable lorsque la fonction

f est de classe C^1 sur U et lorsque sa dérivée f' ne s'annule pas sur U. En restreignant U si nécessaire, on peut satisfaire ces conditions lorsque f est de classe C^1 sur un voisinage du point ω et lorsque sa dérivée f' ne s'annule pas sur ce voisinage. Nous allons exposer son principe en supposant f de classe C^1 sur U, mais sans supposer que f' ne s'annule pas sur U, car nous verrons qu'il peut arriver que la méthode s'applique même lorsque f' s'annule en cetains points de U.

La méthode de Newton consiste à prendre, pour point de départ, un point x_0 de U; on calcule $f(x_0)$ et, si $f(x_0) \neq 0$, on calcule aussi $f'(x_0)$; puis, si $f'(x_0) \neq 0$, on pose

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}\,.$$

Cette formule a une interprétation géométrique simple : la tangente au graphe de la fonction f, au point $\big(x_0, f(x_0)\big)$, est la droite D_0 d'équation

$$y - f(x_0) = f'(x_0)(x - x_0)\,.$$

Nous voyons donc que x_1 est l'abscisse du point d'intersection de la droite D_0 avec l'axe des abscisses.

Supposons x_1 élément de U. Calculons $f(x_1)$. Si $f(x_1) = 0$, nous avons trouvé une solution $\omega = x_1$ de l'équation $(*)$. Dans le cas contraire, calculons $f'(x_1)$ et (en supposant $f'(x_1) \neq 0$), posons

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}\,.$$

De proche en proche, nous construisons ainsi une suite $(x_0\,,\ x_1\,,\ \ldots\,,\ x_k\,,\ \ldots)$ dans U. En supposant les x_i déterminés jusqu'à $i = k$, on peut faire une itération de plus et calculer x_{k+1} si $x_k \in U$ et si $f'(x_k) \neq 0$. Losrsque ces conditions sont satisfaites, x_{k+1} est donné par la formule

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}\,. \qquad (**)$$

Posons

$$N(x) = x - \frac{f(x)}{f'(x)}\,.$$

L'application N, appelée *application de Newton*, est définie sur le complémentaire dans U de l'ensemble des points $x \in U$ tels que $f'(x) = 0$. On voit que $x_k = N(x_{k-1})$. On dit que la suite $(x_0\,,\ x_1\,,\ \ldots\,,\ x_k\,,\ \ldots)$ est l'orbite, issue du point x_0, du *système dynamique discret* obtenu par itération de l'application de Newton N.

Lorsqu'à une certaine étape on obtint un point $x_k \in U$ tel que $f(x_k) = 0$, il est inutile de poursuivre le calcul. D'ailleurs, si $f'(x_k) \neq 0$, on obtient $x_n = x_k$ pour tout $n \geq k$: la suite devient stationnaire à partir du rang k. On a alors trouvé une solution $\omega = x_k$ de l'équation $(*)$. Cela n'arrive qu'exceptionnellement.

L'algorithme s'arrête pour une valeur finie de k, pour laquelle $x_k \in U$, lorsque $f(x_k) \neq 0$ et $f'(x_k) = 0$, ou lorsque x_{k+1}, calculé par application de la formule $(**)$, n'est pas

élément de U. Dans un tel cas, l'algorithme de Newton ne conduit pas à une solution de l'équation $(*)$, du moins avec le point de départ x_0 initialement choisi.
Il peut aussi arriver que par application de l'algorithme de Newton, on obtienne une suite infinie contenue dans U qui ne converge pas dans U. Mais lorsque cette suite infinie converge vers un élément de U, cet élément est solution de $(*)$. Cela résulte en effet de la proposition suivante.

10.2. Proposition. — *Supposons que l'application de l'algorithme de Newton, à partir d'un élément initial x_0 de U, ait permis de construire une suite $(x_n\,,\ n \in \mathbb{N})$ d'éléments de U qui converge vers une limite $l \in U$. Alors $f(l) = 0$, donc $\omega = l$ est une solution de l'équation $(*)$.*

Preuve : Nous avons

$$\frac{f(x_n)}{f'(x_n)} = x_n - x_{n+1}\,.$$

Comme par hypothèse la suite $(x_n\,,\ n \in \mathbb{N})$ converge vers un réel l, la formule ci-dessus montre que $\lim_{n\to+\infty} f(x_n)/\big(f'(x_n)\big) = 0$. Mais par hypothèse f est de classe C^1 sur U, donc f et f' sont continues sur U. Comme par hypothèse $l \in U$, $f(x_n)$ et $f'(x_n)$ convergent vers $f(l)$ et $f'(l)$, respectivement. La formule ci-dessus montre alors que nécessairement $f(l) = 0$. □

Le théorème suivant indique des conditions suffisantes de convergence de l'algorithme de Newton.

10.3. Théorème. — *Soit U un ouvert de $\mathbb{R}$ et $f : U \to \mathbb{R}$ une application de classe C^2. On suppose qu'il existe un élément ω de U tel que $f(\omega) = 0$ et $f'(\omega) \neq 0$. Alors il existe un réel $\eta > 0$ ayant les propriétés suivantes :*

(i) *l'intervalle $[\omega - \eta, \omega + \eta]$ est contenu dans U, la dérivée f' de f ne s'annule pas dans cet intervalle, et le point ω est l'unique solution de l'équation $f(x) = 0$ contenue dans cet intervalle ;*

(ii) *pour tout $x_0 \in [\omega - \eta, \omega + \eta]$, l'application de l'algorithme de Newton avec x_0 comme point initial donne une suite infinie $(x_n\,,\ n \in \mathbb{N})$, contenue dans l'intervalle $[\omega - \eta, \omega + \eta]$, qui converge vers le point ω ; de plus, la convergence est au moins aussi rapide que la convergence vers 0 d'une progression géométrique de raison strictement inférieure à 1.*

Preuve : Puisque U est ouvert, que f' est continue et que $f'(\omega) \neq 0$, il existe $\eta_1 > 0$ tel que l'intervalle $[x - \eta_1, x + \eta_1]$ soit contenu dans U et que f' ne s'annule pas sur cet intervalle. La fonction f étant strictement monotone sur cet intervalle, ω est l'unique racine de l'équation $f(x) = 0$ contenue dans cet intervalle. Nous voyons qu'en choisissant un réel η vérifiant $0 < \eta \leq \eta_1$, les conditions (i) seront satisfaites.

Puisque f est de classe C^2, les dérivées première f' et f'' de f sont continues sur le compact $[\omega-\eta,\omega+\eta]$. De plus, f' ne s'annule pas sur ce compact. Par suite, nous avons

$$M = \inf_{x\in[\omega-\eta,\omega+\eta]} \big|f'(x)\big| > 0\,, \qquad m = \sup_{x\in[\omega-\eta,\omega+\eta]} \big|f''(x)\big| < +\infty\,.$$

Choisissons un réel $\eta > 0$ vérifiant

$$\eta \le \eta_1 \qquad \text{et} \qquad r = \frac{m\eta}{2M} < 1\,.$$

Nous allons prouver que η répond à la question. Supposons que l'application de l'algorithme de Newton, à partir d'un point initial $x_0 \in [\omega-\eta,\omega+\eta]$, ait permis de construire les x_i jusqu'à $i = k$ et que ces points appartiennent à l'intervalle $[\omega-\eta,\omega+\eta]$. Puisque $f'(x_k) \neq 0$, nous pouvons définie x_{k+1} et nous avons, compte tenu de $f(\omega) = 0$,

$$x_{k+1} - \omega = x_k - \frac{f(x_k)}{f'(x_k)} - \omega = \frac{f(\omega) - f(x_k) - f'(x_k)(\omega - x_k)}{f'(x_k)}\,.$$

Mais d'après la formule des accroissements finis avec reste de Lagrange,

$$\begin{aligned}\big|f(\omega) - f(x_k) - f'(x_k)(\omega-x_k)\big| &\le \frac{1}{2}\sup_{0\le t\le 1}\big|f''\big(x_k + t(\omega-x_x)\big)\big|\,|\omega-x_k|^2\\ &\le \frac{m\eta}{2}\,|\omega-x_k|\,.\end{aligned}$$

Nous en déduisons

$$|x_{k+1}-\omega| \le \frac{m\eta}{2\big|f'(x_k)\big|}\,|x_k-\omega| \le \frac{m\eta}{2M}\,|x_k-\omega| = r|x_k-\omega|\,.$$

Comme $0 \le r < 1$, cette dernière inégalité prouve que $x_{k+1} \in [\omega-\eta,\omega+\eta]$. Ainsi, de proche en proche, nous voyons que l'application de l'algorithme de Newton à partir du point x_0 permet de construire une suite infinie $(x_n\,,\ n\in\mathbb{N})$ contenue dans l'intervalle $[\omega-\eta,\omega+\eta]$, et dont le terme général vérifie $|x_n-\omega| \le r^n|x_0-\omega|$. Nous pouvons conclure que la suite $(x_n\,,\ n\in\mathbb{N})$ converge vers ω, au moins aussi vite que la progression géométrique de raison r converge vers 0. □

10.4. La méthode de Newton pour les fonctions de plusieurs variables. — Soient E et F deux espaces vectoriels réel de même dimension finie n, U un ouvert de E et $f : U \to F$ une application de classe C^1. Pour tout $x \in U$, nous notons $Df(x)$ la différentielle de f au point x. Rappelons que c'est une application linéaire de E dans F.
Pour tout $x \in U$ tel que $Df(x)$ soit un isomorphisme de E sur F, posons

$$N(x) = x - \big(Df(x)\big)^{-1}\big(f(x)\big)\,,$$

où nous avons noté $\big(Df(x)\big)^{-1}$ l'isomorphisme de F sur E inverse de l'isomorphisme $Df(x)$. L'application N, ainsi définie sur l'ouvert formé par les éléments $x \in U$ tels que $Df(x)$ soit un isomorphisme, est appelée *application de Newton*.
Pour résoudre, par approximations successives, l'équation

$$f(x) = 0\,,$$

l'algorithme de Newton consite à construire une suite $(x_0\,,\ x_1\,,\ \ldots\,,\ x_k\,,\ \ldots)$ en choisissant un point $x_0 \in U$ et en posant, chaque fois que les x_i ont déjà été définis jusqu'à $i = k$, qu'ils sont éléments de U et que $Df(x_k)$ est un isomorphisme,

$$x_{k+1} = N(x_k) = x_k - \big(Df(x_k)\big)^{-1}\big(f(x_k)\big)\,.$$

Il est facile de voir que la proposition 10.2 et le théorème 10.3 restent applicables moyennant des adaptations évidentes : il suffit de munir E et F de normes et les espaces d'applications linéaires $\mathcal{L}(E,F)$ et $\mathcal{L}(F,E)$ des normes associées, de remplacer judicieusement les valeurs absolues qui apparaissent dans les énoncés et les démonstrations de 10.2 et 10.3 par des expressions formées au moyen de ces normes, et de remplacer dans 10.3 l'intervalle $[\omega-\eta,\omega+\eta]$ par la boule fermée de centre ω et de rayon η.

10.5. La méthode de Newton-Raphson. — La méthode de Newton pour une fonction de plusieurs variables nécessite, à chaque itération, le calcul de l'inverse de l'isomorphisme $Df(x_k)$. On utilise souvent en pratique des variantes de la méthode de Newton dans lesquelles on évite d'avoir à faire ce calcul (assez lourd) à chaque itération. On peut, par exemple, calculer $B = \big(Df(x_0)\big)^{-1}$ seulement pour le point initial x_0, et poser (en supposant que les x_i ont été déterminés jusqu'à $i = k$ et qu'ils appartiennent à l'ouvert U),

$$x_{k+1} = x_k - B\big(f(x_k)\big)\,.$$

On peut aussi convenir de calculer $\big(Df(x_k)\big)^{-1}$ non pas à chaque itération, mais seulement une fois sur 10 (ou une fois sur 100), et faire comme si $\big(Df(x_k)\big)^{-1}$ restait constant pendant 10 itérations (ou 100 itérations, respectivement).

Le lecteur intéressé trouvera une étude de la convergence de ces généralisations de la méthode de Newton dans le livre de P. G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, Paris, 1990.

Chapitre IV

Résolution numérique des équations différentielles

1. Méthodes numériques approchées

Nous avons étudié jusqu'à présent deux types de méthodes numériques : les méthodes directes, qui conduisent à la solution exacte du problème (aux erreurs dues à la précision finie des calculs numériques en nombres réels près), et les méthodes itératives, qui en principe permettent, en augmentant le nombre d'itérations, d'obtenir un résultat aussi proche qu'on le veut de la solution exacte. Remarquons au passage que dans la pratique, les erreurs d'arrondi augmentant avec le nombre d'itérations, les méthodes itératives ne permettent pas vraiment d'approcher arbitrairement près de la solution exacte.
On va maintenant étudier un exemple de méthode d'un autre type : une méthode approchée, qui résoud le problème considéré seulement de manière approximative. Les méthodes de ce type comportent un paramètre que l'on peut librement choisir; par ce choix, on peut, au prix d'un allongement des calculs, diminuer l'écart entre la solution exacte et l'approximation fournie par la méthode. En appliquant la méthode plusieurs fois, avec des valeurs décroissantes (ou croissantes, selon le cas) du paramètre, on obtiendrait, comme avec les méthodes itératives, une suite d'approximations de la solution exacte, convergeant (si la méthode est bonne) vers cette solution exacte. Mais contrairement à ce qui a lieu avec les méthodes itératives, chaque application d'une méthode approchée nécessite de reprendre le problème à son début : on ne peut pas prendre pour point de départ le résultat obtenu lors du calcul approché précédent.
L'exemple de méthode approchée que nous allons étudier concerne la résolution des équations différentielles. D'autres exemples de méthodes approchées sont la méthode des trapèzes pour le calcul d'intégrales, et les méthodes de calcul de transformées de Fourier.

2. Rappel sur les équations différentielles

2.1. Définitions. —

1. *Une équation différentielle sous forme canonique, dans un espace vectoriel E (réel, de dimension finie) est une équation de la forme*

$$\varphi'(t) = f\big(t, \varphi(t)\big)\,, \tag{1}$$

où f est une application d'un ouvert Ω de $\mathbf{R} \times E$ dans E.

2. *Une solution de cette équation est une application φ d'un intervalle ouvert I de* $\mathbf{R}$ *dans E, continue et dérivable en tout point t de I, telle que, pour tout $t \in I$, sa dérivée $\varphi'(t)$ en ce point, soit égale à $f\big(t, \varphi(t)\big)$.*
3. *Une donnée de Cauchy pour l'équation différentielle (1) est un point (t_0, x_0) de Ω. On dit qu'une solution $\varphi : I \to E$ satisfait cette donnée de Cauchy si $t_0 \in I$ et $\varphi(t_0) = x_0$.*
4. *Une solution $\varphi : I \to E$ de l'équation (1) est dite maximale si toute autre solution $\psi : J \to E$, définie sur un intervalle ouvert J de* $\mathbf{R}$ *contenant I, et telle que $\psi\,\big|_I = \varphi$, est en fait égale à $\varphi : J = I$, $\psi = \varphi$. En d'autres termes, une solution est dite maximale lorsqu'on ne peut pas agrandir l'intervalle sur lequel elle est définie.*

On indique ci-dessous, sans démonstration, le très important théorème de Cauchy-Lipschitz (enseigné dans l'unité C2 de la licence).

2.2. Théorème de Cauchy-Lipschitz. — *Soit E un espace vectoriel de dimension finie, Ω un ouvert de* $\mathbf{R} \times E$*, et $f : \Omega \to E$ une application continue, et localement lipschitzienne par rapport à sa seconde variable. Cela signifie que tout point (t_0, x_0) de Ω possède un voisinage U (dans* $\mathbf{R} \times E$*, $U \subset \Omega$), tel qu'il existe $k \in \mathbf{R}$, vérifiant la propriété suivante :*

– *si (t, x) et $(t, x') \in U$, alors $\big\|f(t, x) - f(t, x')\big\| \leq k\|x - x'\|$.*

Alors, pour toute donnée de Cauchy $(t_0, x_0) \in \Omega$, il existe une solution maximale unique de l'équation différentielle (1) vérifiant cette donnée de Cauchy. De plus, toute solution de (1) vérifiant cette donnée de Cauchy est une restriction de cette solution maximale à un sous-intervalle ouvert de son intervalle de définition.

2.3. Remarques

a) On a supposé l'espace vectoriel E muni d'une norme, afin de pouvoir écrire l'inégalité

$$\big\|f(t, x) - f(t, x')\big\| \leq k\|x - x'\|\,.$$

Le choix de cette norme n'a pas d'importance, car E étant supposé de dimension finie, on sait que toutes les normes sur cet espace sont équivalentes.

b) Le théorème de Cauchy-Lipschitz reste valable lorsque E est un espace de Banach de dimension infinie (mais dans ce cas on doit bien préciser la norme dont il est muni, pour laquelle il est complet).

c) Il existe d'autres théorèmes d'existence, pour E de dimension finie, applicables lorsque f est continue mais pas nécessairement localement lipschitzienne par rapport à sa seconde variable. Mais alors, en général, on ne peut pas affirmer l'unicité de la solution maximale satisfaisant une donnée de Cauchy spécifiée.

d) Dans de très nombreux cas, la fonction f est différentiable de classe C^1 (cela signifie qu'elle a des dérivées partielles par rapport à t et par rapport à chaque composante de x (une fois choisie une base de E), qui sont des fonctions continues de (t,x). L'inégalité des accroissements finis (étudiée dans l'unité de valeur C2 de la licence) permet de montrer que dans ce cas, f est localement lipschitzienne par rapport à l'ensemble des deux variables (t,x), donc, *a fortiori*, par rapport à sa seconde variable x. Le théorème de Cauchy-Lipschitz lui est alors applicable.

e) Supposons le théorème de Cauchy-Lipschitz applicable à l'équation (1), et soit $(t_0,x_0) \in \Omega$ une donnée de Cauchy. La solution maximale φ de (1) satisfaisant cette donnée de Cauchy est définie sur un intervalle ouvert $I =]\alpha,\beta[$ de $\mathbf{R}$, avec $\alpha < t_0 < \beta$. On peut, selon les cas, avoir α fini ou égal à $-\infty$, et de même β fini ou égal à $+\infty$. Il n'est pas toujours facile d'évaluer α et β. On montre cependant le résultat suivant :

– Pour tout $r > 0$, on note $B'(x_0,r) = \left\{ x \in E \;\middle|\; \|x - x_0\| \leq r \right\}$ la boule fermée de centre x_0 et de rayon r. Soient $l > 0$ et $r > 0$ tels que $[t_0, t_0+l] \times B'(x_0,r)$ soit contenu dans Ω. Si la restriction de f à $[t_0, t_0+l] \times B'(x_0,r)$ vérifie

$$\sup_{(t,x)\in[t_0,t_0+l]\times B'(x_0,r)} \big\|f(t,x)\big\| \leq \frac{r}{l}\,,$$

alors $t_0 + l < \beta$ et, pour tout $t \in [t_0, t_0+l]$, $\varphi(t) \in B'(x_0,r)$.

De même, si $[t_0 - l, t_0] \times B'(x_0,r)$ est contenu dans Ω et si la restriction de f à $[t_0 - l, t_0] \times B'(x_0,r)$ vérifie

$$\sup_{(t,x)\in[t_0-l,t_0]\times B'(x_0,r)} \big\|f(t,x)\big\| \leq \frac{r}{l}\,,$$

alors $\alpha < t_0 - l$ et, pour tout $t \in [t_0 - l, t_0]$, $\varphi(t) \in B'(x_0,r)$.

3. La méthode d'Euler

3.1. Hypothèses générales et notations. — On va s'intéresser, dans le présent paragraphe et les suivants, à des méthodes qui permettent la détermination approchée de la solution de l'équation différentielle (1) satisfaisant la donnée de Cauchy $(t_0,x_0) \in \Omega$, sur un intervalle $[t_0, t_0+T]$, avec $T > 0$. On pourrait aussi bien appliquer les mêmes méthodes pour la détermination approchée de cette solution sur l'intervalle $[t_0 - T, t_0]$. On supposera pour simplifier qu'il existe $r > 0$ tel que $[t_0, t_0+T] \times B'(x_0,r)$ soit contenu dans Ω et que

$$\sup_{(t,x)\in[t_0,t_0+T]\times B'(x_0,r)} \big\|f(t,x)\big\| \leq \frac{r}{T}\,.$$

D'après la propriété indiquée en 2.3.5, on est alors assuré de l'existence de la solution cherchée sur l'intervalle $[t_0, t_0+T]$. On note φ cette solution.

Soit $h > 0$, $h \leq T$. Soit $J(h)$ le plus grand entier k tel que $kh \leq T$. Pour tout entier i vérifiant $0 \leq i \leq J(h)$, on pose, pour alléger l'écriture,

$$t_i = t_0 + ih\,, \qquad \varphi_i = \varphi(t_i)\,,$$

où φ désigne la solution *exacte* de (1) vérifiant la donnée de Cauchy (t_0,x_0).

3.2. Formules d'Euler. — La méthode d'Euler consiste à prendre, pour valeur approchée des φ_i, les quantités Φ_i ($0 \leq i \leq J(h)$) définies par

$$\begin{cases} \Phi_0 = \varphi_0 = x_0\,, \\ \Phi_{i+1} = \Phi_i + hf(t_i, \Phi_i) \qquad \text{pour } 0 \leq i \leq J(h)\,. \end{cases}$$

Cette méthode a une interprétation simple. On a en effet, pour tous t et t' appartenant à l'intervalle de définition de φ,

$$\varphi(t') = \varphi(t) + \int_t^{t'} f\big(s, \varphi(s)\big)\, ds\,,$$

donc en particulier

$$\varphi_{i+1} = \varphi_i + \int_{t_i}^{t_{i+1}} f\big(s, \varphi(s)\big)\, ds\,.$$

On voit donc que les valeurs approchées Φ_i sont obtenues en remplaçant la fonction intégrée $f\big(s, \varphi(s)\big)$ sur l'intervalle $[t_i, t_{i+1}]$ par la constante $f(t_i, \Phi_i)$. On a en effet :

$$\Phi_{i+1} = \Phi_i + \int_{t_i}^{t_{i+1}} f(t_i, \Phi_i)\, ds = \Phi_i + hf(t_i, \Phi_i)\,,$$

puisque $t_{i+1} - t_i = h$. La méthode d'Euler consiste à faire comme si la dérivée $\dfrac{d\varphi(s)}{ds} = f\big(s, \varphi(s)\big)$ de la solution φ gardait une valeur constante $f(t_i, \Phi_i)$ sur chaque intervalle $[t_i, t_{i+1}]$.

3.3. Théorème. — *L'approximation de la solution φ donnée par la méthode d'Euler converge uniformément vers cette solution lorsque h tend vers 0. Cela signifie que*

$$\lim_{h \to 0} \sup_{0 \leq i \leq J(h)} \|\Phi_i - \varphi_i\| = 0\,.$$

La démonstration de ce théorème repose sur le lemme suivant.

3.4. Lemme de Gronwall discret. — *Soient (a_i) et (b_i) deux suites de réels ≥ 0, avec $0 \leq i \leq N$. On suppose qu'il existe une constante $\lambda \geq 0$ telle que, pour tout i, $0 \leq i \leq N-1$, on ait*

$$a_{i+1} \leq (1 + \lambda) a_i + b_i\,.$$

Alors on a, pour tout i, $0 \leq i \leq N$,

$$a_i \leq e^{\lambda_i} a_0 + \sum_{j=0}^{i-1} e^{\lambda(i-j-1)} b_j\,.$$

Preuve : Puisque $\lambda \geq 0$, on a

$$1 + \lambda \leq e^{\lambda}\,,$$

donc, pour tout j ($0 \leq j \leq N-1$),

$$a_{j+1} \leq e^{\lambda} a_j + b_j\,.$$

Posons

$$a_j = \gamma_j e^{\lambda j} .$$

L'inégalité précédente devient

$$\gamma_{j+1} e^{\lambda(j+1)} \leq \gamma_j e^{\lambda(j+1)} + b_j ,$$

ou

$$\gamma_{j+1} \leq \gamma_j + e^{-\lambda(j+1)} b_j .$$

En additionnant ces inégalités pour j allant de 0 à $i-1$, on obtient

$$\gamma_i \leq \gamma_0 + \sum_{j=0}^{i-1} e^{-\lambda(j+1)} b_j ,$$

ou en revenant aux a_i et en remarquant que $a_0 = \gamma_0$,

$$a_i \leq e^{\lambda_i} a_0 + \sum_{j=0}^{i-1} e^{\lambda(i-j-1)} b_j .$$

□

Preuve du théorème 3.3: On peut écrire, pour tout i ($0 \leq i \leq J(h)-1$),

$$\Phi_{i+1} = \Phi_i + h f(t_i, \Phi_i) ,$$
$$\varphi_{i+1} = \varphi_i + h f(t_i, \varphi_i) + \epsilon_i ,$$

où on a posé

$$\epsilon_i = \int_{t_i}^{t_{i+1}} f\big(s, \varphi(s)\big)\, ds - h f(t_i, \varphi_i) .$$

On en déduit

$$\Phi_{i+1} - \varphi_{i+1} = \Phi_i - \varphi_i + h\big(f(t_i, \Phi_i) - f(t_i, \varphi_i)\big) - \epsilon_i ,$$

d'où

$$\|\Phi_{i+1} - \varphi_{i+1}\| \leq \|\Phi_i - \varphi_i\| + h\|f(t_i, \Phi_i) - f(t_i, \varphi_i)\| + \|\epsilon_i\| .$$

Mais puisque f satisfait les hypothèses du théorème de Cauchy-Lipschitz, elle est localement lipschitzienne par rapport à sa seconde variable. On supposera pour simplifier qu'elle est lipschitzienne par rapport à sa seconde variable (un raisonnement basé sur la compacité de $\varphi\big([t_0, t_0+T]\big)$ permettrait d'étendre le résultat au cas où f est seulement localement lipschitzienne par rapport à sa seconde variable). Il existe donc une constante $K \geq 0$ telle que

$$\big\|f(t_i, \Phi_i) - f(t_i, \varphi_i)\big\| \leq K\|\Phi_i - \varphi_i\| ,$$

d'où

$$\|\Phi_{i+1} - \varphi_{i+1}\| \leq (1 + Kh)\|\Phi_i - \varphi_i\| + \|\epsilon_i\| .$$

En appliquant le lemme de Gronwall discret, on en déduit que pour tout i, $0 \leq i \leq J(h)$,

$$\|\Phi_i - \varphi_i\| \leq e^{hKi}\|\Phi_0 - \varphi_0\| + \sum_{j=0}^{i-1} e^{hK(i-j-1)}\|\epsilon_j\| .$$

Mais on a

$$hKi \leq hKJ(h)\,, \qquad hK(i-j-1) \leq hKJ(h)\,,$$

donc, puisque $hJ(h) \leq T$,

$$hKi \leq KT\,, \qquad hK(i-j-1) \leq KT\,.$$

On a donc

$$\|\Phi_i - \varphi_i\| \leq e^{KT}\Big(\|\Phi_0 - \varphi_0\| + \sum_{j=0}^{i-1} \|\epsilon_j\|\Big)\,.$$

Mais par hypothèse $\|\Phi_0 - \varphi_0\| = 0$. D'autre part on a

$$\begin{aligned}\|\epsilon_j\| &= \Big\|\int_{t_j}^{t_{j+1}} \Big(f\big(t,\varphi(t)\big) - f\big(t_i,\varphi(t_i)\big)\Big)\,dt\Big\| \\ &\leq \int_{t_j}^{t_{j+1}} \Big\|f\big(t,\varphi(t)\big) - f\big(t_i,\varphi(t_i)\big)\Big\|\,dt\,.\end{aligned}$$

La fonction $t \mapsto f\big(t,\varphi(t)\big)$ est continue sur l'intervalle $[t_0, t_0+T]$, qui est compact. Cette fonction est donc uniformément continue. Pour tout $\epsilon > 0$, il existe donc un $\eta > 0$ tel que, si t et $t' \in [t_0, t_0+T]$ vérifient $|t-t'| < \eta$, alors

$$\Big\|f\big(t,\varphi(t)\big) - f\big(t',\varphi(t')\big)\Big\| \leq \epsilon\,.$$

Par suite, si $h \leq \eta$, on a pour tout j, $0 \leq j \leq J(h)-1$,

$$\|\epsilon_j\| \leq \int_{t_j}^{t_{j+1}} \Big\|f\big(t,\varphi(t)\big) - f\big(t_j,\varphi(t_j)\big)\Big\|\,dt \leq h\epsilon\,.$$

On a donc

$$\sum_{j=0}^{i-1} \|\epsilon_j\| \leq ih\epsilon \leq J(h)h\epsilon \leq \epsilon T\,.$$

D'où

$$\|\Phi_i - \varphi_i\| \leq Te^{KT}\epsilon\,.$$

Ceci prouve que

$$\lim_{h\to 0} \sup_{0\leq i\leq J(h)} \|\Phi_i - \varphi_i\| = 0\,.$$

□

3.5. Remarques

a) La somme

$$\sum_{i=0}^{J(h)-1} hf\big(t_i,\varphi(t_i)\big)$$

est une somme de Cauchy-Riemann pour l'intégrale

$$\int_{t_0}^{t_0+T} f\big(t,\varphi(t)\big)\,dt\,.$$

La propriété démontrée ci-dessus, selon laquelle

$$\lim_{h\to 0} \sum_{i=0}^{J(h)-1} \|\epsilon_i\| = 0\,,$$

est une conséquence de théorèmes établis lors de l'étude de l'intégrale de Riemann d'une fonction continue.

b) Dans la pratique, la méthode d'Euler n'est pas utilisable sauf dans des cas très particuliers, sa convergence étant trop lente. C'est pourquoi nous allons étudier d'autres méthodes.

4. Notion générale de schéma à un pas

On considère l'équation différentielle, dans l'espace vectoriel E de dimension finie,

$$\varphi'(t) = f\big(t, \varphi(t)\,, \tag{1}$$

la fonction f satisfaisant les hypothèses du théorème de Cauchy-Lipschitz. On note φ la solution exacte de cette équation vérifiant la donnée de Cauchy $\varphi(t_0) = x_0$. Comme précédemment, on suppose que son domaine de définition contient l'intervalle $[t_0, t_0+T]$. On choisit un réel h vérifiant $0 < h < T$, et on note $J(h)$ le plus grand entier tel que $hJ(h) \leq T$. On note, pour tout entier i vérifiant $0 \leq i \leq J(h)$,

$$t_i = t_0 + ih\,, \qquad \varphi_i = \varphi(t_i)\,.$$

Comme dans le paragraphe précédent, on cherche des valeurs approchées Φ_i des φ_i.

4.1. Définition. — *On dit que les valeurs approchées Φ_i, des $\varphi_i = \varphi(t_i)$ $(0 \leq i \leq J(h))$ sont données par un schéma à un pas lorsqu'elles sont données par des formules de la forme*

$$\Phi_0 = \varphi_0\,,$$
$$\Phi_{i+1} = \Phi_i + hF(t_i, \Phi_i, h)\,, \quad 0 \leq i \leq J(h) - 1\,,$$

où F est une fonction des trois variables $t \in \mathbf{R}$, $x \in E$, $h \in \mathbf{R}^+$.

4.2. Exemple. — Le schéma d'Euler, étudié dans le paragraphe précédent, est un schéma à un pas, dans lequel la fonction F est

$$F(t, x, h) = f(t, x)\,.$$

4.3. Définition. — *Le schéma à un pas défini par la fonction F est dit stable s'il existe une constante $M \geq 0$ telle que, pour tous U_0 et $V_0 \in E$, tout $h \geq 0$ assez petit et toute suite (ϵ_i), $0 \leq i \leq J(h)$, d'éléments de E, les suites (U_i) et (V_i) définies par*

$$\begin{cases} U_{i+1} = U_i + hF(t_i, U_i, h)\,, \\ V_{i+1} = V_i + hF(t_i, V_i, h) + \epsilon_i\,, \end{cases}$$

vérifient, pour tout i $(0 \leq i \leq J(h))$,

$$\|U_i - V_i\| \leq M\Big(\|u_0 - V_0\| + \sum_{j=0}^{i-1} \|\epsilon_j\|\Big) .$$

4.4. Exemple. — Dans le paragraphe précédent, en appliquant le lemme de Gronwall discret, on a prouvé la stabilité du schéma d'Euler.

4.5. Définition. — *Le schéma à un pas défini par* F *est dit consistant avec l'équation* (1) *si pour toute solution exacte* φ *de cette équation on a*

$$\lim_{h\to 0} \sum_{i=0}^{J(h)-1} \Big\|\varphi(t_{i+1}) - \varphi(t_i) - hF\big(t_i, \varphi(t_i), h\big)\Big\| = 0 .$$

4.6. Définition. — *On dit que le schéma numérique donnant les* Φ_i $(0 \leq i \leq J(h))$ *comme valeurs approchées des* $\varphi_i = \varphi(t_i)$ *converge, sur* $[t_0, t_0 + T]$, *vers la solution exacte, si*

$$\lim_{h\to 0} \sup_{0\leq i\leq J(h)} \|\Phi_i - \varphi_i\| = 0 .$$

4.7. Théorème. — *Un schéma à un pas stable et consistant avec l'équation* (1) *est convergent vers la solution exacte de cette équation.*

Preuve : Posons $U_i = \Phi_i$, $V_i = \varphi_i$. En raison de la stabilité,

$$\|\Phi_i - \varphi_i\| \leq M\Big(\|\Phi_0 - \varphi_0\| + \sum_{j=0}^{i-1} \big\|\varphi_{j+1} - \varphi_j - hF(t_j, \varphi_j, h)\big\|\Big) .$$

Mais on prendra $\Phi_0 = \varphi_0$, et la somme des termes figurant au second membre tend vers 0 lorsque $h \to 0$ à cause de la consistance du schéma avec l'équation. □

On va maintenant indiquer des conditions nécessaires et suffisantes, ou simplement suffisantes, pour qu'un schéma à un pas soit consistant avec une équation, ou soit stable.

4.8. Théorème. — *Considérons un schéma à un pas défini par une fonction* F *continue. Ce schéma est consistant avec l'équation* (1) *si et seulement si*

$$F(t, u, 0) = f(t, u) .$$

Preuve : Notons φ la solution exacte de l'équation et posons

$$\epsilon_i = \varphi(t_{i+1}) - \varphi(t_i) - hF\big(t_i, \varphi(t_i), h\big) .$$

On a

$$\epsilon_i = \int_{t_i}^{t_{i+1}} \Big(f\big(t, \varphi(t)\big) - F\big(t_i, \varphi(t_i), h\big)\Big)\, dt .$$

On peut écrire

$$\begin{aligned} f\big(t,\varphi(t)\big) - F\big(t_i,\varphi(t_i),h\big) &= f\big(t,\varphi(t)\big) - f\big(t_i,\varphi(t_i)\big) \\ &\quad + f\big(t_i,\varphi(t_i)\big) - F\big(t_i,\varphi(t_i),0\big) \\ &\quad + F\big(t_i,\varphi(t_i),0\big) - F\big(t_i,\varphi(t_i),h\big)\,, \end{aligned}$$

d'où

$$\begin{aligned} \epsilon_i = &\int_{t_i}^{t_{i+1}} \Big(f\big(t,\varphi(t)\big) - f\big(t_i,\varphi(t_i)\big)\Big)\, dt \\ &+ h\Big(f\big(t_i,\varphi(t_i)\big) - F\big(t_i,\varphi(t_i),0\big)\Big) \\ &+ h\Big(F\big(t_i,\varphi(t_i),0\big) - F\big(t_i,\varphi(t_i),h\big)\Big)\,. \end{aligned} \qquad (*)$$

On en déduit la majoration :

$$\begin{aligned} \sum_{i=0}^{J(h)-1} \|\epsilon_i\| \le &\sum_{i=0}^{J(h)-1} \int_{t_i}^{t_{i+1}} \Big\| f\big(t,\varphi(t)\big) - f\big(t_i,\varphi(t_i)\big)\Big\|\, dt \\ &+ h \sum_{i=0}^{J(h)-1} \Big\| f\big(t_i,\varphi(t_i)\big) - F\big(t_i,\varphi(t_i),0\big)\Big\| \\ &+ h \sum_{i=0}^{J(h)-1} \Big\| F\big(t_i,\varphi(t_i),0\big) - F\big(t_i,\varphi(t_i),h\big)\Big\|\,. \end{aligned}$$

Le premier terme du membre de droite tend vers 0 lorsque $h \to 0$, ainsi qu'on l'a déjà prouvé lors de l'étude de la convergence du schéma d'Euler. La fonction $(t,h) \mapsto F\big(t,\varphi(t),h\big)$ est continue, donc $h \mapsto F\big(t,\varphi(t),h\big)$ est continue, uniformément par rapport à t sur l'intervalle $[t_0, t_0+T]$. Par suite, le troisième terme du membre de droite tend vers 0 lorsque $h \to 0$.

Supposons que pour tous t et u, $F(t,u,0) = f(t,u)$. Alors le second terme du membre de droite de l'inégalité ci-dessus est nul, et par suite $\displaystyle\sum_{i=0}^{J(h)-1} \|\epsilon_i\|$ tend vers 0 lorsque $h \to 0$, ce qui exprime que le schéma est consistant avec l'équation.

Réciproquement, supposons le schéma consistant avec l'équation. En réordonnant l'égalité $(*)$, on obtient la majoration

$$\begin{aligned} h \sum_{i=0}^{J(h)-1} &\Big\| f\big(t_i,\varphi(t_i)\big) - F\big(t_i,\varphi(t_i),0\big)\Big\| \\ &\le \sum_{i=0}^{J(h)-1} \int_{t_i}^{t_{i+1}} \Big\| f\big(t,\varphi(t)\big) - f\big(t_i,\varphi(t_i)\big)\Big\|\, dt + \sum_{i=0}^{J(h)-1} \|\epsilon_i\| \\ &\quad + h \sum_{i=0}^{J(h)-1} \Big\| F\big(t_i,\varphi(t_i),0\big) - F\big(t_i,\varphi(t_i),h\big)\Big\|\,. \end{aligned}$$

Les trois termes du second membre tendent vers 0 lorsque $h \to 0$. Il en est donc de même du premier membre. Mais d'après la théorie de l'intégrale de Riemann, celui-ci tend, lorsque $h \to 0$, vers l'intégrale

$$\int_{t_0}^{t_0+T} \Big\| f\big(t,\varphi(t)\big) - F\big(t,\varphi(t),0\big)\Big\| \, dt \,.$$

Cette intégrale est donc nulle. La fonction intégrée étant continue et ≥ 0, est nécessairement identiquement nulle. On a donc

$$f\big(t,\varphi(t)\big) = F\big(t,\varphi(t),0\big) \,.$$

D'après le théorème de Cauchy-Lipschitz, pour tout (θ, x) appartenant au domaine de définition de f, il existe une solution φ telle que $\varphi(\theta) = x$. On a, en appliquant le résultat qui précède à cette solution,

$$f(\theta, x) = F(\theta, x, 0) \,,$$

ce qui est le résultat qu'on voulait démontrer. □

4.9. Théorème. — *Pour que le schéma à un pas défini par la fonction F soit stable, il suffit que cette fonction soit lipschitzienne par rapport à sa seconde variable.*

Preuve : Supposons F lipschitzienne par rapport à sa seconde variable. Il existe alors une constante $K \geq 0$ telle que, pour tous $t \in \mathbf{R}$, u et $v \in E$, $h \in \mathbf{R}^+$,

$$\big\|F(t,u,h) - F(t,v,h)\big\| \leq \|u - v\| \,.$$

Soient U_0 et V_0 deux éléments de E, et (ϵ_i), avec $0 \leq i \leq J(h) - 1$, une suite d'éléments de E. Posons, pour tout entier i vérifiant $0 \leq i \leq J(h) - 1$,

$$U_{i+1} = U_i + hF(t_i, U_i, h) \,, \qquad V_{i+1} = V_i + hF(t_i, V_i, h) + \epsilon_i \,.$$

On a

$$\begin{aligned} \|U_{i+1} - V_{i+1}\| &\leq \|U_i - V_i\| + h\big\|F(t_i, U_i, h) - F(t_i, V_i, h)\big\| + \|\epsilon_i\| \\ &\leq (1 + hK)\|U_i - V_i\| + \|\epsilon_i\| \,. \end{aligned}$$

En appliquant le lemme de Gronwall discret, on obtient

$$\begin{aligned} \|U_i - V_i\| &\leq e^{hKi}\|U_0 - V_0\| + \sum_{j=0}^{i-1} e^{hK(i-j-1)}\|\epsilon_j\| \\ &\leq e^{KT}\big(\|U_0 - V_0\| + \sum_{j=0}^{i-1} \|\epsilon_j\|\big) \,, \end{aligned}$$

ce qui prouve la stabilité du schéma. □

La notion de schéma d'ordre p (relativement à l'équation (1)) définie ci-dessous précise celle de schéma consistant avec cette équation.

4.10. Définition. — *Le schéma à un pas défini par la fonction F est dit d'ordre p (avec p réel > 0) relativement à l'équation (1) si, pour toute solution exacte φ de (1), il existe une constante $c > 0$ telle que*

$$\sum_{i=0}^{J(h)-1} \Big\| \varphi(t_{i+1}) - \varphi(t_i) - hF\big(t_i, \varphi(t_i), h\big) \Big\| \leq Ch^p \,.$$

4.11. Théorème. — *Si le schéma à un pas défini par la fonction F est stable et d'ordre p relativement à l'équation (1), il converge vers la solution exacte de l'équation au moins aussi vite que la suite (h^p) converge vers 0 lorsque $h \to 0$.*

Preuve : Le même calcul que celui fait pour prouver la convergence d'un schéma stable et consistant conduit à l'inégalité

$$\|\Phi_i - \varphi_i\| \leq M\big(\|\Phi_0 - \varphi_0\| + Ch^p\big) \,,$$

ce qui implique le résultat puisque $\Phi_0 = \varphi_0$. □

4.12. Théorème. — *Soit l'équation différentielle*

$$\varphi'(t) = f\big(t, \varphi(t)\big) \,. \tag{1}$$

On suppose f de classe C^p, avec $p \geq 1$. On considère un schéma à un pas défini par une fonction $F(t, x, h)$, continue par rapport à (t, x, h), p fois dérivable par rapport à h, et dont les dérivées (par rapport à h) jusqu'à l'ordre p sont toutes des fonctions continues de (t, x, h). Ce schéma est d'ordre p relativement à l'équation (1) si et seulement si, pour tout entier k vérifiant $0 \leq k \leq p-1$, on a

$$\frac{\partial^k F(t, x, h)}{\partial h^k}\Big|_{h=0} = \frac{f_k(t, x)}{k+1} \,,$$

où $f_k(t, x)$ est défini par

$$\begin{cases} f_0(t, x) = f(t, x) \,, \\ f_{k+1}(t, x) = \dfrac{\partial f_k(t, x)}{\partial t} + D_2 f_k(t, x)\big(f(t, x)\big) \,, \quad 0 \leq k \leq p-1 \,, \end{cases}$$

où $D_2 f_k(t, x)$ désigne la différentielle partielle de f_k par rapport à sa seconde variable x.

Preuve : Faisons une remarque préliminaire. Une solution exacte φ de l'équation (1) est de classe C^{p+1}. On a en effet

$$\varphi'(t) = f\big(t, \varphi(t)\,.$$

Comme φ est dérivable, elle est continue; f étant aussi continue, l'expression ci-dessus montre que φ' est continue, donc que φ est de classe C^1. Supposons φ de classe C^k (avec $1 \leq k \leq p$), (hypothèse de récurrence). L'expression ci-dessus montre que φ' est aussi de classe C^k, donc que φ est de classe C^{k+1}. On a ainsi prouvé par récurrence que φ est de classe C^{p+1}. On vérifie aisément que ses dérivées successives ont pour expression

$$\varphi^{(k)}(t) = f_{k-1}\big(t, \varphi(t)\big) \,.$$

Posons, pour tout entier i ($0 \leq i \leq J(h)-1$),

$$\epsilon_i = \varphi(t_{i+1}) - \varphi(t_i) - hF\big(t_i, \varphi(t_i), h\big)\,.$$

En appliquant la formule de Taylor avec reste intégral on obtient

$$\begin{aligned}\epsilon_i = \varphi(t_i) + h\varphi'(t_i) + \cdots + \frac{h^p}{p!}\varphi^{(p)}(t_i) + \int_{t_i}^{t_{i+1}} \frac{(t_{i+1}-t)^p}{p!}\varphi^{(p+1)}(t)\,dt \\ - \varphi(t_i) - h\bigg(F\big(t_i,\varphi(t_i),0\big) + h\frac{\partial F\big(t_i,\varphi(t_i,0\big)}{\partial h} + \cdots \\ + \frac{h^{p-1}}{(p-1)!}\frac{\partial^{(p-1)}F\big(t_i,\varphi(t_i),0\big)}{\partial h^{p-1}} + \int_0^h \frac{(h-s)^{p-1}}{(p-1)!}\frac{\partial^p F\big(t_i,\varphi(t_i),s\big)}{\partial h^p}\,ds\bigg)\,.\end{aligned}$$

En groupant les termes de même degré en h, on peut écrire

$$\begin{aligned}\epsilon_i = \sum_{k=1}^{p}\frac{\beta_i^k}{(k-1)!}h^k + \int_{t_i}^{t_{i+1}}\frac{(t_{i+1}-t)^p}{p!}\varphi^{(p+1)}(t)\,dt \\ - h\int_0^h \frac{(h-s)^{p-1}}{(p-1)!}\frac{\partial^p F\big(t_i,\varphi(t_i),s\big)}{\partial h^p}\,ds\,,\end{aligned}$$

où on a posé

$$\begin{aligned}\beta_i^1 &= \varphi'(t_i) - F\big(t_i,\varphi(t_i),0\big)\,,\\ \beta_i^k &= \frac{\varphi^{(k)}(t_i)}{k} - \frac{\partial^{k-1}F\big(t_i,\varphi(t_i),0\big)}{\partial h^{k-1}}\,,\quad 2\leq k\leq p\,.\end{aligned}$$

Si, pour tout entier k vérifiant $0 \leq k \leq p-1$, on a

$$\frac{\partial^k F(t,x,h)}{\partial h^k}\Big|_{h=0} = \frac{f_k(t,x)}{k+1}\,,$$

les β_i^k sont tous nuls, donc

$$\|\epsilon_i\| \leq \int_{t_i}^{t_{i+1}}\frac{(t_{i+1}-t)^p}{p!}\big\|\varphi^{(p+1)}(t)\big\|\,dt + h\int_0^h \frac{(h-s)^{p-1}}{(p-1)!}\left\|\frac{\partial^p F\big(t_i,\varphi(t_i),s\big)}{\partial h^p}\right\|\,ds\,.$$

Mais $\big\|\varphi^{(p+1)}(t)\big\|$ et $\left\|\dfrac{\partial^p F\big(t_i,\varphi(t_i,s)\big)}{\partial h^p}\right\|$ sont bornés lorsque $t\in[t_0,t_0+T]$, $s\in[0,h]$.

On en déduit une majotration de $\|\epsilon_i\|$ de la forme

$$|\epsilon_i\| \leq Mh^{p+1}\,,$$

où M est une constante ne dépendant pas de i. D'où

$$\sum_{i=0}^{J(h)-1}\|\epsilon_i\| \leq MJ(h)h^{p+1} \leq MTh^p\,,$$

puisque $hJ(h)\leq T$. Cela prouve que le schéma est d'ordre p relativement à l'équation (1).

Réciproquement, si le schéma est d'ordre p relativement à (1), un raisonnement par récurrence sur k analogue à celui fait pour montrer que si un schéma est consistant avec l'équation, la fonction F vérifie $F(t,u,0) = f(t,u)$, montre qu'on a nécessairement, pour tout entier k vérifiant $0 \leq k \leq p-1$,

$$\frac{\partial^k F(t,x,h)}{\partial h^k}\Big|_{h=0} = \frac{f_k(t,x)}{k+1}\,.$$

□

5. Les méthodes de Runge-Kutta

Les méthodes de Runge-Kutta sont parmi les plus utilisées pour la résolution approchée des équations différentielles. Elles font partie des schémas à un pas. Il en existe de tous les ordres. Nous allons décrire successivement les méthodes de Runge-Kutta d'ordre 2 et d'ordre 4.

5.1. La méthode de Runge-Kutta d'ordre 2. — Elle consiste à poser

$$\Phi_{i,1} = \Phi_i\,,$$

$$\Phi_{i,2} = \Phi_i + \frac{h}{2} f(t_i, \Phi_{i,1})\,,$$

$$\Phi_{i+1} = \Phi_i + h f(t_i + \frac{h}{2}, \Phi_{i,2})\,.$$

C'est une méthode à un pas, associée à la fonction

$$F(t,x,h) = f\left(t + \frac{h}{2}, x + \frac{h}{2} f(t,x)\right)\,.$$

Cette méthode est stable, consistante avec l'équation considérée et, lorsque f est de classe C^2, d'ordre 2 relativement à cette équation. On le vérifie aisément en appliquant les théorèmes du paragraphe précédent. Ainsi par exemple, en supposant f lipschitzienne de rapport K relativement à sa seconde variable,

$$\begin{aligned}
\big\|F(t,u,h) - F(t,v,h)\big\| &= \Big\|f\Big(t + \frac{h}{2}, u + \frac{h}{2} f(t,u)\Big) - f\Big(t + \frac{h}{2}, v + \frac{h}{2} f(t,v)\Big)\Big\| \\
&\leq K\Big\|u + \frac{h}{2} f(t,u) - v - \frac{h}{2} f(t,v)\Big\| \\
&\leq K\Big(\|u - v\| + \frac{h}{2}\big\|f(t,u) - f(t,v)\big\|\Big) \\
&\leq K\left(1 + \frac{hK}{2}\right)\|u - v\|\,,
\end{aligned}$$

ce qui prouve que F est lipschitzienne par rapport à sa seconde variable, donc que le schéma est stable. De même on a

$$F(t,x,0) = f(t,x)\,,$$

ce qui prouve que le schéma est consistant avec l'équation (1). En dérivant par rapport à h, on obtient

$$F'(t,x,h) = \frac{1}{2}\frac{\partial f}{\partial t}\left(t + \frac{h}{2}, x + \frac{h}{2} f(t,x)\right) + \frac{1}{2} D_2 f\left(t + \frac{h}{2}, x + \frac{h}{2} f(t,x)\right) f(t,x)\,.$$

On en déduit

$$F'(t,x,0) = \frac{1}{2}\left(\frac{\partial f(t,x)}{\partial t} + D_2 f(t,x)\big(f(t,x)\big)\right) = \frac{1}{2} f_1(t,x)\,,$$

ce qui prouve que le schéma est d'ordre 2.

5.2. La méthode de Runge-Kutta d'ordre 4. — C'est la plus utilisée en pratique. Elle consiste à poser

$$\begin{aligned}
\Phi_{i,1} &= \Phi_i\,,\\
\Phi_{i,2} &= \Phi_i + \frac{h}{2} f\Big(t_i + \frac{h}{2}, \Phi_{i,1}\Big)\,,\\
\Phi_{i,3} &= \Phi_i + \frac{h}{2} f\Big(t_i + \frac{h}{2}, \Phi_{i,2}\Big)\,,\\
\Phi_{i,4} &= \Phi_i + h f\Big(t_i + h, \Phi_{i,3}\Big)\,,\\
\Phi_{i+1} &= \Phi_i + \frac{h}{6}\bigg(f(t_i,\Phi_i) + 2f\Big(t_i + \frac{h}{2}, \Phi_{i,2}\Big) + 2f\Big(t_i + \frac{h}{2}, \Phi_{i,3}\Big)\\
&\qquad\qquad + f(t_i + h, \Phi_{i,4})\bigg)\,.
\end{aligned}$$

C'est encore une méthode à un pas. On laisse au lecteur le soin d'établir l'expression de la fonction F correspondante. On vérifie aisément que le schéma est stable et consistant avec l'équation. On vérifie aussi, au prix de calculs assez laborieux, que lorsque f est de classe C^4, le schéma est d'ordre 4.