ALGÈBRE ET

ANALYSE APPROFONDIES II

MT242

Année 1998-1999

Chapitre 0. Introduction générale

Dans cette introduction nous allons commenter les principales notions contenues dans le cours du second semestre, leurs relations entre elles et avec le cours du premier semestre.

La modélisation de nombreux problèmes concrets conduit souvent à la recherche du minimum (ou du maximum) d'une quantité f dépendant d'une ou plusieurs variables, par exemple $f(x)$, avec x réel, ou bien $f(x_1, x_2)$, avec x_1, x_2 réels. Si f ne dépend que d'une seule variable, on peut attaquer le problème par les méthodes déjà vues au lycée : rechercher les zéros de la dérivée, et chercher ensuite si le point trouvé (ou l'un des points trouvés) est effectivement un minimum. On a vu en première année comment étudier cette deuxième question avec un développement limité (en abrégé : DL) d'ordre 2 au voisinage du point candidat ; si g est de classe C^2 au voisinage de a et si $g'(a) = 0$ (condition nécessaire pour avoir un minimum au point a) on pourra écrire par exemple avec Taylor-Young la relation

$$g(a+h) = g(a) + g''(a)\frac{h^2}{2} + o(h^2)$$

qui montre que la condition $g''(a) \geq 0$ est nécessaire pour que g admette un minimum au point a, et que la condition $g''(a) > 0$ est suffisante pour que a soit un minimum local. L'étude en plusieurs variables fait apparaître la notion de *différentielle* (pour remplacer la dérivée en une variable), qui sera introduite au chapitre 3. Par exemple, un DL à l'ordre 1 de la fonction $f(x_1, x_2)$ de deux variables réelles x_1, x_2 au voisinage du point origine $(0, 0)$ sera une expression de la forme

$$f(x_1, x_2) = f(0, 0) + ax_1 + bx_2 + o(x),$$

et la différentielle (de la fonction f au point $(0, 0)$) sera la *fonction linéaire* ℓ donnée par la partie linéaire du développement limité, c'est à dire qu'ici la fonction ℓ sera définie par $\ell(x_1, x_2) = ax_1 + bx_2$. Par une extension assez naturelle du cas de la dimension 1, la condition nécessaire pour que f admette un minimum au point $(0, 0)$ s'exprimera maintenant par la nullité de l'application différentielle ℓ de f au point $(0, 0)$, en l'occurence elle s'exprimera par les conditions $a = b = 0$. Pour aller plus loin, on écrira un DL d'ordre 2 qui sera de la forme

$$f(x_1, x_2) = f(0, 0) + ax_1 + bx_2 + cx_1^2 + 2dx_1x_2 + ex_2^2 + o(\|x\|^2),$$

qui se réduira à

$$f(x_1, x_2) = f(0, 0) + cx_1^2 + 2dx_1x_2 + ex_2^2 + o(\|x\|^2)$$

quand la différentielle s'annule. Pour que f admette un minimum en 0, il est alors nécessaire que $cx_1^2 + 2dx_1x_2 + ex_2^2$ soit toujours ≥ 0. Une telle expression est une *forme quadratique*, dont l'étude fait l'objet du chapitre 1. On verra notamment dans le chapitre 1 la *méthode de décomposition de Gauss*, une méthode simple qui donne en particulier le signe d'une forme quadratique.

Pour allonger la liste des outils de manipulation des fonctions de plusieurs variables, la théorie de l'intégrale double complète le calcul différentiel. Le théorème fondamental dans cette direction est le théorème 4.3.3, qui ramène le calcul d'une intégrale double à deux intégrales simples successives.

Un autre outil de modélisation très important est la notion d'équation différentielle. Chacun sait que de nombreux phénomènes physiques sont décrits au moyen d'équation différentielles, par exemple la chute des corps, le mouvement des planètes, l'oscillation d'un ressort... Ceci sera vu au chapitre 5. Le théorème d'existence de solutions de Cauchy-Lipschitz (théorème 5.3.3) est un théorème important, dont la preuve est une belle illustration de l'utilisation des théorèmes de continuité et de dérivabilité sur les séries normalement convergentes de fonctions, théorèmes vus au premier semestre. Le cas des systèmes différentiels linéaires (à coefficients constants), de la forme

$$y' = \mathrm{A}y$$

utilise la série exponentielle de matrice $\mathrm{e}^{t\mathrm{A}}$. Une question naturelle qui se pose pour un phénomène $y(t)$ dépendant du temps t et régi par une telle équation différentielle linéaire est de connaître le comportement de la solution lorsque le temps t tend vers $+\infty$; cette étude utilise toute la théorie de la réduction des matrices, vue au premier semestre.

Le poly se termine par un index terminologique et un index des notations.

Chapitre 1. Formes quadratiques

Dans ce chapitre, on désignera par E un espace vectoriel sur un corps $\mathbb{K}$; pour l'essentiel, le lecteur pourra considérer que $\mathbb{K}$ sera égal à $\mathbb{R}$ ou à $\mathbb{C}$; l'amateur de mathématiques pourra néanmoins observer avec intérêt que la plupart des résultats sont valables pour un corps $\mathbb{K}$ (presque) quelconque ; on se risquera une fois ou deux à des allusions à des cas plus exotiques, en particulier au cas du corps à deux éléments $\mathbb{K} = \mathbb{Z}/2\mathbb{Z}$, pour lequel justement certaines des méthodes qui seront développées pour les formes quadratiques *ne marchent pas*.

1.1. Formes linéaires. Espace dual

Une *forme linéaire* sur E est une application $\mathbb{K}$-linéaire de E dans $\mathbb{K}$. Etant données deux formes linéaires f et g sur E, on définit leur somme $f + g$ par l'opération usuelle de somme de deux fonctions,

$$\forall x \in \mathrm{E}, \quad (f+g)(x) = f(x) + g(x).$$

Il est facile de vérifier que $f+g$ est encore une forme linéaire. Pour tout $\lambda \in \mathbb{K}$, on définit aussi la fonction λf par la formule $(\lambda f)(x) = \lambda f(x)$. Cette fonction λf est linéaire, et muni de ces deux opérations, l'ensemble des formes linéaires sur E est un espace vectoriel sur $\mathbb{K}$. C'est *l'espace dual* de E, noté E^*.

L'image $f(\mathrm{E})$ d'une forme linéaire f est un $\mathbb{K}$-sous-espace vectoriel de $\mathbb{K}$, donc elle est égale à $\{0\}$ ou à $\mathbb{K}$. Une forme linéaire f sur E est non nulle si et seulement s'il existe un vecteur $x \in \mathrm{E}$ tel que $f(x) = 1$. Si E est de dimension n et si f est une forme linéaire non nulle sur E, le noyau $\ker f$ est un sous-espace vectoriel de dimension $n-1$ de E. On appelle *hyperplan* (vectoriel) de E tout sous-espace vectoriel F de E qui est de *codimension* 1 c'est à dire tel que $\mathrm{F} \neq \mathrm{E}$ et que $\mathrm{E} = \mathrm{F} + \mathbb{K}\,x$ pour tout vecteur $x \notin \mathrm{F}$ (autrement dit, quand on a F, il ne manque plus qu'une dimension pour arriver à E tout entier). Si E est de dimension finie $n > 0$, un sous-espace F de E est un hyperplan de E si et seulement si $\dim \mathrm{F} = n - 1$. Un *hyperplan affine* est un sous-ensemble obtenu en translatant un hyperplan vectoriel. Le noyau d'une forme linéaire f sur E non nulle est donc un hyperplan de E, et les ensembles de la forme $\mathrm{H}_c = \{x \in \mathrm{E} : f(x) = c\}$ sont des hyperplans affines. Dans le cas réel (lorsque $\mathbb{K} = \mathbb{R}$), cet hyperplan affine H_c sépare l'espace E en deux *demi-espaces*, à savoir $\{x \in \mathrm{E} : f(x) \leq c\}$ et $\{x \in \mathrm{E} : f(x) \geq c\}$.

Exercice facile. Montrer sans supposer la dimension de E finie que le noyau d'une forme linéaire non nulle est de codimension 1.

Base duale d'une base de E

Supposons que l'espace vectoriel E soit de dimension finie $n > 0$, et supposons donnée une base $\mathbf{e} = (e_1, \ldots, e_n)$ de l'espace E ; on définit un système $\mathbf{e}^*$ de formes linéaires sur E à partir de cette base, de la façon suivante : pour chaque $i = 1, \ldots, n$, on désigne par e_i^* la fonction scalaire définie sur E, qui associe à chaque vecteur x de E sa $i$ème coordonnée dans la base $\mathbf{e}$, et on pose $\mathbf{e}^* = (e_1^*, \ldots, e_n^*)$. On peut définir toutes ces fonctions (e_i^*) par une formule (implicite) unique,

$$(*) \qquad \forall x \in \mathrm{E}, \quad x = \sum_{i=1}^{n} e_i^*(x)\, e_i.$$

On remarque que

$$e_i^*(e_j) = \delta_{i,j},$$

où $\delta_{i,j}$, appelé *symbole de Kronecker*, est égal à 1 si $i = j$ et à 0 sinon (en d'autres termes, la matrice des coefficients $(\delta_{i,j})$ est la matrice unité I_n).

Proposition 1.1.1. *Le système* $\mathbf{e}^*$ *est une base de l'espace dual* E^*. *En conséquence, lorsque* E *est de dimension finie, on a* $\dim \mathrm{E}^* = \dim \mathrm{E}$.

On dit que $\mathbf{e}^*$ est la *base duale* de la base $\mathbf{e}$ de E.

Démonstration. Soit x^* une forme linéaire sur E ; en appliquant x^* à la décomposition d'un vecteur $x \in \mathrm{E}$ quelconque donnée par la formule $(*)$, on obtient

$$\forall x \in \mathrm{E}, \quad x^*(x) = \sum_{i=1}^n e_i^*(x) x^*(e_i) = \Bigl(\sum_{i=1}^n x^*(e_i) e_i^*\Bigr)(x).$$

En termes de fonctions sur E, ceci signifie que $x^* = \sum_{i=1}^n x^*(e_i)\, e_i^*$, et montre que le système de formes linéaires $\mathbf{e}^*$ est générateur pour l'espace vectoriel dual E^*. Si on écrit une combinaison linéaire $x^* = \sum_{i=1}^n c_i e_i^*$, on trouve, en appliquant la forme linéaire x^* au vecteur e_j, la relation $c_j = x^*(e_j)$ qui montre que les coefficients (c_j) de la combinaison linéaire sont uniquement déterminés, donc $\mathbf{e}^*$ est une base du dual E^*.

Calcul des coordonnées d'une forme linéaire dans une base duale

Si $\mathbf{e}^*$ est la base duale d'une base $\mathbf{e} = (e_1, \ldots, e_n)$ de E, et si $x^* = c_1 e_1^* + \cdots + c_n e_n^*$, on vient de dire qu'on trouve les coefficients (c_i) au moyen de la formule

$$c_i = x^*(e_i), \quad i = 1, \ldots, n\,;$$

cette formule évidente sera utilisée à plusieurs reprises dans ce chapitre.

Remarque. Etant donnés un système de vecteurs $\mathbf{x} = (x_1, \ldots, x_n)$ dans E (espace vectoriel de dimension n) et un système de formes linéaires $(y_1^*, \ldots, y_n^*)$ dans E^*, les relations

$$\forall i, j = 1, \ldots, n, \quad y_i^*(x_j) = \delta_{i,j}$$

impliquent que le système $\mathbf{x} = (x_1, \ldots, x_n)$ est une base de E et que $(y_1^*, \ldots, y_n^*)$ est la base duale de la base $\mathbf{x}$ de E.

Proposition 1.1.2. *Soit f une forme linéaire non nulle sur* E, *et soit* $\mathrm{H} = \ker f$ *son noyau ; si g est une forme linéaire sur* E *qui est nulle en tout point de* H, *il existe un scalaire $\lambda \in \mathbb{K}$ tel que $g = \lambda f$ (autrement dit, g est proportionnelle à f).*

Démonstration. Puisque f est non nulle, on peut trouver un vecteur $x_0 \in \mathrm{E}$ tel que $f(x_0) = 1$. Supposons que g soit une forme linéaire nulle sur le noyau H de f, et posons $\lambda = g(x_0)$. Soit y un vecteur quelconque de E ; écrivons

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

Posons $z = y - f(y)\, x_0$. On a $f(z) = f(y) - f(y) f(x_0) = 0$, donc $z \in \mathrm{H}$ et $g(z) = 0$ par hypothèse. On a donc $g(y) = g(z) + f(y) g(x_0) = f(y) g(x_0) = \lambda f(y)$, ce qui montre bien que $g = \lambda f$.

On peut résumer la démonstration ainsi : la forme linéaire $g - \lambda f$ est nulle sur E, parce qu'elle est nulle sur H et sur $\mathbb{K}x_0$, et que $\mathrm{E} = \mathrm{H} \oplus \mathbb{K}x_0$; elle est nulle sur H puisque pour tout $y \in \mathrm{H}$, on a $(g - \lambda f)(y) = g(y) - \lambda g(y) = 0 - 0 = 0$, et nulle sur $\mathbb{K}x_0$ puisque $(g - \lambda f)(x_0) = \lambda - \lambda = 0$.

Corollaire 1.1.1. *Deux formes linéaires qui ont le même noyau sont proportionnelles.*

Proposition 1.1.3. *Soit* E *un espace vectoriel de dimension finie ; pour tout vecteur non nul* $x \in \mathrm{E}$*, il existe une forme linéaire* $x^* \in \mathrm{E}^*$ *telle que* $x^*(x) \neq 0$*.*

Démonstration. Puisque x est non nul, on peut trouver d'après le théorème de la base incomplète une base $\mathbf{e} = (e_1, \ldots, e_n)$ de E telle que $e_1 = x$. La forme linéaire e_1^* de la base duale $\mathbf{e}^*$ convient, puisque $e_1^*(x) = e_1^*(e_1) = 1 \neq 0$.

Proposition 1.1.4. *Soient* $f_1, \ldots, f_k$ *des formes linéaires indépendantes sur un espace vectoriel* E *de dimension finie ; posons*

$$\mathrm{M} = \{x \in \mathrm{E} : \forall i = 1, \ldots, k,\ f_i(x) = 0\}$$

(c'est l'intersection des noyaux $\ker f_j$ *des formes linéaires considérées). La dimension du sous-espace vectoriel* M *est égale à* $\dim \mathrm{E} - k$*.*

Démonstration. Posons $n = \dim \mathrm{E}$, et complétons le système $(f_1, \ldots, f_k)$ en une base $(f_1, \ldots, f_n)$ du dual E^*. Considérons l'application linéaire $u : \mathrm{E} \to \mathbb{K}^n$ définie par

$$u(x) = (f_1(x), \ldots, f_n(x)) \in \mathbb{K}^n.$$

Si $x \in \ker u$, on a $f_i(x) = 0$ pour tout $i = 1, \ldots, n$, donc $f(x) = 0$ pour toute forme linéaire $f \in \mathrm{E}^*$ (micro-exercice), donc $x = 0_\mathrm{E}$ d'après la proposition précédente 1.1.3. Il en résulte que u est injective, donc u est un isomorphisme puisque $\dim \mathrm{E}^* = \dim \mathrm{E} = n = \dim \mathbb{K}^n$. On voit alors que M est l'image par l'isomorphisme inverse u^{-1} du sous-espace M' de $\mathbb{K}^n$ de dimension $n - k$ défini par

$$\mathrm{M}' = \{(t_1, \ldots, t_n) \in \mathbb{K}^n : t_1 = t_2 = \cdots = t_k = 0\}$$

donc $\dim \mathrm{M} = \dim \mathrm{M}' = n - k$.

Remarque. Si on ne suppose pas que les formes $(f_1, \ldots, f_k)$ sont indépendantes, on verra que $\dim \mathrm{M} = \dim \mathrm{E} - \dim[f_1, \ldots, f_k]$ (où la notation $[f_1, \ldots, f_k]$ désigne le sous-espace vectoriel de E^* engendré par $f_1, \ldots, f_k$). En particulier, on a toujours $\dim \mathrm{M} \geq \dim \mathrm{E} - k$.

Corollaire 1.1.2. *Si* $(f_1, \ldots, f_n)$ *est une base de* E^**, il existe une base* $x_1, \ldots, x_n$ *de* E *telle que* $f_i(x_j) = \delta_{i,j}$ *pour tous* $i, j = 1, \ldots, n$ *(toute base de* E^* *est donc la base duale d'une base de* E*).*

Démonstration. Comme dans la démonstration de la proposition précédente, on introduit l'isomorphisme u de E sur $\mathbb{K}^n$ défini par $u(x) = (f_1(x), \ldots, f_n(x)) \in \mathbb{K}^n$ pour tout $x \in \mathrm{E}$. Pour trouver les vecteurs (x_j) il suffit de considérer les images par l'inverse u^{-1} des vecteurs de la base canonique de $\mathbb{K}^n$.

Application linéaire transposée

Soit $a : \mathrm{E} \to \mathrm{F}$ une application linéaire ; on définit une application ${}^t a : \mathrm{F}^* \to \mathrm{E}^*$ par la formule

$$\forall y^* \in \mathrm{F}^*,\quad {}^t a(y^*) = y^* \circ a \in \mathrm{E}^*.$$

On vérifie facilement que cette application ${}^t a$ est linéaire de F^* dans E^*. Si on se donne une deuxième application linéaire $b : \mathrm{F} \to \mathrm{G}$, on voit que

$${}^t(b \circ a) = {}^t a \circ {}^t b$$

(bien noter l'interversion de a et b). La transposée de l'identité de E est l'identité de E^*. Il résulte des deux propriétés précédentes que la transposée d'un isomorphisme est un isomorphisme, et que l'on a dans ce cas $({}^t a)^{-1} = {}^t(a^{-1})$.

Supposons que E et F soient de dimensions finies m et n, que $\mathbf{e}$ soit une base de E et $\mathbf{f}$ une base de F. Considérons la matrice (en général rectangulaire) $\mathrm{A} = \mathrm{mat}(a, \mathbf{e}, \mathbf{f})$ de l'application linéaire a par rapport à ces deux bases. La matrice $\mathrm{B} = \mathrm{mat}({}^t a, \mathbf{f}^*, \mathbf{e}^*)$ est alors égale à la matrice transposée de A. En effet, la colonne i de la matrice B contient les coordonnées de la forme linéaire ${}^t a(f_i^*)$ dans la base $\mathbf{e}^*$. On a vu que la coordonnée j d'une forme linéaire x^* dans cette base $\mathbf{e}^*$ est égale à $x^*(e_j)$, donc en appliquant ceci à $x^* = {}^t a(f_i^*)$ on obtient

$$\mathrm{B}_{j,i} = {}^t a(f_i^*)(e_j) = f_i^*(a(e_j)) = \mathrm{A}_{i,j}.$$

Transposition d'un changement de base : si V est la matrice de changement de base de $\mathbf{e}$ vers $\mathbf{f}$, la matrice transposée ${}^t\mathrm{V}$ est la matrice de changement de $\mathbf{f}^*$ vers $\mathbf{e}^*$ (attention au changement de l'ordre des bases !).

Bidual d'un espace vectoriel

A tout vecteur $x \in \mathrm{E}$, on peut associer une fonction scalaire $j_\mathrm{E}(x)$, définie sur l'espace dual E^* au moyen de la formule

$$\forall x^* \in \mathrm{E}^*, \quad (j_\mathrm{E}(x))(x^*) = x^*(x) \in \mathbb{K}.$$

Il est facile de vérifier que $j_\mathrm{E}(x)$ est une fonction linéaire de E^* dans $\mathbb{K}$, c'est à dire un élément du dual $(\mathrm{E}^*)^*$ de E^*, qu'on appelle le *bidual* de E, et que l'on note E^{**}. De plus, l'application $j_\mathrm{E} : x \to j_\mathrm{E}(x)$ est linéaire de E dans E^{**}.

Théorème 1.1.1. *Si* E *est de dimension finie, l'application* j_E *est un isomorphisme de* E *sur son bidual* E^{**}.

Le caractère “naturel” de la définition justifie que l'on appelle j_E l'isomorphisme *canonique* de E sur E^{**} (la définition de j_E ne dépend que de E et de la définition du dual E^*, et ne dépend d'aucun choix auxiliaire).

Démonstration. On sait déjà que $\dim \mathrm{E}^{**} = \dim \mathrm{E}^* = \dim \mathrm{E}$ d'après la proposition 1.1.1. Pour savoir que j_E est un isomorphisme, il suffit donc de vérifier que j_E est injective, c'est à dire de voir que pour tout vecteur $x \in \mathrm{E}$ non nul, l'image $j_\mathrm{E}(x)$ est non nulle. Mais d'après la proposition 1.1.3, si $x \in \mathrm{E}$ est non nul il existe une forme linéaire x_0^* telle que $x_0^*(x) \neq 0$, donc $(j_\mathrm{E}(x))(x_0^*) = x_0^*(x) \neq 0$, ce qui montre que la fonction $j_\mathrm{E}(x)$ n'est pas identiquement nulle sur E^*, ce qui signifie que $j_\mathrm{E}(x) \neq 0_{\mathrm{E}^*}$.

Remarque. Retrouvons le fait que toute base $\mathbf{b} = (x_1^*, \ldots, x_n^*)$ de E^* est la base duale d'une base $\mathbf{e}$ de E. Considérons la base duale $(x_1^{**}, \ldots, x_n^{**})$ de la base $\mathbf{b}$; c'est une base du bidual E^{**}. Pour chaque $i = 1, \ldots, n$ il existe un vecteur $x_i \in \mathrm{E}$ tel que $x_i^{**} = j_\mathrm{E}(x_i)$. On a alors pour tous $i, j = 1, \ldots, n$

$$x_i^*(x_j) = x_j^{**}(x_i^*) = \delta_{i,j},$$

donc la base $\mathbf{b}$ est la base duale de la base $(x_1, \ldots, x_n)$ de E.

1.2. Formes bilinéaires

Une *application bilinéaire* est une application f définie sur le produit $\mathrm{E}\times\mathrm{F}$ de deux espaces vectoriels, à valeurs dans un troisième espace vectoriel G, et telle que pour tous $x_0 \in \mathrm{E}$ et $y_0 \in \mathrm{F}$ fixés, les applications $x \to f(x, y_0)$ et $y \to f(x_0, y)$ soient linéaires, respectivement de E dans G et de F dans G. On va surtout étudier le cas des *formes bilinéaires*, c'est à dire le cas où $\mathrm{G} = \mathbb{K}$.

Définition 1.2.1. Soient E et F deux espaces vectoriels sur le même corps $\mathbb{K}$; on dit qu'une application φ de $\mathrm{E}\times\mathrm{F}$ dans $\mathbb{K}$ est une *forme bilinéaire* sur $\mathrm{E}\times\mathrm{F}$ si

– pour tout $x_0 \in \mathrm{E}$ fixé, l'application $y \to \varphi(x_0, y)$ est linéaire de F dans $\mathbb{K}$,
– pour tout $y_0 \in \mathrm{F}$ fixé, l'application $x \to \varphi(x, y_0)$ est linéaire de E dans $\mathbb{K}$.

Exemples.

1. L'application φ définie par $\varphi(x, y) = xy$ est une forme bilinéaire sur $\mathbb{R}\times\mathbb{R}$. Plus généralement, si ℓ_1 est une forme linéaire sur E et ℓ_2 une forme linéaire sur F, l'application φ définie par $\varphi(x, y) = \ell_1(x)\ell_2(y)$ est une forme bilinéaire sur $\mathrm{E}\times\mathrm{F}$.

2. Si E désigne l'espace des fonctions réelles continues sur $[0, 1]$, on peut définir sur $\mathrm{E}\times\mathrm{E}$ la forme bilinéaire φ

$$\varphi(f, g) = \int_0^1 f(t)g(t)\, dt.$$

3. Sur $\mathbb{R}^n$, on a la forme bilinéaire usuelle donnée par le produit scalaire,

$$\varphi(x, y) = x \,.\, y = x_1y_1 + \cdots + x_ny_n.$$

4. Si $v = (x, y, z, t)$ et $v' = (x', y', z', t')$ sont deux éléments de $\mathbb{R}^4$, posons

$$\varphi(v, v') = xx' + yy' + zz' - tt'.$$

On définit ainsi une forme bilinéaire sur $\mathbb{R}^4$, qu'on qualifie d'*hyperbolique*. Cette forme bilinéaire met en évidence des phénomènes intéressants que nous étudierons plus loin. Elle joue un rôle dans la théorie de la relativité restreinte.

5. Forme bilinéaire canonique sur $\mathrm{E}^*\times\mathrm{E}$: on pose pour tous $x^* \in \mathrm{E}^*$ et $x \in \mathrm{E}$

$$\varphi(x^*, x) = x^*(x).$$

6. Si φ est une forme bilinéaire sur $\mathrm{E}\times\mathrm{F}$, l'application $\widetilde{\varphi}$ définie sur $\mathrm{F}\times\mathrm{E}$ par $\widetilde{\varphi}(y, x) = \varphi(x, y)$ est bilinéaire sur $\mathrm{F}\times\mathrm{E}$.

7. Si u est une application linéaire de F dans E^*, on peut définir une forme bilinéaire sur $\mathrm{E}\times\mathrm{F}$ en posant pour tout couple $(x, y) \in \mathrm{E}\times\mathrm{F}$

$$\varphi(x, y) = u(y)(x).$$

Remarque. Application linéaire $\mathrm{D}_\varphi : \mathrm{F} \to \mathrm{E}^*$ associée à φ (à droite) : l'exemple 7 ci-dessus est tout à fait général. Supposons en effet donnée une forme bilinéaire φ sur $\mathrm{E}\times\mathrm{F}$. A tout vecteur $y \in \mathrm{F}$ fixé, associons une forme linéaire $f(y) \in \mathrm{E}^*$, définie sur E par la formule

$$f(y)(x) = \varphi(x, y),$$

c'est à dire que $f(y)$ est l'application linéaire $x \to \varphi(x, y)$ de E dans $\mathbb{K}$. L'application $y \to f(y)$ est alors linéaire de F dans E^*. On posera $f = \mathrm{D}_\varphi$. Sa relation de définition est donc

$$\forall x \in \mathrm{E}, \forall y \in \mathrm{F}, \quad (\mathrm{D}_\varphi(y))(x) = \varphi(x, y).$$

La forme bilinéaire définie à partir de D_φ par la méthode de l'exemple 7 est égale à φ.

En d'autres termes, l'espace vectoriel $B(E, F)$ des formes bilinéaires sur $E \times F$ est isomorphe à l'espace $\mathcal{L}(F, E^*)$ des applications linéaires de F dans E^*. Si on échange le rôle des variables, on voit que $B(E, F)$ est aussi isomorphe à $B(F, E)$, donc à $\mathcal{L}(E, F^*)$.

Exemple. Pour la forme bilinéaire canonique sur $E^* \times E$ de l'exemple 5, l'application D_φ de E dans $(E^*)^* = E^{**}$ est l'application j_E introduite à propos du bidual. L'application D_φ est donc un isomorphisme dans cet exemple.

Cas de la dimension finie. Matrice d'une forme bilinéaire par rapport à deux bases

On suppose que E et F sont de dimensions finies m et n, que $\mathbf{e}$ est une base de E et $\mathbf{f}$ une base de F. Une forme bilinéaire φ sur $E \times F$ est complètement connue si on connaît toutes les valeurs des $\varphi(e_i, f_j)$, pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$. En effet, si $x = \sum_{i=1}^m a_i e_i \in E$ et $y = \sum_{j=1}^n b_j f_j \in F$, on obtiendra en développant par bilinéarité

$$\varphi(x, y) = \sum_{i,j} \varphi(e_i, f_j)\, a_i b_j.$$

On introduit la matrice Φ de la forme bilinéaire par rapport aux bases $\mathbf{e}$ et $\mathbf{f}$, en posant

$$\Phi_{i,j} = \varphi(e_i, f_j).$$

On notera $\Phi = \mathrm{Mat}(\varphi, \mathbf{e}, \mathbf{f})$. Si X est le vecteur colonne des coordonnées de x dans la base $\mathbf{e}$ et Y celui des coordonnées de y dans la base $\mathbf{f}$, on a

$$\varphi(x, y) = {}^t\mathrm{X}\, \Phi\, \mathrm{Y}.$$

La matrice Φ est aussi la matrice de l'application linéaire $D_\varphi : F \to E^*$ par rapport à la base $\mathbf{f}$ de F et à la base duale $\mathbf{e}^*$. En effet, si $B = \mathrm{mat}(D_\varphi, \mathbf{f}, \mathbf{e}^*)$, l'élément $B_{i,j}$ est la $i$ème coordonnée dans la base duale $\mathbf{e}^*$ du vecteur image $D_\varphi(f_j)$, c'est à dire

$$B_{i,j} = (D_\varphi(f_j))(e_i) = \varphi(e_i, f_j) = \Phi_{i,j}.$$

Il est évident que la matrice de $\widetilde{\varphi}$ (exemple 6) par rapport aux bases $\mathbf{f}$ et $\mathbf{e}$ est la transposée de Φ,

$$\mathrm{Mat}(\widetilde{\varphi}, \mathbf{f}, \mathbf{e}) = {}^t\,\mathrm{Mat}(\varphi, \mathbf{e}, \mathbf{f}),$$

puisque $\widetilde{\varphi}(f_i, e_j) = \varphi(e_j, f_i) = \Phi_{j,i}$.

Changement de base pour une forme bilinéaire

Supposons que $X = PX'$ et $Y = QY'$ expriment des changements de base dans E et dans F respectivement. On aura alors

$$\varphi(x, y) = {}^t(\mathrm{PX}')\, \Phi\, \mathrm{QY}' = {}^t\mathrm{X}'\, {}^t\mathrm{P}\, \Phi\, \mathrm{QY}' = {}^t\mathrm{X}'\, ({}^t\mathrm{P}\, \Phi\, \mathrm{Q})\mathrm{Y}'$$

ce qui montre que la matrice Φ' dans les nouvelles bases est donnée par

$$\Phi' = {}^t\mathrm{P}\, \Phi\, \mathrm{Q}.$$

Dans le cas où $E = F$, on choisit le plus souvent de prendre la même base pour les deux côtés. Le changement de base s'exprime alors par $\Phi' = {}^t\mathrm{P}\, \Phi\, \mathrm{P}$.

Orthogonalité pour une forme bilinéaire

Soient A une partie de E et φ une forme bilinéaire sur $\mathrm{E} \times \mathrm{F}$; on pose

$$\mathrm{A}^{\perp d} = \{y \in \mathrm{F} : \forall x \in \mathrm{A}, \quad \varphi(x, y) = 0\}.$$

L'ensemble $\mathrm{A}^{\perp d}$ est un sous-espace vectoriel de F, même si A ne l'est pas. On convient que l'orthogonal de $\mathrm{A} = \emptyset$ est F tout entier. On définit de la même façon l'orthogonal d'une partie B de F : c'est le sous-espace vectoriel $\mathrm{B}^{\perp g}$ de E formé de tous les vecteurs $x \in \mathrm{E}$ tels que $\varphi(x, y) = 0$ pour tout $y \in \mathrm{B}$.

Si $\mathrm{E} = \mathrm{F}$ et si φ n'est pas symétrique, il faut faire attention à distinguer l'orthogonalité à gauche et à droite ; c'est pour cela que nous avons introduit ces notations un peu lourdes (au lieu de noter $\mathrm{A}^{\perp}$ tout simplement). L'orthogonal de A est égal à l'orthogonal de Vect(A). En particulier, si E_1 est un sous-espace de dimension finie de E muni d'une base $(e_1, \ldots, e_k)$, on voit facilement que l'orthogonal $\mathrm{E}_1^{\perp d}$ est égal à

$$\mathrm{E}_1^{\perp d} = \{y \in \mathrm{F} : \varphi(e_i, y) = 0,\ i = 1, \ldots, k\}.$$

Exemple. L'orthogonal $\mathrm{E}^{\perp d}$ de $\mathrm{A} = \mathrm{E}$ est le sous-espace vectoriel de F formé de tous les vecteurs $y \in \mathrm{F}$ tels que l'application $x \to \varphi(x, y)$ soit identiquement nulle sur E : c'est donc le noyau de l'application linéaire D_φ.

On peut calculer la dimension de l'orthogonal d'un sous-espace vectoriel E_1 de E, moyennant une hypothèse simple.

Proposition 1.2.1. *On suppose que φ est une forme bilinéaire sur $\mathrm{E} \times \mathrm{F}$, où E et F sont de dimension finie. Pour tout sous-espace vectoriel E_1 de E, tel qu'aucun vecteur non nul $x \in \mathrm{E}_1$ ne puisse être orthogonal à F tout entier, (c'est à dire que $\mathrm{E}_1 \cap \mathrm{F}^{\perp g} = \{0_\mathrm{E}\}$) on a*

$$\dim \mathrm{E}_1^{\perp d} = \dim \mathrm{F} - \dim \mathrm{E}_1.$$

Démonstration. Supposons que E_1 soit de dimension k, muni d'une base $(e_1, \ldots, e_k)$. On a dit que

$$\mathrm{E}_1^{\perp d} = \{y \in \mathrm{F} : \varphi(e_i, y) = 0,\ i = 1, \ldots, k\}.$$

Pour chaque $i = 1, \ldots, k$, soit ℓ_i la forme linéaire sur F définie par $\ell_i(y) = \varphi(e_i, y)$; on a aussi $\mathrm{E}_1^{\perp d} = \{y \in \mathrm{F} : \ell_i(y) = 0,\ i = 1, \ldots, k\}$. Si nous savons que $\ell_1, \ldots, \ell_k$ sont indépendantes, on en déduira que $\dim \mathrm{E}_1^{\perp d} = \dim \mathrm{F} - k$ d'après la proposition 1.1.4. Supposons donc que $\sum_{i=1}^k c_i \ell_i = 0$, et montrons que $c_1 = \cdots = c_k = 0$. L'équation précédente signifie que

$$0 = \Bigl(\sum_{i=1}^k c_i \ell_i\Bigr)(y) = \sum_{i=1}^k c_i \varphi(e_i, y) = \varphi\Bigl(\sum_{i=1}^k c_i e_i, y\Bigr)$$

pour tout $y \in \mathrm{F}$. Le vecteur $x = \sum_{i=1}^k c_i e_i$ est donc un vecteur de E_1 tel que $\varphi(x, y) = 0$ pour tout vecteur y de F, c'est à dire que $x \in \mathrm{F}^{\perp g}$. D'après notre hypothèse, il en résulte que $x = 0_\mathrm{E}$, donc $c_1 = \cdots = c_k = 0$ puisque $\mathbf{e}$ est une base de E_1.

Exercice. En général, $\dim \mathrm{E}_1^{\perp d} + \dim \mathrm{E}_1 = \dim \mathrm{F} + \dim(\mathrm{E}_1 \cap \mathrm{F}^{\perp g})$.

On suppose dans ce paragraphe que les espaces E et F sont de même dimension finie, $\dim \mathrm{E} = \dim \mathrm{F} = n > 0$, et on suppose donnée une forme bilinéaire φ sur $\mathrm{E} \times \mathrm{F}$ telle que pour tout $y \neq 0_{\mathrm{F}}$, il existe un vecteur $x \in \mathrm{E}$ tel que $\varphi(x,y) \neq 0$. Cette propriété signifie exactement que $\mathrm{E}^{\perp d} = \{0_{\mathrm{F}}\}$, ce qui équivaut encore à dire que l'application D_φ est injective. Puisque $\dim \mathrm{F} = \dim \mathrm{E} = \dim \mathrm{E}^*$, cela équivaut donc à dire que D_φ est un isomorphisme de F sur E^*. La forme bilinéaire φ vérifie donc l'hypothèse si et seulement si sa matrice par rapport à une base de E et une base de F est inversible. On voit qu'alors $\widetilde{\varphi}$ vérifie la même hypothèse, puisque sa matrice dans les bases données est la transposée de celle de φ. Lorsque $\dim \mathrm{F} = \dim \mathrm{E}$, on peut donc encore dire que D_φ est inversible si et seulement si pour tout $x \neq 0_{\mathrm{E}}$, il existe un vecteur $y \in \mathrm{F}$ tel que $\varphi(x,y) \neq 0$, c'est à dire que $\mathrm{F}^{\perp g} = \{0_{\mathrm{E}}\}$.

Exemple. On a vu que pour la forme bilinéaire canonique φ sur $\mathrm{E}^* \times \mathrm{E}$ définie dans l'exemple 5, l'application linéaire $\mathrm{D}_\varphi = j_{\mathrm{E}} : \mathrm{E} \to \mathrm{E}^{**}$ est un isomorphisme (c'est l'isomorphisme canonique entre E et son bidual).

Dans le cas où D_φ est bijective, on peut calculer la dimension de l'orthogonal d'un sous-espace quelconque, et obtenir un résultat pour le biorthogonal.

Corollaire 1.2.1. *On suppose que* $\dim \mathrm{E} = \dim \mathrm{F} = n$ *et que* φ *est une forme bilinéaire sur* $\mathrm{E} \times \mathrm{F}$ *telle que* D_φ *soit bijective. Pour tout sous-espace vectoriel* E_1 *de* E*, on a*

$$\dim \mathrm{E}_1^{\perp d} = n - \dim \mathrm{E}_1 \, ; \quad (\mathrm{E}_1^{\perp d})^{\perp g} = \mathrm{E}_1.$$

Les mêmes résultats sont valables de façon symétrique pour les sous-espaces vectoriels de l'espace F.

Démonstration. Puisque $\mathrm{D}_{\widetilde{\varphi}}$ est bijective, aucun vecteur non nul de E_1 ne peut être orthogonal à F tout entier, ce qui permet d'appliquer la proposition précédente 1.2.1, et d'en déduire que $\dim \mathrm{E}_1^{\perp d} = \dim \mathrm{F} - \dim \mathrm{E}_1 = n - \dim \mathrm{E}_1$. Puisque l'application linéaire D_φ est elle aussi bijective, on peut appliquer le même argument à $\mathrm{E}_1^{\perp d}$, et obtenir ainsi que $\dim(\mathrm{E}_1^{\perp d})^{\perp g} = \dim \mathrm{E} - (n - \dim \mathrm{E}_1) = \dim \mathrm{E}_1$. Pour conclure à l'égalité des deux sous-espaces, il suffit de remarquer que $\mathrm{E}_1 \subset (\mathrm{E}_1^{\perp d})^{\perp g}$, ce que le lecteur fera facilement.

1.3. Formes quadratiques

Avant de donner la définition, on va essayer de montrer sur deux exemples à quoi ressemblent les formes quadratiques. Commençons par un petit exemple sur $\mathbb{R}^3$. La fonction Q définie pour tout $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ par

$$\mathrm{Q}(x) = \mathrm{Q}(x_1, x_2, x_3) = x_1^2 + 2x_1x_3 + x_2^2$$

est une forme quadratique. On pourrait dire que c'est un polynôme homogène de degré deux dans l'ensemble des variables, mais cette idée ne couvre pas tous les cas envisageables. Prenons en effet un autre exemple, en dimension infinie cette fois. Soit E l'espace des fonctions réelles de classe C^1 sur l'intervalle $[0,1]$, et posons pour tout élément $f \in \mathrm{E}$

$$\mathrm{Q}(f) = \int_0^1 \big(f(t)^2 + 2f(t)f'(t)\big)\, dt.$$

Ce sera encore une forme quadratique, que l'on ne peut plus raisonnablement voir comme polynôme de certaines variables. On va donc poser une définition un peu indirecte.

Définition 1.3.1. Soit E un espace vectoriel sur $\mathbb{K}$; on dit qu'une fonction Q de E dans $\mathbb{K}$ est une *forme quadratique* sur E s'il existe une forme bilinéaire φ sur $\mathrm{E} \times \mathrm{E}$ telle que $\mathrm{Q}(x) = \varphi(x,x)$ pour tout $x \in \mathrm{E}$.

Exemples.

a. Prenons sur $\mathbb{K}^2$ la fonction Q définie pour tout $x = (x_1, x_2)$ par $\mathrm{Q}(x) = x_1x_2$; si on prend $\varphi(x,y) = x_1y_2$, il est clair que φ est une forme bilinéaire sur $\mathbb{K}^2 \times \mathbb{K}^2$ et que

$Q(x) = \varphi(x, x)$ pour tout $x \in \mathbb{K}^2$. Il est clair qu'il y a au moins une autre solution, en prenant $\varphi(x, y) = x_2 y_1$.

b. Posons $Q(x) = x_1^2 + 3x_1x_2 + 5x_2x_3$ pour tout $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ et montrons que Q vérifie la définition ci-dessus. On trouve facilement deux formes bilinéaires solutions de notre question,

$$\varphi_1(x, y) = x_1y_1 + 3x_1y_2 + 5x_2y_3 \quad \text{et} \quad \varphi_2(x, y) = x_1y_1 + 3x_2y_1 + 5x_3y_2$$

ce qui donne aussi une solution symétrique en prenant la demi-somme des deux,

$$\varphi(x, y) = x_1y_1 + \frac{3}{2}(x_1y_2 + x_2y_1) + \frac{5}{2}(x_2y_3 + x_3y_2).$$

Donnons deux propriétés des formes quadratiques qui découlent immédiatement de la définition. Soit Q une forme quadratique sur un espace vectoriel E ; pour tout $x \in E$ et tout scalaire $\lambda \in \mathbb{K}$, on a

$$Q(\lambda x) = \lambda^2 Q(x).$$

D'autre part, la fonction sur $E \times E$ définie par

$$f(x, y) = Q(x + y) - Q(x) - Q(y)$$

est une forme bilinéaire sur $E \times E$, puisque $f(x, y) = \varphi(x, y) + \varphi(y, x)$.

Exercice. Soit E un espace vectoriel de dimension finie sur $\mathbb{K}$; si Q est une fonction sur E telle que $Q(\lambda x) = \lambda^2 Q(x)$ pour tous $\lambda \in \mathbb{K}$, $x \in E$ et telle que $f(x, y) = Q(x+y) - Q(x) - Q(y)$ soit bilinéaire sur $E \times E$, montrer que Q est une forme quadratique (on prendra une base $(e_1, \ldots, e_n)$ de E et on cherchera une forme bilinéaire φ telle que $\varphi(x, x) = Q(x)$ pour tout $x \in E$, vérifiant la condition supplémentaire $\varphi(e_i, e_j) = 0$ pour tous $i > j$). On pourrait donc choisir de définir les formes quadratiques par les deux propriétés ci-dessus.

Dans l'exemple *b* ci-dessus, exemple où $\mathbb{K} = \mathbb{R}$, on a remarqué qu'on pouvait trouver une forme bilinéaire *symétrique* pour définir la forme quadratique étudiée. Si on cherche à faire la même chose avec un corps $\mathbb{K}$ absolument quelconque, on rencontre une difficulté un peu surprenante ; dans un tel corps $\mathbb{K}$, il est naturel de considérer que la notation 2 représente l'élément $1_{\mathbb{K}} + 1_{\mathbb{K}}$ du corps $\mathbb{K}$. Si $2 = 1_{\mathbb{K}} + 1_{\mathbb{K}} \neq 0_{\mathbb{K}}$, on peut par définition d'un corps trouver un inverse de 2 pour la multiplication, qu'on notera $1/2 \in \mathbb{K}$. Si $2 \neq 0$, on peut toujours supposer que la forme bilinéaire φ de la définition ci-dessus est symétrique, en la remplaçant par la forme symétrique $\psi = \frac{1}{2}(\varphi + \widetilde{\varphi})$. Puisque $\psi(x, x) = \varphi(x, x)$, la forme bilinéaire symétrique ψ définit la même forme quadratique que φ.

Le lecteur sensé pourra se demander comment $1 + 1$ pourrait être nul ! Cela se produit quand on utilise un corps $\mathbb{K}$ tel que le corps à deux éléments $\mathbb{Z}/2\mathbb{Z}$, ce qu'heureusement on ne fait pas tous les jours, au moins en DEUG. On notera que si $2 \neq 0_{\mathbb{K}}$, on a aussi $4 \neq 0_{\mathbb{K}}$, ce qui jouera un tout petit rôle un peu plus loin.

Pour les deux exemples introductifs de cette section, on trouvera pour forme bilinéaire symétrique associée :

$$\varphi(x, y) = x_1y_1 + x_1y_3 + x_3y_1 + x_2y_2$$

pour le premier exemple, et pour le second

$$\varphi(f, g) = \int_0^1 \bigl(f(t)g(t) + f(t)g'(t) + f'(t)g(t)\bigr)\, dt.$$

Remarque. Si $\mathbb{K} = \mathbb{Z}/2\mathbb{Z}$, la forme quadratique $Q(x) = x_1x_2$ sur $\mathbb{K}^2$ de l'exemple *a ne peut pas* être représentée par une forme bilinéaire *symétrique* sur $\mathbb{K}^2 \times \mathbb{K}^2$.

Supposons donc que $2 \neq 0$, et supposons que φ soit une forme bilinéaire symétrique telle que $\varphi(x,x) = \mathrm{Q}(x)$ pour tout $x \in \mathrm{E}$. On a alors pour tous $x, y \in \mathrm{E}$ la relation $\mathrm{Q}(x+y) = \mathrm{Q}(x) + 2\varphi(x,y) + \mathrm{Q}(y)$. Ceci permet de trouver φ à partir de Q,

$$2\varphi(x,y) = \mathrm{Q}(x+y) - \mathrm{Q}(x) - \mathrm{Q}(y)\,; \quad 4\varphi(x,y) = \mathrm{Q}(x+y) - \mathrm{Q}(x-y)$$

Cette relation montre que φ (symétrique) est complètement déterminée par Q. On dira que φ est la forme bilinéaire symétrique associée à la forme quadratique Q, obtenue par *polarisation* de la forme quadratique Q ; on dit que φ est la *forme polaire* de Q.

Résumé. *Si* $1_{\mathbb{K}} + 1_{\mathbb{K}} \neq 0$, *pour toute forme quadratique* Q *sur* E *il existe une forme bilinéaire symétrique unique* φ *sur* $\mathrm{E} \times \mathrm{E}$ *telle que* $\mathrm{Q}(x) = \varphi(x,x)$ *pour tout* $x \in \mathrm{E}$. *Cette forme bilinéaire* φ *est définie par*

$$\varphi(x,y) = \frac{1}{4}\big(\mathrm{Q}(x+y) - \mathrm{Q}(x-y)\big) = \frac{1}{2}\big(\mathrm{Q}(x+y) - \mathrm{Q}(x) - \mathrm{Q}(y)\big).$$

Exercice élémentaire : si $2 \neq 0$, trouver la forme polaire des monômes $\mathrm{Q}(x) = x_i x_j$, ou bien $\mathrm{Q}(x) = x_i^2$ (où $x = (x_1, \ldots, x_n) \in \mathbb{K}^n$, et $i, j = 1, \ldots, n$).

Si E est de dimension finie, on considère en général la matrice de φ (qu'on appellera aussi matrice de Q) en prenant deux fois la même base pour E. Soit $\mathbf{e}$ une base de E ; puisque φ est symétrique, il est clair que la matrice $\Phi = \mathrm{Mat}(\varphi, \mathbf{e}, \mathbf{e})$, dont les coefficients sont $\Phi_{i,j} = \varphi(e_i, e_j)$, est symétrique (${}^t\Phi = \Phi$). Si X désigne le vecteur colonne des coordonnées de $x \in \mathrm{E}$ dans la base $\mathbf{e}$, on a

$$\mathrm{Q}(x) = {}^t\mathrm{X}\,\Phi\,\mathrm{X}.$$

Si on effectue le changement de base $\mathrm{X} = \mathrm{P}\mathrm{X}'$, la matrice devient

$$\Phi' = {}^t\mathrm{P}\,\Phi\,\mathrm{P}.$$

Orthogonalité pour une forme bilinéaire symétrique

Soit φ une forme bilinéaire symétrique sur $\mathrm{E} \times \mathrm{E}$; on dit que deux vecteurs x et y sont orthogonaux *par rapport à* φ, ou bien φ-orthogonaux si $\varphi(x,y) = 0$. On a aussi (et heureusement !) $\varphi(y,x) = 0$ puisque φ est symétrique. Si x et y sont φ-orthogonaux, on a $\mathrm{Q}(x+y) = \mathrm{Q}(x) + \mathrm{Q}(y)$ pour la forme quadratique Q associée à φ, définie par $\mathrm{Q}(x) = \varphi(x,x)$ pour tout $x \in \mathrm{E}$. Soit A une partie de l'espace E ; on pose

$$\mathrm{A}^{\perp} = \{y \in \mathrm{E} : \forall x \in \mathrm{A}, \quad \varphi(x,y) = 0\}.$$

L'ensemble $\mathrm{A}^{\perp}$ est un sous-espace vectoriel de E. Si $\mathrm{A}_1 \subset \mathrm{A}_2$, on a $\mathrm{A}_2^{\perp} \subset \mathrm{A}_1^{\perp}$. On a toujours $\mathrm{A} \subset \mathrm{A}^{\perp\perp}$.

Noyau d'une forme bilinéaire symétrique

Soit φ une forme bilinéaire symétrique sur E ; l'orthogonal de E est le sous-espace

$$\mathrm{N}_{\varphi} = \mathrm{E}^{\perp} = \{y \in \mathrm{E} : \forall x \in \mathrm{E}, \quad \varphi(x,y) = 0\}.$$

On l'appelle le *noyau* de la forme φ (ou de la forme quadratique Q). On voit facilement que c'est aussi le noyau de l'application linéaire D_{φ} de E dans E^*.

Il est facile de trouver l'expression matricielle du noyau de φ : si $\Phi = \mathrm{Mat}(\varphi, \mathbf{e}, \mathbf{e})$ et si Y est le vecteur colonne des coordonnées de $y \in \mathrm{E}$ dans la base $\mathbf{e}$, on voit que $y \in \mathrm{N}_{\varphi}$ si et seulement si $\Phi\mathrm{Y} = 0$.

Définition 1.3.2. Soient E un espace vectoriel de dimension finie, φ une forme bilinéaire symétrique sur $\mathrm{E} \times \mathrm{E}$ et Q la forme quadratique sur E associée à φ. On appelle *rang de la forme bilinéaire symétrique* φ, ou bien rang de la forme quadratique Q la quantité $r = \dim \mathrm{E} - \dim \mathrm{N}_\varphi$.

Dans n'importe quelle base $\mathbf{e}$ de E, le rang de φ est égal au rang de la matrice $\Phi = \mathrm{Mat}(\varphi, \mathbf{e}, \mathbf{e})$. En effet, le noyau N_Φ de l'application $\mathrm{Y} \to \Phi\mathrm{Y}$ est en bijection linéaire avec N_φ, et le rang de la matrice Φ est égal à $n - \dim \mathrm{N}_\Phi = \dim \mathrm{E} - \dim \mathrm{N}_\varphi$.

Formes bilinéaires symétriques non dégénérées.

On suppose ici que E est de dimension finie > 0.

Définition 1.3.3. On dit qu'une forme bilinéaire symétrique φ sur $\mathrm{E} \times \mathrm{E}$ est non dégénérée si son noyau N_φ est réduit à $\{0_\mathrm{E}\}$.

Cela signifie que pour tout $y \neq 0_\mathrm{E}$, il existe un vecteur $x \in \mathrm{E}$ tel que $\varphi(x, y) \neq 0$. Cela équivaut encore à dire que l'application D_φ est injective de E dans E^*, donc bijective puisque les deux espaces ont la même dimension. La forme bilinéaire φ est donc non dégénérée si et seulement si sa matrice par rapport à une base quelconque de E est inversible. Cela équivaut aussi à dire que $\mathrm{rang}\,\varphi = \dim \mathrm{E}$.

Exemple. La forme hyperbolique de l'exemple 4 est non dégénérée : pour tout vecteur $v = (x, y, z, t)$ non nul, on a $\varphi(v, e_j) \neq 0$ pour au moins un vecteur e_j de la base canonique de $\mathbb{R}^4$. On peut dire aussi que la matrice de φ par rapport à la base canonique est inversible : c'est la matrice diagonale de coefficients diagonaux $1, 1, 1, -1$.

Orthogonalité pour une forme bilinéaire non dégénérée

La proposition 1.2.1 nous donne en particulier

On suppose que φ *est une forme bilinéaire symétrique sur* $\mathrm{E} \times \mathrm{E}$, *où* E *est de dimension finie. Pour tout sous-espace vectoriel* F *de* E, *tel qu'aucun vecteur non nul* $x \in \mathrm{F}$ *ne puisse être orthogonal à* E *tout entier, (c'est à dire que* $\mathrm{F} \cap \mathrm{N}_\varphi = \{0_\mathrm{E}\}$*) on a*

$$\dim \mathrm{F}^\perp = \dim \mathrm{E} - \dim \mathrm{F}.$$

Si φ est non dégénérée, l'hypothèse ci-dessus est vérifiée pour tout sous-espace F de l'espace E. On obtient donc :

Corollaire 1.3.1. *On suppose que* $\dim \mathrm{E} < +\infty$ *et que* φ *est une forme bilinéaire symétrique non dégénérée sur* $\mathrm{E} \times \mathrm{E}$. *Pour tout sous-espace vectoriel* F *de* E, *on a*

$$\dim \mathrm{F}^\perp = \dim \mathrm{E} - \dim \mathrm{F}\,; \quad (\mathrm{F}^\perp)^\perp = \mathrm{F}.$$

Supposons maintenant que φ soit une forme bilinéaire symétrique sur $\mathrm{E} \times \mathrm{E}$, et que F soit un sous-espace vectoriel de E. On peut considérer la restriction de φ au produit $\mathrm{F} \times \mathrm{F}$. Dire que cette restriction ψ est non dégénérée sur $\mathrm{F} \times \mathrm{F}$ signifie exactement qu'aucun vecteur non nul de F ne peut être orthogonal à F tout entier, ce qui se traduit par

$$\mathrm{F} \cap \mathrm{F}^\perp = \{0_\mathrm{E}\}.$$

Exemple. Reprenons la forme bilinéaire non dégénérée

$$\varphi(x, y) = x_1 y_1 + x_2 y_2 + x_3 y_3 - x_4 y_4$$

sur $\mathbb{R}^4$. Considérons le vecteur non nul $v_0 = (1, 0, 0, 1)$. On a $\varphi(v_0, v_0) = 0$, donc le vecteur non nul v_0 est φ-orthogonal à lui-même ! On dit que v_0 est un vecteur *isotrope*. De plus si on pose $\mathrm{F} = \mathbb{R}v_0$ on voit que la restriction ψ de φ à $\mathrm{F} \times \mathrm{F}$ est nulle, donc ψ est dégénérée.

Il ne faut pas confondre le fait que la forme bilinéaire φ est non dégénérée (ce qui est vrai dans l'exemple présent) et le fait que $\varphi(x, x)$ n'est pas nul si $x \neq 0_\mathrm{E}$ (qui est faux dans ce même exemple). Pour tout compliquer, on verra plus loin que pour les formes bilinéaires *positives*, il y a identité de ces deux propriétés !

Proposition 1.3.1. *Soient* E *un espace vectoriel de dimension finie,* φ *une forme bilinéaire symétrique sur* $\mathrm{E} \times \mathrm{E}$ *et* F *un sous-espace vectoriel de* E *; la restriction* ψ *de* φ *à* $\mathrm{F} \times \mathrm{F}$ *est non dégénérée si et seulement si* $\mathrm{E} = \mathrm{F} \oplus \mathrm{F}^\perp$.

Démonstration. Supposons d'abord la restriction ψ non dégénérée. On sait alors que l'on a $\mathrm{F} \cap \mathrm{F}^\perp = \{0\}$. *A fortiori*, aucun vecteur non nul de F ne peut être orthogonal à l'espace E tout entier, donc la proposition 1.2.1 s'applique, et permet de voir que $\dim \mathrm{F}^\perp = \dim \mathrm{E} - \dim \mathrm{F}$. Puisque l'intersection de F et $\mathrm{F}^\perp$ est nulle, on trouve que $\mathrm{E} = \mathrm{F} \oplus \mathrm{F}^\perp$.

Réciproquement, dire que la somme de F et $\mathrm{F}^\perp$ est une somme directe sous-entend que $\mathrm{F} \cap \mathrm{F}^\perp = \{0\}$, ce qui équivaut à dire que ψ est non dégénérée.

Exemples.

1. Posons $\mathrm{Q}(x) = x_1^2$ pour tout $x \in \mathbb{R}^2$. Le noyau est alors $\mathrm{N} = \mathbb{R}e_2$, donc Q est dégénérée. Mais si on prend $\mathrm{F} = \mathbb{R}e_1$, la restriction de Q à F est non dégénérée ; on obtient $\mathrm{F}^\perp = \mathbb{R}e_2$ qui est bien de dimension $2 - 1 = 1$, et de plus $\mathrm{F} \cap \mathrm{N} = \{0\}$, ce qui correspond à $\mathbb{R}^2 = \mathrm{F} \oplus \mathrm{F}^\perp$. En revanche si on part du sous-espace $\mathrm{G} = \mathbb{R}e_2$, l'orthogonal $\mathrm{G}^\perp$ est $\mathbb{R}^2$ tout entier et la somme $\mathrm{G} + \mathrm{G}^\perp$ n'est pas directe.

2. Posons $\mathrm{Q}(x) = x_1^2 + x_2^2 - x_3^2$. Cette forme est non dégénérée. Posons $v = (1, 0, 1)$. L'orthogonal de $\mathrm{F} = \mathbb{R}v$ est bien de dimension $2 = \dim \mathrm{E} - \dim \mathrm{F} = 3 - 1$ puisque Q est non dégénérée, mais $\mathbb{R}v \subset (\mathbb{R}v)^\perp$ parce que v est isotrope, donc la somme $\mathrm{F} + \mathrm{F}^\perp$ n'est pas directe. Effectivement, la restriction de φ à $\mathrm{F} = \mathbb{R}v$ est dégénérée.

Expression dans une base φ-orthogonale

Soient φ une forme bilinéaire symétrique sur $\mathrm{E} \times \mathrm{E}$ et Q la forme quadratique sur E associée à φ ; si $e_1, \ldots, e_k$ sont deux à deux φ-orthogonaux, on a

$$\mathrm{Q}(c_1 e_1 + \cdots + c_k e_k) = c_1^2\, \mathrm{Q}(e_1) + \cdots + c_k^2\, \mathrm{Q}(e_k).$$

Si $(e_1, \ldots, e_n)$ est une base de E formée de vecteurs deux à deux φ-orthogonaux (on verra plus loin que si E est de dimension finie et $2 \neq 0$, on peut toujours trouver une telle base φ-orthogonale pour E), la matrice $\Phi = \mathrm{Mat}(\varphi, \mathbf{e}, \mathbf{e})$ est diagonale. Les coefficients diagonaux sont $\mathrm{Q}(e_1), \ldots, \mathrm{Q}(e_n)$.

Supposons que $\mathrm{Q}(e_i) \neq 0$ pour $i = 1, \ldots, s$ et que $\mathrm{Q}(e_i) = 0$ pour $i > s$, où s est un entier tel que $0 \leq s \leq n$ (si $s = 0$, on veut dire que tous les $(\mathrm{Q}(e_i))$ sont nuls, et si $s = n$, tous les $(\mathrm{Q}(e_i))$ sont non nuls). Il est facile de déterminer le noyau de N_φ en utilisant la base $\mathbf{e}$. On voit que

$$\mathrm{N}_\varphi = [e_{s+1}, \ldots, e_n]$$

ce qui montre aussi que le rang de φ est égal à s.

Si des vecteurs $x_1, \ldots, x_p$ de E sont deux à deux orthogonaux pour la forme φ et si on a $x = c_1 x_1 + \cdots + c_p x_p$, on voit que

$$\varphi(x, x_j) = c_j \varphi(x_j, x_j), \quad \varphi(x, x) = \sum_{j=1}^{p} c_j^2 \, \varphi(x_j, x_j).$$

Si tous les vecteurs sont non isotropes, $\varphi(x_j, x_j) \neq 0$ pour $j = 1, \ldots, p$ on voit que les coefficients de la combinaison linéaire x sont uniquement déterminés,

$$c_i = \frac{\varphi(x, x_i)}{\varphi(x_i, x_i)}.$$

Lemme. *Si des vecteurs sont deux à deux φ-orthogonaux et non isotropes, ils sont linéairement indépendants.*

Remarque. Le résultat n'est pas vrai pour des vecteurs isotropes : dans l'exemple 4, si $v_0 = (1, 0, 0, 1)$, le système (v_0, v_0) est un système de vecteurs φ-orthogonaux non nuls, mais bien sûr pas indépendants !

Proposition 1.3.2. *On suppose que $2 \neq 0$. Soit φ une forme bilinéaire symétrique sur un espace vectoriel* E *de dimension finie ; il existe une base $e_1, \ldots, e_n$ de* E *formée de vecteurs deux à deux φ-orthogonaux.*

Démonstration. Par récurrence sur la dimension de E. Si E est de dimension 1, on choisit $e_1 \neq 0$ dans E ; ce vecteur constitue une base de E, et puisqu'il est seul, il n'y a pas d'orthogonalité à vérifier. Supposons maintenant $n > 1$, et supposons le résultat établi pour toute forme bilinéaire symétrique ψ sur un espace vectoriel G de dimension $< n$. Soit φ une forme bilinéaire symétrique sur E, avec $\dim \mathrm{E} = n$, et posons $\mathrm{Q}(x) = \varphi(x, x)$ pour tout $x \in \mathrm{E}$; si φ est nulle, il suffit de prendre une base quelconque $(e_1, \ldots, e_n)$ de E : les vecteurs de base seront automatiquement φ-orthogonaux. Sinon, φ est non nulle, et puisque $\varphi(x, y) = \frac{1}{2}\bigl(\mathrm{Q}(x+y) - \mathrm{Q}(x) - \mathrm{Q}(y)\bigr)$, la forme quadratique Q ne peut pas être identiquement nulle sur E. Il existe donc un vecteur $v \in \mathrm{E}$ tel que $\varphi(v, v) \neq 0$; posons $e_1 = v$ et

$$\mathrm{G} = \{x \in \mathrm{E} : \varphi(x, e_1) = 0\} = \{e_1\}^{\perp}.$$

Puisque la forme linéaire $x \to \varphi(x, e_1)$ n'est pas nulle, son noyau G est de dimension $n - 1$, et $e_1 \notin \mathrm{G}$, donc $\mathrm{E} = \mathrm{G} \oplus \mathbb{K} e_1$. Définissons sur G la forme bilinéaire symétrique ψ par $\psi(x, y) = \varphi(x, y)$ pour tous $x, y \in \mathrm{G}$ (c'est la restriction de φ à $\mathrm{G} \times \mathrm{G}$). D'après l'hypothèse de récurrence, il existe une base $(e_2, \ldots, e_n)$ de G formée de vecteurs deux à deux ψ-orthogonaux. Le système $(e_1, e_2, \ldots, e_n)$ est alors une base de E. Par définition de G on a $\varphi(e_i, e_1) = 0$ pour tout $i > 1$, et pour $2 \leq i, j$, $i \neq j$ on a $\varphi(e_i, e_j) = \psi(e_i, e_j) = 0$, donc on a trouvé une base φ-orthogonale pour E.

Remarque. La matrice de φ dans la base φ-orthogonale $(e_1, \ldots, e_n)$ est diagonale, les éléments diagonaux étant $\varphi(e_1, e_1), \ldots, \varphi(e_n, e_n)$.

ATTENTION. Il ne s'agit pas ici d'une diagonalisation de matrice au sens de la théorie des valeurs propres et des vecteurs propres. Si Φ est la matrice de φ dans une première base, et si Φ' est la matrice de φ dans une base φ-orthogonale, la matrice Φ' est diagonale, mais ses valeurs diagonales ne sont pas nécessairement des valeurs propres de Φ (ces deux matrices n'ont pas en général le même polynôme caractéristique ; la formule de changement de base n'est pas celle qui donne l'invariance du polynôme caractéristique, c'est ici la formule $\Phi' = {}^t\mathrm{P}\,\Phi\,\mathrm{P}$. Si par exemple Φ est déjà diagonale et si P est égale à $c\,\mathrm{I}_n$, les coefficients diagonaux seront multipliés par c^2 et n'auront visiblement aucune raison d'être des valeurs propres de Φ).

Le cas réel

Supposons dans ce paragraphe que $\mathbb{K} = \mathbb{R}$. On va simplifier au maximum l'expression que l'on peut trouver pour la matrice de φ dans une base φ-orthogonale. Supposons donc que φ soit bilinéaire symétrique sur un espace vectoriel réel E de dimension finie, et soit $\mathbf{e} = (e_1, \ldots, e_n)$ une base φ-orthogonale ; on va remplacer cette base par une base de la forme $\mathbf{e}' = (t_1 e_1, \ldots, t_n e_n)$, avec $t_i > 0$. Si $\mathrm{Q}(e_i) \neq 0$, on pose $t_i = |\mathrm{Q}(e_i)|^{-1/2}$, et si $\mathrm{Q}(e_i) = 0$, on pose par exemple $t_i = 1$ (mais ça n'a en fait aucune importance). Dans la nouvelle base $(e'_1, \ldots, e'_n)$, on a encore l'orthogonalité, puisque $\varphi(t_i e_i, t_j e_j) = t_i t_j \varphi(e_i, e_j) = 0$ si $i \neq j$, et d'autre part $\mathrm{Q}(e'_i) = \mathrm{Q}(t_i e_i) = t_i^2 \mathrm{Q}(e_i)$ est nul si $\mathrm{Q}(e_i) = 0$, sinon
$$\mathrm{Q}(e'_i) = \mathrm{Q}(t_i e_i) = \mathrm{Q}(e_i)/|\mathrm{Q}(e_i)| = \pm 1,$$
selon le signe de $\mathrm{Q}(e_i)$. En résumé :

Lorsque $\mathbb{K} = \mathbb{R}$, *il existe une base φ-orthogonale dans laquelle la matrice de φ est diagonale, avec des coefficients diagonaux égaux* à 1, -1 *ou* 0.

Dans le cas complexe, on peut toujours trouver $t_i \in \mathbb{C}$ tel que $t_i^2 = \mathrm{Q}(e_i)$. On obtient alors :

Lorsque $\mathbb{K} = \mathbb{C}$, *il existe une base φ-orthogonale dans laquelle la matrice de φ est diagonale, avec des coefficients diagonaux égaux* à 1 *ou* 0.

Les résultats précédents se traduisent matriciellement de la façon suivante : désignons par $\mathrm{J}_{p,q,n}$ la matrice diagonale de taille $n \times n$ dont les p premiers coefficients diagonaux sont égaux à 1, les q suivants égaux à -1 et les restants égaux à 0 (on suppose $p, q \geq 0$ et $p + q \leq n$). Pour toute matrice symétrique réelle A de taille $n \times n$, il existe une matrice réelle inversible P telle que ${}^t\mathrm{P\,A\,P}$ soit égale à l'une des matrices $\mathrm{J}_{p,q,n}$. Pour toute matrice symétrique complexe A de taille $n \times n$, il existe une matrice complexe inversible P telle que ${}^t\mathrm{P\,A\,P}$ soit égale à l'une des matrices $\mathrm{J}_{p,0,n}$.

Expliquons le cas réel. Considérons la forme bilinéaire symétrique définie pour $\mathrm{X}, \mathrm{Y} \in \mathbb{R}^n$ par la formule $\varphi(\mathrm{X}, \mathrm{Y}) = {}^t\mathrm{X\,A\,Y}$. On peut trouver une base $\mathbf{f}$ de $\mathbb{R}^n$ qui soit φ-orthogonale, et deux entiers $p, q \geq 0$ tels que $p + q \leq n$ et tels que $\varphi(f_j, f_j) = 1$ pour $1 \leq j \leq p$, $\varphi(f_j, f_j) = -1$ pour $p + 1 \leq j \leq p + q$ et $\varphi(f_j, f_j) = 0$ si $j > p + q$. Dans cette base $\mathbf{f}$ la matrice de φ est égale à $\mathrm{J}_{p,q,n}$. Si P désigne la matrice de passage, P est une matrice réelle inversible et ${}^t\mathrm{P\,A\,P} = \mathrm{J}_{p,q,n}$. Le cas complexe s'explique de façon analogue.

Dans le cas réel, le rang de la forme φ est égal à $p + q$, et on verra plus loin dans le paragraphe *signature* que les deux entiers p et q ne dépendent que de A, et pas de la base φ-orthogonale particulière.

Combinaisons linéaires de carrés de formes linéaires

Revenons au cas général. Supposons que $\mathbf{e}$ soit une base φ-orthogonale de l'espace vectoriel E. Si $x = \sum_{i=1}^n c_i e_i$, on aura par orthogonalité $\mathrm{Q}(x) = \sum_{i=1}^n \mathrm{Q}(e_i)\, c_i^2$ soit encore en introduisant la base duale et en utilisant la relation $c_i = e_i^*(x)$
$$\mathrm{Q} = \sum_{i=1}^n \mathrm{Q}(e_i)\, (e_i^*)^2.$$
On a donc exprimé Q comme combinaison linéaire de carrés de formes linéaires. Inversement, soient $\ell_1, \ldots, \ell_k$ des formes linéaires sur E et $a_1, \ldots, a_k$ des coefficients dans $\mathbb{K}$.

Posons

$$\mathrm{Q}(x) = \sum_{i=1}^{k} a_i(\ell_i(x))^2.$$

Cette forme quadratique admet clairement pour forme polaire

$$\varphi(x, y) = \sum_{i=1}^{k} a_i\, \ell_i(x)\ell_i(y).$$

Si $\ell_1, \ldots, \ell_k$ sont indépendantes, on peut prolonger en une base $(\ell_1, \ldots, \ell_n)$ de E^*, puis trouver une base $\mathbf{e} = (e_1, \ldots, e_n)$ de E telle que $\ell_i(e_j) = \delta_{i,j}$ pour tous $i, j = 1, \ldots, n$ (utiliser le corollaire 1.1.2). On vérifie alors que

$$\varphi(e_i, e_j) = \sum_{s=1}^{k} a_i\, \ell_s(e_i)\ell_s(e_j) = 0$$

si $i \neq j$ (on remarque que $\ell_s(e_i)\ell_s(e_j) = \delta_{s,i}\delta_{s,j} = 0$ pour tout s lorsque $i \neq j$), donc $\mathbf{e}$ est une base φ-orthogonale. On a ainsi vu le rapport entre la recherche d'une base φ-orthogonale et l'expression de la forme quadratique Q comme combinaison linéaire de carrés de formes linéaires indépendantes.

Décomposition de Gauss

La méthode de Gauss est une méthode pratique très simple qui permet de trouver une base orthogonale pour une forme bilinéaire symétrique φ. En réalité, il s'agit d'une méthode qui va permettre d'exprimer la forme quadratique Q associée à φ sous forme de combinaison linéaire de carrés de formes linéaires indépendantes. On pourra ensuite en déduire une base φ-orthogonale comme on l'a expliqué dans le paragraphe précédent.

Commençons par un exemple très simple, où la forme quadratique Q est définie sur $\mathbb{R}^2$ par $\mathrm{Q}(x) = x_1^2 + x_1x_2 + x_2^2$. On écrit le début d'un carré, $(x_1 + \frac{1}{2}x_2)^2 = x_1^2 + x_1x_2 + \frac{1}{4}x_2^2$, donc

$$\mathrm{Q}(x) = \bigl(x_1 + \frac{1}{2}x_2\bigr)^2 + \frac{3}{4}\, x_2^2 = \ell_1(x)^2 + \ell_2(x)^2.$$

A partir de cette expression, on peut trouver une base orthogonale de la façon suivante : les deux formes linéaires sont ℓ_1 et ℓ_2, définies par $\ell_1(x) = x_1 + \frac{1}{2}x_2$ et $\ell_2(x) = \frac{\sqrt{3}}{2}x_2$. On cherche la base voulue en résolvant les équations $\ell_i(e_j) = \delta_{i,j}$, $i, j = 1, 2$. On obtient ainsi

$$e_1 = (1, 0)\,; \quad e_2 = (-1/\sqrt{3}, 2/\sqrt{3}).$$

Dans cet exemple, on voit que Q est une *somme* de deux carrés. Il en résulte que l'expression $\mathrm{Q}(x)$ est toujours ≥ 0 (en fait $\mathrm{Q}(x) > 0$ pour tout $x \neq 0$, car les relations $\ell_1(x) = 0$ et $\ell_2(x) = 0$ entraînent ici que $x = 0$). Une des applications de la méthode de Gauss est donc l'étude du signe d'une forme quadratique réelle, une question que nous reverrons en Calcul Différentiel à propos de la formule de Taylor à l'ordre deux pour une fonction de plusieurs variables. Mais la méthode de Gauss s'applique aussi à des corps généraux (pour lesquels il n'y a pas de notion de *signe* d'une expression), à condition toutefois que $2 \neq 0$.

Proposition 1.3.3. *Soit $\mathbb{K}$ un corps tel que $2 \neq 0$, et soit Q une forme quadratique sur $\mathbb{K}^n$. Il existe n formes linéaires indépendantes $\ell_1, \ldots, \ell_n$ sur $\mathbb{K}^n$ et des coefficients*

$c_1, \ldots, c_n$ dans $\mathbb{K}$ tels que pour tout $x \in \mathbb{K}^n$

$$\mathrm{Q}(x) = \sum_{i=1}^{n} c_i \, (\ell_i(x))^2.$$

Le rang de Q est égal au nombre des coefficients c_i non nuls.

Démonstration. On démontre l'existence de la décomposition par récurrence sur le nombre de variables. Soit Q une forme quadratique sur $\mathbb{K}^n$, considérée comme fonction $\mathrm{Q}(x_1, \ldots, x_n)$ de n variables $x_1, \ldots, x_n \in \mathbb{K}$. On peut écrire

$$\mathrm{Q}(x) = \sum_{i=1}^{n} a_i x_i^2 + \sum_{i<j} b_{i,j} x_i x_j$$

Supposons d'abord que l'un des coefficients a_i soit non nul, et pour simplifier l'écriture, supposons que $i = 1$. On écrit alors Q sous la forme

$$\mathrm{Q}(x) = a_1 x_1^2 + \ell(y) \, x_1 + q(y),$$

où $a_1 \neq 0$, $y = (x_2, \ldots, x_n)$, ℓ est une forme linéaire qui ne dépend pas de x_1 et q une forme quadratique qui ne dépend pas de x_1. Ensuite,

$$\mathrm{Q}(x) = a_1 \bigl(x_1 + \frac{1}{2a_1} \, \ell(y)\bigr)^2 + \bigl(q(y) - \frac{1}{4a_1} \, \ell(y)^2\bigr),$$

et on applique l'hypothèse de récurrence à la forme quadratique $\widetilde{\mathrm{Q}}(y) = q(y) - \frac{1}{4a_1} \, \ell(y)^2$, qui ne dépend plus que des $(n-1)$ variables $x_2, \ldots, x_n$. Cette forme quadratique est par hypothèse de récurrence combinaison linéaire de carrés de formes linéaires indépendantes $\ell_2, \ldots, \ell_n$ qui ne dépendent que des variables $x_2, \ldots, x_n$. Ces formes sont alors indépendantes de la forme linéaire $\ell_1(x) = x_1 + \frac{1}{2a_1} \, \ell(y)$ (parce que ℓ_1 dépend explicitement de la variable x_1 ; pour mieux se convaincre, on pourra voir quelle est la forme de la matrice $n \times n$ dont les lignes contiennent les coefficients des formes linéaires $\ell_1, \ldots, \ell_n$).

Le deuxième cas est plus embêtant. C'est le cas où tous les coefficients a_i sont nuls, sans que la forme Q soit nulle. Soit alors (i, j) un couple tel que $b_{i,j}$ soit non nul, et supposons pour simplifier l'écriture que $(i, j) = (1, 2)$. On écrit maintenant

$$\mathrm{Q}(x) = b_{1,2} \, x_1 x_2 + \ell(y) x_1 + m(y) x_2 + q(y),$$

où $y = (x_3, \ldots, x_n)$, ℓ et m sont des formes linéaires qui ne dépendent pas de x_1, x_2 et q une forme quadratique qui ne dépend pas de x_1, x_2. On pose $u_1 = (x_1 + x_2)/2$, $u_2 = (x_1 - x_2)/2$, et on transforme l'expression précédente en

$$\mathrm{Q}(x) = b_{1,2} \, (u_1^2 - u_2^2) + \ell(y)(u_1 + u_2) + m(y)(u_1 - u_2) + q(y),$$

que l'on traite par la méthode précédente, appliquée aux nouvelles variables

$$(u_1, u_2, x_3, \ldots, x_n).$$

On termine en revenant aux variables initiales.

Exercices.

1. Traiter l'exemple $\mathrm{Q}(x) = x^2 + yz + y^2 - 2axz + z^2$.

2. Un exemple sans carré : $xy + yz + czx$. On pose $x = u + v$ et $y = u - v$, on fait le calcul avec les variables u, v, z et on change à la fin. En déduire une base φ-orthogonale de $\mathbb{R}^3$.

Le cas réel

On supposera désormais que $\mathbb{K} = \mathbb{R}$. Dans la méthode de Gauss, on pourra simplifier chacun des termes $c\,\ell(x)^2$ (où $c \in \mathbb{R}$) de la proposition précédente de la façon suivante : on pose $m(x) = \sqrt{|c|}\,\ell(x)$; si $c > 0$, on remplace $c\,\ell(x)^2$ par $m(x)^2$, et si $c < 0$, on remplace $c\,\ell(x)^2$ par $-m(x)^2$. Si c est nul, on n'écrit rien du tout (et en fait dans l'application pratique de la méthode de Gauss, on ne calcule en général pas celles des formes linéaires ℓ_j de la proposition précédente pour lesquelles $c_j = 0$).

Si on a une forme quadratique Q qui n'est pas définie sur $\mathbb{R}^n$ mais sur un espace vectoriel réel “abstrait” de dimension n, on choisit une base de E et on effectue les calculs avec les coordonnées $x_1, \ldots, x_n$ dans cette base. En résumé :

Etant donnée une forme quadratique Q *définie sur un espace vectoriel réel* E *de dimension finie, il existe des formes linéaires indépendantes* $\ell_1, \ldots, \ell_p, \ell_{p+1}, \ldots, \ell_{p+q}$ *telles que*

$$\forall x \in \mathrm{E}, \quad \mathrm{Q}(x) = \ell_1(x)^2 + \cdots + \ell_p(x)^2 - \ell_{p+1}(x)^2 - \cdots - \ell_{p+q}(x)^2.$$

Bien entendu, il est possible que $p = 0$ (il n'y a alors pas de signe $+$), ou bien $q = 0$ (il n'y a alors pas de signe $-$), ou bien les deux (si $\mathrm{Q} = 0$). Si on complète ce système de formes linéaires en une base $(\ell_1, \ldots, \ell_n)$ de E^*, on peut trouver une base $(e_1, \ldots, e_n)$ de E telle que $\ell_i(e_j) = \delta_{i,j}$. Dans cette base la matrice $\Phi = \mathrm{Mat}(\varphi, \mathbf{e}, \mathbf{e})$ est diagonale, avec $\Phi_{1,1} = \cdots = \Phi_{p,p} = 1$, $\Phi_{p+1,p+1} = \cdots = \Phi_{p+q,p+q} = -1$ et $\Phi_{i,i} = 0$ si $i > p + q$. Le rang de la forme φ est égal à $p + q$. On voit donc que φ est non-dégénérée si et seulement si $p + q = n$. Exprimé en termes de base φ-orthogonale, le résultat précédent donne :

Il existe une base φ-orthogonale dans laquelle la matrice de φ est diagonale, avec p coefficients diagonaux égaux à 1, *et q égaux à* -1, *les autres étant* 0.

La décomposition de Gauss permet d'étudier le signe de la forme quadratique Q. Si $q = 0$, on a $\mathrm{Q}(x) \geq 0$ pour tout x ; si $p = 0$, on a $\mathrm{Q}(x) \leq 0$ pour tout x. Dans le cas où $p > 0$ **et** $q > 0$, on est sûr que Q change de signe : en effet, $\ell_1(e_1) > 0$ mais $\ell_{p+1}(e_{p+1}) < 0$. Si $p + q < n$, il existe des vecteurs non nuls tels que $\mathrm{Q}(x) = 0$, par exemple $x = e_{p+q+1}$. Si $p = n$, on aura $\mathrm{Q}(x) > 0$ pour tout $x \neq 0$, et si $q = n$, on aura $\mathrm{Q}(x) < 0$ pour tout $x \neq 0$.

Signature d'une forme quadratique réelle

Définition 1.3.4. Soit Q une forme quadratique sur un espace vectoriel réel E et soit φ sa forme polaire ; on dit que φ (ou Q) est *positive* sur E si $\mathrm{Q}(x) \geq 0$ pour tout vecteur $x \in \mathrm{E}$. On dit que φ (ou Q) est *définie positive* sur E si $\mathrm{Q}(x) > 0$ pour tout vecteur $x \in \mathrm{E}$ non nul.

Lemme. *Soit φ une forme bilinéaire symétrique sur un espace vectoriel réel* E *de dimension finie, et soit $(x_1, \ldots, x_n)$ une base φ-orthogonale de* E.

– *Si on a $\varphi(x_i, x_i) \geq 0$ pour tout $i = 1, \ldots, n$, la forme φ est positive sur* E.
– *Si on a $\varphi(x_i, x_i) > 0$ pour tout $i = 1, \ldots, n$, la forme φ est définie positive sur* E.

Démonstration. Elle est très facile ; expliquons uniquement la seconde partie, qui est très légèrement plus délicate. Soit x un vecteur de E, décomposé dans la base $\mathbf{x} = (x_1, \ldots, x_n)$ sous la forme $x = \sum_{i=1}^n c_i x_i$. Puisque la base $\mathbf{x}$ est φ-orthogonale, on obtient

$$\varphi(x, x) = \sum_{i=1}^n c_i^2\,\varphi(x_i, x_i)$$

quantité positive puisque les c_i sont réels et $\varphi(x_i, x_i) > 0$ par hypothèse. Si $\varphi(x, x) = 0$, il faut que tous les termes de la somme soient nuls, c'est à dire $c_i^2\varphi(x_i, x_i) = 0$ pour $i = 1, \ldots, n$, ce qui implique $c_i = 0$ puisque $\varphi(x_i, x_i) \neq 0$, donc $x = 0_{\mathrm{E}}$.

On définit de même la notion de forme bilinéaire symétrique *négative*, ou bien *définie négative*, et on a bien sûr le lemme analogue à celui qui précède (il suffit de remplacer φ par $-\varphi$).

Exercices.

1. Soit A la matrice dans la base canonique d'une forme bilinéaire symétrique φ sur $\mathbb{R}^2$. Montrer que φ est définie positive si et seulement si $\mathrm{A}_{1,1} > 0$ et $\det \mathrm{A} > 0$.

2. Si Φ est la matrice d'une forme bilinéaire symétrique définie positive sur $\mathbb{R}^n$, montrer que :

det $\Phi > 0$ (introduire une base orthogonale et changer de base) ;

pour tout $k = 1, \ldots, n$, la matrice Φ_k de taille $k \times k$ obtenue en ne gardant que les coefficients $\Phi_{i,j}$ tels que $1 \leq i, j \leq k$ définit une forme bilinéaire définie positive sur $\mathbb{R}^k$;

il existe une matrice triangulaire supérieure inversible **réelle** U telle que $\Phi = {}^t\mathrm{U}\,\mathrm{U}$. Réciproque ?

On dira que Q est de *signature* (p, q) si la décomposition dans une base orthogonale contient p termes tels que $\mathrm{Q}(e_i) = 1$ (*somme* de carrés) et q termes tels que $\mathrm{Q}(e_i) = -1$ (*différence* de carrés). Cette définition n'a de sens que si on montre que ces deux nombres ne dépendent que de Q, et pas de la décomposition particulière.

Supposons donc que $(e_1, \ldots, e_n)$ soit une base φ-orthogonale de E, avec $\varphi(e_i, e_i) > 0$ pour $i = 1, \ldots, p$ et $\varphi(e_i, e_i) \leq 0$ pour $i > p$. La forme Q est alors définie positive sur le sous-espace F de dimension p engendré par $e_1, \ldots, e_p$ et négative sur le sous-espace G de dimension $n - p$ engendré par les vecteurs $e_{p+1}, \ldots, e_n$. Soit $\mathbf{e}' = (e'_1, \ldots, e'_n)$ une autre base φ-orthogonale, et supposons que maintenant $\varphi(e'_i, e'_i) > 0$ pour $i = 1, \ldots, p'$, et $\varphi(e'_i, e'_i) \leq 0$ pour $i > p'$. Désignons par F′ le sous-espace engendré par $(e'_1, \ldots, e'_{p'})$ et par G′ le sous-espace engendré par les autres vecteurs de la base $\mathbf{e}'$. La forme φ est alors négative sur G′. Il en résulte que $\mathrm{F} \cap \mathrm{G}' = \{0\}$ (en effet, si $x \in \mathrm{F}$ et $x \neq 0$, on a $\mathrm{Q}(x) > 0$, donc $x \notin \mathrm{G}'$). On en déduit que F et G′ forment une somme directe, donc $p + (n - p') = \dim(\mathrm{F} + \mathrm{G}') \leq \dim \mathrm{E} = n$, ce qui donne $p - p' \leq 0$. En raisonnant de même sur F′ et G, on obtient aussi $p' - p \leq 0$, donc $p = p'$. Le raisonnement est le même pour montrer que $q = q'$.

Définition 1.3.5. On dit que la forme quadratique Q définie sur l'espace vectoriel réel de dimension finie E, et de forme polaire φ, est *de signature* (p, q) si sa matrice dans une base φ-orthogonale contient p coefficients > 0 et q coefficients < 0 sur la diagonale.

Autrement dit, la forme quadratique Q est de signature (p, q) s'il existe une base φ-orthogonale $(e_1, \ldots, e_n)$ pour E telle qu'il existe p indices i pour lesquels $\mathrm{Q}(e_i) > 0$ et q indices tels que $\mathrm{Q}(e_i) < 0$. Dans ce cas, on aura la même situation pour toute base φ-orthogonale de E. Si $\dim \mathrm{E} = n$, la forme Q est définie positive si et seulement si elle est de signature $(n, 0)$, et elle est définie négative si et seulement si elle est de signature $(0, n)$.

Exemples. La forme hyperbolique sur $\mathbb{R}^4$ de l'exemple 4 est de signature $(3, 1)$. Le produit scalaire usuel sur $\mathbb{R}^n$ est de signature $(n, 0)$.

Formes bilinéaires positives. Inégalité de Cauchy-Schwarz

Soit E un espace vectoriel réel et soit φ une forme bilinéaire symétrique positive sur l'espace E ; soient $x, y \in$ E ; on a

$$\varphi(tx+y, tx+y) \geq 0,$$

pour tout nombre réel t. On a donc un trinôme réel qui ne change pas de signe quand t décrit $\mathbb{R}$,

$$t^2\varphi(x,x) + 2t\,\varphi(x,y) + \varphi(y,y),$$

donc son discriminant est ≤ 0, ce qui donne l'inégalité de Cauchy-Schwarz

$$\big(\varphi(x,y)\big)^2 \leq \varphi(x,x)\,\varphi(y,y).$$

Exemples.

1. Dans le cas du produit scalaire usuel sur $\mathbb{R}^n$, on a bien une forme bilinéaire positive puisque $\varphi(x,x) = \sum_{i=1}^n x_i^2 \geq 0$. Il en résulte que le produit scalaire de deux vecteurs est majoré par le produit des normes de ces vecteurs.

2. Considérons la forme bilinéaire φ définie sur l'espace des fonctions réelles continues sur $[0,1]$ par la formule

$$\varphi(f,g) = \int_0^1 f(t)g(t)\,dt$$

pour toutes fonctions réelles continues f et g. Il est clair que cette forme bilinéaire est positive, donc on peut lui appliquer l'inégalité de Cauchy-Schwarz. On obtient ainsi l'inégalité

$$\Big|\int_0^1 f(t)g(t)\,dt\Big| \leq \Big(\int_0^1 f(t)^2\,dt\Big)^{1/2}\Big(\int_0^1 g(t)^2\,dt\Big)^{1/2}.$$

Lemme. *Supposons que φ soit une forme bilinéaire sur un espace vectoriel réel* E. *Si φ est* **positive** *sur* E*, le noyau de φ coïncide avec l'ensemble des vecteurs isotropes.*

Démonstration. Dans tous les cas, le noyau d'une forme bilinéaire est contenu dans l'ensemble des vecteurs isotropes ; inversement, si φ est positive et si $\varphi(x,x) = 0$, on déduit de l'inégalité de Cauchy-Schwarz $\varphi(x,y)^2 \leq \varphi(x,x)\,\varphi(y,y)$ que $\varphi(x,y) = 0$ pour tout $y \in$ E, donc x appartient au noyau de E.

1.4. Un peu de géométrie. Division harmonique

Supposons donnée une droite affine Δ, et sur cette droite considérons quatre points A, B, C et D. Supposons ces points deux à deux distincts. On dit que le quadruplet $(\mathrm{A},\mathrm{B},\mathrm{C},\mathrm{D})$ forme une division harmonique si

$$\frac{\overline{\mathrm{CA}}}{\overline{\mathrm{CB}}} = -\frac{\overline{\mathrm{DA}}}{\overline{\mathrm{DB}}}.$$

En écrivant cette relation sous la forme

$$\overline{\mathrm{CA}}\;\overline{\mathrm{DB}} = -\overline{\mathrm{DA}}\;\overline{\mathrm{CB}}$$

on peut inclure un certain nombre de cas particuliers. Par exemple, si $\mathrm{C} = \mathrm{B}$, on doit prendre $\mathrm{D} = \mathrm{B}$ aussi, donc $(\mathrm{A},\mathrm{B},\mathrm{B},\mathrm{B})$ est un cas particulier de division harmonique.

Remarque. La situation est réversible : si $(\mathrm{A},\mathrm{B},\mathrm{C},\mathrm{D})$ est une division harmonique, il en est de même pour $(\mathrm{C},\mathrm{D},\mathrm{A},\mathrm{B})$.

Si on prend pour C le milieu de AB, on ne peut pas trouver de point D qui vérifie la relation de division harmonique. Cependant, si on considère un point C' différent du milieu mais qui s'approche du milieu de AB, on voit que le point D' correspondant s'éloigne à l'infini quand C' tend vers le milieu L. On dit alors que $(\mathrm{A}, \mathrm{B}, \mathrm{L}, \infty)$ forme encore une division harmonique. Cette notion de point à l'infini se formalise précisément avec la notion de droite projective. Supposons que Δ soit représentée comme droite affine du plan, droite ne passant pas par 0, d'équation

$$ux_1 + vx_2 = 1.$$

La *droite projective* $\mathbb{P}_1$ est l'ensemble des droites vectorielles du plan. On obtient une visualisation de la droite projective en associant à chaque $\mathrm{D} \in \mathbb{P}_1$ le point d'intersection x_{D} de D avec la droite Δ. En fait ce procédé ne s'applique pas à la droite exceptionnelle D_0 dont l'équation est $ux_1+vx_2 = 0$, qui est parallèle à Δ et ne coupe donc pas Δ. On considère que cette droite D_0 parallèle à Δ représente le "point à l'infini" sur Δ. Une droite vectorielle D est formée de points de la forme (td, td'), où t varie dans $\mathbb{R}$ et où (d, d') est un vecteur directeur pour D (il s'agit donc d'un vecteur non nul). On dit que (d, d') est un système de *coordonnées homogènes* pour le "point" D de l'espace $\mathbb{P}_1$. Les coordonnées homogènes ne sont pas uniques. Ainsi, $(1, 1)$, $(2, 2)$ ou $(-1, -1)$ représentent le même "point" de l'espace projectif. La notion de point à l'infini n'a pas de sens intrinsèque : elle dépend de la droite Δ choisie pour représenter $\mathbb{P}_1$.

Si on utilise les coordonnées homogènes, la relation de division harmonique devient

$$(c'a - a'c)(d'b - b'd) = -(d'a - a'd)(c'b - b'c),$$

où chacun des points $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}$ a des coordonnées homogènes, (a, a') pour A, (b, b') pour B, etc...

On suppose maintenant donnée une droite affine Δ dans un espace vectoriel réel E de dimension ≥ 2, et sur cette droite deux points distincts A et B. On considère deux autres points C et D de la droite, qui sont définis par

$$\overrightarrow{\mathrm{AC}} = \lambda\overrightarrow{\mathrm{AB}}, \quad \overrightarrow{\mathrm{AD}} = \mu\overrightarrow{\mathrm{AB}},$$

où λ et μ sont deux nombres réels. Dire que $(\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D})$ forme une division harmonique revient alors à dire que $-\lambda(\mu - 1) = \mu(\lambda - 1)$, soit encore

$$\lambda + \mu = 2\lambda\mu.$$

Les conditions suivantes sont équivalentes.

(a) Les points $(\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D})$ forment une division harmonique, c'est à dire $(\lambda+\mu)-2\lambda\mu = 0$.

(b) Pour tout point Ω, pour toute forme quadratique Q de forme polaire φ_{Q}, telle que $\mathrm{Q}(\overrightarrow{\Omega\mathrm{A}}) = \mathrm{Q}(\overrightarrow{\Omega\mathrm{B}}) = 0$, on a $\varphi_{\mathrm{Q}}(\overrightarrow{\Omega\mathrm{C}}, \overrightarrow{\Omega\mathrm{D}}) = 0$.

(c) Pour tout point Ω hors de Δ, il existe une forme quadratique Q de forme polaire φ_{Q}, non identiquement nulle sur le plan vectoriel $[\overrightarrow{\Omega\mathrm{A}}, \overrightarrow{\Omega\mathrm{B}}]$, et telle que $\mathrm{Q}(\overrightarrow{\Omega\mathrm{A}}) = \mathrm{Q}(\overrightarrow{\Omega\mathrm{B}}) = 0$, et $\varphi_{\mathrm{Q}}(\overrightarrow{\Omega\mathrm{C}}, \overrightarrow{\Omega\mathrm{D}}) = 0$.

(d) Il existe un point Ω hors de Δ, une forme quadratique Q de forme polaire φ_{Q}, non identiquement nulle sur le plan vectoriel $[\overrightarrow{\Omega\mathrm{A}}, \overrightarrow{\Omega\mathrm{B}}]$, et telle que $\mathrm{Q}(\overrightarrow{\Omega\mathrm{A}}) = \mathrm{Q}(\overrightarrow{\Omega\mathrm{B}}) = 0$, et $\varphi_{\mathrm{Q}}(\overrightarrow{\Omega\mathrm{C}}, \overrightarrow{\Omega\mathrm{D}}) = 0$.

Etant donné un point Ω, posons $x = \overrightarrow{\Omega\mathrm{A}}$ et $y = \overrightarrow{\Omega\mathrm{B}}$. On a alors $\overrightarrow{\mathrm{AB}} = y - x$. Pour tout $t \in \mathbb{R}$, définissons le point $\mathrm{M}(t)$ de la droite Δ par la relation $\overrightarrow{\mathrm{AM}(t)} = t(y - x)$, et posons $z(t) = \overrightarrow{\Omega\mathrm{M}(t)}$. On a donc $z(t) = x + t(y - x) = (1 - t)x + ty$. Notons que $\mathrm{C} = \mathrm{M}(\lambda)$, $\overrightarrow{\Omega\mathrm{C}} = z(\lambda)$, $\mathrm{D} = \mathrm{M}(\mu)$, $\overrightarrow{\Omega\mathrm{D}} = z(\mu)$, $x = z(0)$ et $y = z(1)$. Remarquons qu'une forme quadratique Q telle que $\mathrm{Q}(x) = \mathrm{Q}(y) = 0$, de forme polaire φ, est nulle sur le plan $[x, y]$ si on ajoute la condition $\varphi(x, y) = 0$ (en effet, sa matrice dans la base (x, y) du plan est alors nulle).

Partons de la première hypothèse. Supposons que $\lambda + \mu = 2\lambda\mu$, et soit Q une forme quadratique vérifiant les conditions de (b) ; on suppose donc que $\mathrm{Q}(x) = \mathrm{Q}(y) = 0$. On aura

$$\varphi(z(\lambda), z(\mu)) = \varphi((1 - \lambda)x + \lambda y, (1 - \mu)x + \mu y) = ((1 - \lambda)\mu + \lambda(1 - \mu))\varphi(x, y) = 0,$$

ce qui montre (b).

Supposons maintenant que la propriété (b) soit satisfaite. Soit Ω un point hors de Δ ; il faut trouver une forme quadratique Q qui s'annule pour x et y, non identiquement nulle sur le plan vectoriel $[x, y]$. Posons $e_1 = \frac{1}{2}\overrightarrow{\mathrm{AB}}$. Soit L le milieu de AB et définissons e_2 par $\overrightarrow{\Omega\mathrm{L}} = e_2$. Définissons Q égale à $x_1^2 - x_2^2$, où x_1, x_2 sont les deux premières coordonnées dans une base de E commençant par e_1, e_2. On a alors $\mathrm{Q}(\overrightarrow{\Omega\mathrm{A}}) = \mathrm{Q}(e_2 - e_1) = 0$, et de même pour le point B, on a $\mathrm{Q}(\overrightarrow{\Omega\mathrm{B}}) = \mathrm{Q}(e_1 + e_2) = 0$. D'après l'hypothèse (b), on en déduit que $\varphi(\overrightarrow{\Omega\mathrm{C}}, \overrightarrow{\Omega\mathrm{D}}) = 0$. Par ailleurs $\mathrm{Q}(\overrightarrow{\Omega\mathrm{L}}) = -1 \neq 0$, donc on a bien montré la propriété (c).

L'implication $(c) \Rightarrow (d)$ est évidente. Supposons pour finir que (d) soit satisfaite. On a alors $\varphi(x, y) \neq 0$ puisque Q n'est pas nulle sur le plan $[x, y]$. De plus

$$\varphi_{\mathrm{Q}}(z(s), z(t)) = (s(1-t) + t(1-s))\varphi(x, y),$$

et puisque notre dernière hypothèse est $\varphi(z(\lambda), z(\mu)) = 0$, on obtient bien la relation $\lambda(1-\mu) + \mu(1-\lambda) = 0$.

Conséquence. Faisceau de droites passant par des points en division harmonique

On suppose que Δ est une droite du plan, et que sur cette droite quatre points $(\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D})$ forment une division harmonique. Par un point Ω hors de Δ on trace quatre droites $\Delta_\alpha, \Delta_\beta, \Delta_\gamma, \Delta_\delta$ qui coupent Δ en $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}$ et une autre droite Δ_1 en $\mathrm{A}_1, \mathrm{B}_1, \mathrm{C}_1, \mathrm{D}_1$. Ces quatre nouveaux points forment une division harmonique.

En effet, puisque Ω est hors de Δ il existe une forme quadratique non triviale qui vérifie les conditions de (c) pour les points $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}$. Les vecteurs correspondants $\overrightarrow{\Omega\mathrm{A}_1}$, $\overrightarrow{\Omega\mathrm{B}_1}$, $\overrightarrow{\Omega\mathrm{C}_1}$ et $\overrightarrow{\Omega\mathrm{D}_1}$ étant des multiples des vecteurs $\overrightarrow{\Omega\mathrm{A}}$, $\overrightarrow{\Omega\mathrm{B}}$, $\overrightarrow{\Omega\mathrm{C}}$ et $\overrightarrow{\Omega\mathrm{D}}$, les bonnes propriétés de Q sur $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}$ se transmettent à $\mathrm{A}_1, \mathrm{B}_1, \mathrm{C}_1, \mathrm{D}_1$: on aura donc $\mathrm{Q}(\overrightarrow{\Omega\mathrm{A}_1}) = 0$, $\mathrm{Q}(\overrightarrow{\Omega\mathrm{B}_1}) = 0$, $\varphi(\overrightarrow{\Omega\mathrm{C}_1}, \overrightarrow{\Omega\mathrm{D}_1}) = 0$, ce qui montre que $(\mathrm{A}_1, \mathrm{B}_1, \mathrm{C}_1, \mathrm{D}_1)$ est une division harmonique.

On appelle *faisceau harmonique de droites* un ensemble de quatre droites du plan avec les propriétés de $\Delta_\alpha, \Delta_\beta, \Delta_\gamma, \Delta_\delta$. Pour toute droite affine Δ ne passant pas par le point d'intersection Ω des quatre droites du faisceau, les quatre points d'intersection de Δ avec les droites du faisceau forment une division harmonique.

Coniques

Considérons une ellipse dans le plan, d'équation

$$ax_1^2 + bx_2^2 = 1,$$

avec $a, b > 0$. On place une “copie” de ce plan dans $\mathbb{R}^3$, à l'altitude $x_3 = 1$. Désignons par Π le plan de $\mathbb{R}^3$ d'équation $x_3 = 1$. La conique Γ du plan Π est l'intersection de Π avec l'ensemble (C) de $\mathbb{R}^3$ défini par l'équation

$$ax_1^2 + bx_2^2 - x_3^2 = 0,$$

c'est à dire l'ensemble des vecteurs isotropes d'une forme quadratique de signature $(2, 1)$ sur $\mathbb{R}^3$ donnée par

$$\mathrm{Q}(x) = ax_1^2 + bx_2^2 - x_3^2 = 0.$$

La forme polaire est alors égale à

$$\varphi(x, y) = a\, x_1 y_1 + b\, x_2 y_2 - x_3 y_3.$$

Expliquons le point de vue projectif. Le *plan projectif* est l'ensemble des droites vectorielles de $\mathbb{R}^3$. L'ensemble (C) s'identifie à un ensemble de droites passant par 0, c'est à dire à un ensemble de points du plan projectif. C'est une *conique du plan projectif*. Si on coupe (C) par d'autres plans que Π (plans affines qui ne passent pas par 0), on obtient des visualisations différentes de la conique du plan projectif. On peut obtenir ainsi une section plane de (C) qui soit une hyperbole ou une parabole. Par exemple, si on coupe (C) par le plan $x_2 = 1$, la section, projetée sur le plan (x_1, x_3) a pour équation

$$ax_1^2 - x_3^2 = -b,$$

qui est une hyperbole (exercice : trouver un plan qui coupe (C) suivant une parabole).

Conjugaison par rapport à Γ et divisions harmoniques. On dit que deux points M et P du plan Π sont conjugués par rapport à (Γ) si les vecteurs $\overrightarrow{\mathrm{OM}}$ et $\overrightarrow{\mathrm{OP}}$ sont φ-orthogonaux. Un point A de Γ est conjugué de lui-même. L'ensemble des conjugués d'un point $\mathrm{M} \in \Pi$ est l'intersection de Π avec le plan vectoriel $\overrightarrow{\mathrm{OM}}^{\perp}$. C'est donc une droite de Π, appelée *droite polaire* du point M. Supposons que M ne soit pas sur (Γ), que Δ soit une droite passant par M qui recoupe la polaire de M en P, et l'ellipse en A et B. Les points $(\mathrm{A}, \mathrm{B}, \mathrm{M}, \mathrm{P})$ forment une division harmonique. En effet, $\mathrm{Q}(\overrightarrow{\mathrm{OA}}) = \mathrm{Q}(\overrightarrow{\mathrm{OB}}) = 0$ puisque A et B sont deux points de (Γ), et $\varphi(\overrightarrow{\mathrm{OM}}, \overrightarrow{\mathrm{OP}}) = 0$ puisque P appartient à la droite polaire de M.

Exercices.

1. Dans le cas où $\mathrm{M} = \mathrm{A}$ est un point de (Γ), montrer que la droite polaire de M passe par A et ne passe par aucun autre point de (Γ). C'est donc la tangente en A à l'ellipse.
2. Si d'un point M hors de (Γ) on peut tracer deux tangentes à l'ellipse qui touchent l'ellipse en A et B, montrer que la polaire de M est la droite qui passe par A et B.
3. Si on trace des cordes parallèles à une droite Δ du plan Π, montrer que l'ensemble des milieux est situé sur une droite, qui est la polaire du point à l'infini correspondant à la droite vectorielle de même direction que Δ.

Chapitre 2. Espaces euclidiens

2.1. Produit scalaire et normes euclidiennes

Définition 2.1.1. Un *espace euclidien* est un espace vectoriel réel E **de dimension finie** muni d'une forme bilinéaire symétrique φ définie positive.

On dit que φ définit le *produit scalaire* de l'espace euclidien E, et on notera en général le produit scalaire de deux vecteurs x et y de E par

$$x \cdot y = \varphi(x, y).$$

En termes précis, un espace euclidien n'est pas seulement un espace vectoriel, mais un couple (E, φ) d'un espace vectoriel E et d'une forme bilinéaire φ définie sur $\mathrm{E} \times \mathrm{E}$.

Il peut être instructif de remarquer qu'un certain nombre des propriétés qui seront vues dans ce chapitre ne dépend pas du fait que l'espace vectoriel E soit de dimension finie. On appelle *espace préhilbertien* un espace vectoriel réel E muni d'une forme bilinéaire symétrique φ définie positive.

Exemples.

1. Dans la quasi-totalité de nos exemples, l'espace vectoriel E sera de dimension finie, mais on peut garder un œil un peu ouvert sur d'autres cas, par exemple celui de l'espace $\mathrm{E} = \mathrm{C}([0, 1])$ des fonctions continues sur $[0, 1]$ muni de la forme bilinéaire φ définie pour tous $f, g \in \mathrm{E}$ par

$$\varphi(f, g) = \int_0^1 f(t)g(t)\, dt.$$

Le couple (E, φ) est un espace préhilbertien. En effet, il est clair que φ est bilinéaire, symétrique et positive. Le fait qu'elle soit définie positive provient du résultat suivant : si l'intégrale d'une fonction ≥ 0 continue est nulle, la fonction est nulle.

2. L'exemple le plus commun sera $\mathbb{R}^n$ avec son produit scalaire usuel, défini par

$$\varphi(x, y) = x \cdot y = x_1 y_1 + \cdots + x_n y_n$$

pour deux vecteurs quelconques $x = (x_1, \ldots, x_n)$ et $y = (y_1, \ldots, y_n)$ de $\mathbb{R}^n$.

3. Espace ℓ_2 : c'est l'espace vectoriel (de dimension infinie) formé des suites réelles $(x_n)_{n=0}^{\infty}$ de carré sommable, c'est à dire telles que la série numérique $\sum x_n^2$ converge ; on pose alors

$$x \cdot y = \sum_{n=0}^{\infty} x_n\, y_n.$$

C'est encore un espace préhilbertien. C'est l'extension “naturelle” de l'exemple 2 à la dimension infinie.

4. Si E est un espace euclidien et si F est un sous-espace vectoriel de E, la restriction du produit scalaire de E aux couples de vecteurs de F permet de munir F d'une structure d'espace euclidien. Cette remarque évidente sera appliquée à plusieurs reprises dans ce chapitre.

On rappelle l'inégalité de Cauchy-Schwarz pour les formes bilinéaires symétriques positives,

$$\varphi(x, y)^2 \leq \varphi(x, x)\, \varphi(y, y).$$

Pour un espace euclidien ou préhilbertien E fixé, on notera $\|x\| = \sqrt{x \,.\, x}$ et l'inégalité de Cauchy-Schwarz donne dans ce cas

$$\bigl|x \,.\, y\bigr| \le \|x\| \, \|y\|.$$

Lemme. *Soit* E *un espace préhilbertien ; l'application* $x \to \|x\|$ *est une norme sur* E.

Démonstration. On voit que $\|x\| = 0$ signifie que $x \,.\, x = \varphi(x, x) = 0$, donc $x = 0_{\mathrm{E}}$ puisque φ est définie positive. A partir de l'inégalité de Cauchy-Schwarz, on voit facilement que

$$\|x + y\| \le \|x\| + \|y\|$$

(inégalité triangulaire) ; en effet, on obtient en développant le carré scalaire de $x + y$ et en utilisant la symétrie $(x \,.\, y = y \,.\, x)$

$$\|x + y\|^2 = (x + y) \,.\, (x + y) = x \,.\, x + 2\, x \,.\, y + y \,.\, y \le$$
$$x \,.\, x + 2 \sqrt{x \,.\, x} \sqrt{y \,.\, y} + y \,.\, y = (\sqrt{x \,.\, x} + \sqrt{y \,.\, y})^2 = (\|x\| + \|y\|)^2.$$

Identité du parallélogramme

Pour deux vecteurs quelconques x, y de l'espace E, on a obtenu ci-dessus la relation $\|x+y\|^2 = \|x\|^2 + 2\, x \,.\, y + \|y\|^2$. En additionnant avec la relation analogue obtenue pour les vecteurs x et $-y$, on obtient *l'identité du parallélogramme :*

Etant donnés deux vecteurs x, y d'un espace euclidien (ou préhilbertien), on a

$$\Bigl\|\frac{x + y}{2}\Bigr\|^2 + \Bigl\|\frac{x - y}{2}\Bigr\|^2 = \frac{\|x\|^2 + \|y\|^2}{2}.$$

Orthogonalité dans un espace euclidien

On dit que deux vecteurs x et y de E sont **orthogonaux** si $x \,.\, y = 0$. Cette relation est symétrique. On note $x \perp y$. On dit que deux sous-espaces vectoriels E_1 et E_2 sont orthogonaux si on a $x_1 \perp x_2$ pour tous $x_1 \in \mathrm{E}_1$, $x_2 \in \mathrm{E}_2$.

Si $x \perp y$, la relation ci-dessus donne

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2,$$

qu'il est plus ou moins raisonnable d'appeler le *théorème de Pythagore*.

Proposition 2.1.1. Relation de Pythagore. *Si $x_1, \ldots, x_k$ sont deux à deux orthogonaux, on a*

$$\|x_1 + \cdots + x_k\|^2 = \sum_{i=1}^{k} \|x_i\|^2.$$

Démonstration. On a vérifié d'abord le cas $k = 2$, et on peut ensuite faire une démonstration par récurrence sur $k \ge 2$. On remarquera que $(x_1 + \cdots + x_{k-1})$ est orthogonal à x_k, ce qui fournit le résultat en utilisant l'hypothèse de récurrence et le cas de deux vecteurs,

$$\|(x_1 + \cdots + x_{k-1}) + x_k\|^2 = \|x_1 + \cdots + x_{k-1}\|^2 + \|x_k\|^2 = \cdots$$

Corollaire 2.1.1. *Si des vecteurs d'un espace euclidien ou préhilbertien sont deux à deux orthogonaux et* **non nuls**, *ils sont linéairement indépendants.*

Corollaire 2.1.2. *Soit* E *un espace euclidien ou préhilbertien ; si les sous-espaces vectoriels* $\mathrm{E}_1, \mathrm{E}_2, \ldots, \mathrm{E}_k$ *de* E *sont deux à deux orthogonaux, ils forment une somme directe* $\mathrm{E}_1 \oplus \cdots \oplus \mathrm{E}_k$ *dans* E.

Base orthonormée

Soit E un espace euclidien ; on dit qu'un système $(x_1, \ldots, x_n)$ de vecteurs de E est une **base orthonormée** de E si c'est une base de E et si les vecteurs (x_i) sont deux à deux orthogonaux et de norme 1,

$$\forall i \neq j, \quad x_i \,.\, x_j = 0\,; \quad \forall i = 1, \ldots, n, \;\; \|x_i\| = 1.$$

Théorème 2.1.1. *Tout espace euclidien admet une base orthonormée.*

Démonstration. On a vu dans le chapitre 1 que pour toute forme bilinéaire symétrique φ sur un espace E de dimension finie sur le corps des réels, il existe une base $(e_1, \ldots, e_n)$ de E formée de vecteurs deux à deux φ-orthogonaux telle que $\varphi(e_i, e_i)$ soit égal à 1, -1 ou 0. Ici, puisque φ est définie positive, le seul cas possible est $\varphi(e_i, e_i) = 1$, ce qui termine la démonstration. Mais plutôt que de renvoyer ainsi à la lecture du chapitre 1, il est peut être plus agréable pour le lecteur de montrer directement l'existence d'une base orthonormée pour E en reprenant rapidement la démonstration de la partie du résultat du chapitre 1 dont nous avons besoin ici : le résultat est évident si $\dim \mathrm{E} = 1$. Soit $n > 1$, et supposons le résultat démontré pour tout espace euclidien F de dimension $< n$; soit E un espace euclidien de dimension n ; on choisit un vecteur e_1 de norme 1 dans E. L'ensemble des vecteurs $y \in \mathrm{E}$ tels que $y \perp e_1$ forme un espace euclidien F de dimension $n - 1$, qui admet donc une base orthonormée $(e_2, \ldots, e_n)$ d'après l'hypothèse de récurrence. Alors $(e_1, e_2, \ldots, e_n)$ est une base orthonormée de E.

La méthode de Gram-Schmidt qui est présentée plus loin donnera une variante de démonstration pour le théorème 2.1.1.

Coordonnées dans une base orthonormée

Soit E un espace euclidien, et calculons les coordonnées d'un vecteur x de E dans une base orthonormée $(e_1, \ldots, e_n)$. Ecrivons $x = \sum_{i=1}^n c_i e_i$. En calculant le produit scalaire $x \,.\, e_j$, on obtient $c_j = x \,.\, e_j$. On voit donc que :

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

D'après Pythagore on aura $\|x\|^2 = \sum_{i=1}^n \|c_i e_i\|^2 = \sum_{i=1}^n c_i^2$. On en déduit le calcul de la norme du vecteur x à partir des coordonnées dans une base orthonormée,

$$\|x\|^2 = \sum_{i=1}^{n} (x \,.\, e_i)^2.$$

Ces calculs de coordonnées s'appliquent au calcul de la matrice d'un endomorphisme $u \in \mathcal{L}(\mathrm{E})$ dans une base orthonormée $(e_1, \ldots, e_n)$ de E : la matrice $\mathrm{A} = (a_{i,j})$ admet pour coefficients de la colonne j les coordonnées du vecteur image $u(e_j)$, par conséquent

$$a_{i,j} = u(e_j) \,.\, e_i.$$

Projection orthogonale

Soit E un espace euclidien (ou préhilbertien) ; si x est un vecteur de E et F un sous-espace vectoriel de E, on dit qu'un vecteur y est la *projection orthogonale* de x sur F si

$$y \in \mathrm{F} \quad \text{et} \quad x - y \perp \mathrm{F}.$$

Le vecteur y est uniquement déterminé par ces deux conditions : si y_1 et y_2 vérifient ces conditions, alors $x - y_1$ et $x - y_2$ sont orthogonaux à F, donc la différence $y_1 - y_2$ est orthogonale à F. Mais puisque $y_1 - y_2 \in \mathrm{F}$, on peut écrire $(y_1 - y_2) \,.\, (y_1 - y_2) = 0$, ce qui implique que $y_1 = y_2$.

Lorsque y existe, on note $y = \mathrm{P_F}(x)$. Si le sous-espace F est de dimension finie, la projection $\mathrm{P_F}(x)$ existe toujours : si on choisit une base orthonormée $(e_1, \ldots, e_k)$ de F, on peut décrire l'application $\mathrm{P_F}$ de projection orthogonale de E sur F en posant pour tout $x \in \mathrm{E}$

$$\mathrm{P_F}(x) = \sum_{i=1}^{k} (x \,.\, e_i)\, e_i \in \mathrm{F}.$$

On voit de plus par cette formule que $\mathrm{P_F}$ est une application linéaire ; il est clair que pour tout vecteur $y \in \mathrm{F}$ on a $\mathrm{P_F}(y) = y$, ce qui entraîne que $\mathrm{P_F}(\mathrm{P_F}(x)) = \mathrm{P_F}(x)$ pour tout $x \in \mathrm{E}$, c'est à dire que $(\mathrm{P_F})^2 = \mathrm{P_F}$ (l'application linéaire $\mathrm{P_F}$ est un *projecteur*).

Si z est un vecteur quelconque de F, on pourra écrire $x - z = (x - \mathrm{P_F}(x)) + (\mathrm{P_F}(x) - z)$ et puisque $\mathrm{P_F}(x) - z \in \mathrm{F}$ on a par Pythagore

$$\|x - z\|^2 = \|x - \mathrm{P_F}(x)\|^2 + \|\mathrm{P_F}(x) - z\|^2.$$

Cette relation montre que $\|x - z\| \geq \|x - \mathrm{P_F}(x)\|$ pour tout $z \in \mathrm{F}$, donc $\mathrm{P_F}(x)$ est le point de F qui est le plus proche de x.

Exercice. Montrer que cette propriété de plus courte distance caractérise $\mathrm{P_F}(x)$.

Si F est de dimension infinie, il est possible que la projection de $x \in \mathrm{E}$ sur F n'existe pas, mais elle existe toujours lorsque F est *complet* pour la distance $d(x, y) = \|x - y\|$.

En prenant $z = 0_\mathrm{E}$ dans la relation précédente, on obtient aussi $\|x\|^2 \geq \|\mathrm{P_F}(x)\|^2$, ce qui donne par linéarité $\|\mathrm{P_F}(x_1) - \mathrm{P_F}(x_2)\| \leq \|x_1 - x_2\|$ pour tous vecteurs $x_1, x_2 \in \mathrm{E}$: l'application $\mathrm{P_F}$ diminue les distances.

On définit l'orthogonal $\mathrm{F}^\perp$ d'un sous-espace vectoriel F de E par

$$\mathrm{F}^\perp = \{x \in \mathrm{E} : x \perp \mathrm{F}\}.$$

On voit que $\mathrm{F}^\perp$ est un sous-espace vectoriel de E.

Proposition 2.1.2. *Soit* E *un espace euclidien (ou préhilbertien) et soit* F *un sous-espace vectoriel de dimension finie de* E *; on a*

$$\mathrm{E} = \mathrm{F} \oplus \mathrm{F}^\perp \quad \textit{et} \quad (\mathrm{F}^\perp)^\perp = \mathrm{F}.$$

Démonstration. Tout vecteur $x \in \mathrm{E}$ peut s'écrire $x = \mathrm{P_F}(x) + (x - \mathrm{P_F}x)$, avec $\mathrm{P_F}(x) \in \mathrm{F}$ et $x - \mathrm{P_F}(x) \in \mathrm{F}^\perp$, ce qui montre que $\mathrm{E} = \mathrm{F} + \mathrm{F}^\perp$; de plus, il est clair que $\mathrm{F} \cap \mathrm{F}^\perp = \{0\}$, donc la somme est directe. Montrons la deuxième propriété de l'énoncé. On a toujours

$\mathrm{F} \subset \mathrm{F}^{\perp\perp}$. Inversement, si $x \in \mathrm{F}^{\perp\perp}$, on écrit $x = f + g$, avec $f \in \mathrm{F}$ et $g \perp \mathrm{F}$. Puisque x est orthogonal à $\mathrm{F}^{\perp}$, on aura $0 = x \,.\, g = f \,.\, g + g \,.\, g = g \,.\, g$, donc $g = 0_{\mathrm{E}}$ et $x = f \in \mathrm{F}$.

Si E est un espace euclidien et si F est un sous-espace différent de E, son orthogonal $\mathrm{F}^{\perp}$ est donc $\neq \{0\}$.

Ce résultat n'est pas vrai en toute généralité dans le cas préhilbertien de dimension infinie. Sous certaines conditions topologiques qui sont largement hors programme, on peut étendre la proposition 2.1.2 : si E est préhilbertien *complet* (on dit alors que E est un *espace de Hilbert*) et si F est un sous-espace vectoriel *fermé* de E, la décomposition $\mathrm{E} = \mathrm{F} \oplus \mathrm{F}^{\perp}$ reste valable.

Corollaire 2.1.3. *Soient* E *un espace euclidien et* F *un sous-espace vectoriel de* E *; on a*

$$\mathrm{E} = \mathrm{F} \oplus \mathrm{F}^{\perp} \quad \textit{et} \quad (\mathrm{F}^{\perp})^{\perp} = \mathrm{F}.$$

Procédé d'orthonormalisation de Gram-Schmidt

Donnons d'abord le principe général : soient E un espace euclidien, F un sous-espace vectoriel de E dont on connaît une base orthonormée $(e_1, \ldots, e_k)$, et soit $x \notin \mathrm{F}$; alors le vecteur $y = x - \mathrm{P}_{\mathrm{F}}(x)$ est non nul, orthogonal à F et $e_{k+1} = y/\|y\|$ permet d'allonger notre système orthonormé.

On peut utiliser ce principe dans le cas où une base de E est déjà donnée (mais pas orthogonale), pour construire une nouvelle base de E qui soit orthonormée. On suppose donc donnée une base quelconque $(f_1, f_2, \ldots, f_n)$ de E, et on va construire à partir de cette base une base orthonormée $(e_1, \ldots, e_n)$; la construction se fera de façon que pour chaque $k = 1, \ldots, n$ on ait

$$\mathrm{F}_k = [f_1, \ldots, f_k] = [e_1, \ldots, e_k].$$

On commence en posant $e_1 = f_1/\|f_1\|$. Si les vecteurs $e_1, \ldots, e_{k-1}$ sont déjà déterminés, avec $\mathrm{F}_{k-1} = [e_1, \ldots, e_{k-1}]$, on considère $u_k = f_k - \mathrm{P}_{\mathrm{F}_{k-1}}(f_k)$. Puisque $\mathbf{f}$ est une base, on sait que $f_k \notin \mathrm{F}_{k-1}$, donc $u_k \neq 0_{\mathrm{E}}$, et on peut poser $e_k = u_k/\|u_k\|$. On vérifie facilement que $[e_1, \ldots, e_k] = [f_1, \ldots, f_k]$. Quand on arrive à $k = n$ on a $[e_1, \ldots, e_n] = [f_1, \ldots, f_n]$, ce qui montre que $(e_1, \ldots, e_n)$ est une base de E. Elle est orthonormée par construction.

Si on part d'une base orthonormée $(e_1, \ldots, e_k)$ d'un sous-espace F, complétée par une famille $(f_{k+1}, \ldots, f_n)$ pour former une base de E, on peut appliquer la méthode à partir de l'étape qui définit e_{k+1} pour trouver une famille $(e_{k+1}, \ldots, e_n)$ qui complète le système $(e_1, \ldots, e_k)$ en une base orthonormée de E.

Le procédé peut aussi s'appliquer en dimension infinie, quand on dispose au départ d'une suite infinie (f_n) de vecteurs d'un espace préhilbertien E, telle que $(f_1, \ldots, f_n)$ soit libre pour tout entier n : on produira dans ce cas une suite orthonormée infinie (e_n) dans l'espace préhilbertien.

Exemple. Polynômes orthogonaux sur $[-1, +1]$.

Considérons l'espace préhilbertien $\mathrm{E} = \mathrm{C}([-1, 1])$ des fonctions continues sur $[-1, 1]$, muni du produit scalaire $f \,.\, g = \int_{-1}^{1} f(t)g(t)\,dt$; on part de la suite (f_n) des fonctions polynomiales $1, x, \ldots, x^n, \ldots$ Les termes successifs produits par le procédé de Gram-Schmidt fournissent des polynômes, dont les premiers sont

$$\frac{1}{\sqrt{2}},\ \sqrt{\frac{3}{2}}\,\mathrm{X},\ \sqrt{\frac{5}{8}}\ (3\mathrm{X}^2 - 1),\ \sqrt{\frac{7}{8}}\,(5\mathrm{X}^3 - 3\mathrm{X}), \ldots$$

et les calculs de la méthode de Gram-Schmidt deviennent très vite terrrifiants.

Il est plus facile de trouver ces polynômes orthogonaux en faisant un peu plus de mathématiques… On montre à cet effet que les polynômes

$$\mathrm{P}_n = \frac{d^n}{dx^n}\,(\mathrm{X}^2-1)^n$$

sont deux à deux orthogonaux, et on vérifie que $\deg \mathrm{P}_n = n$ pour tout entier $n \ge 0$. Les polynômes cherchés sont alors simplement des multiples convenables des P_n.

Pour montrer que les (P_n) sont orthogonaux, on remarquera que les dérivées d'ordre $k < n$ de $(\mathrm{X}^2-1)^n$ s'annulent aux points -1 et 1. On utilisera cette remarque, jointe à des intégrations par parties convenables, pour montrer que P_n est orthogonal à P_m lorsque $n < m$ (on dérive P_n jusqu'à ce que la dérivée $(n+1)$ième s'annule, et on intègre P_m sous la forme $\frac{d^{m-k}}{dx^{m-k}}\,(\mathrm{X}^2-1)^m$, avec k variant de 1 à $n < m$).

Exercice : polynômes de Hermite. Montrer que la formule

$$\mathrm{e}^{tx-t^2/2} = \sum_{n=0}^{+\infty} \frac{t^n}{n!}\,\mathrm{H}_n(x)$$

pour tous x, t réels, définit une suite (H_n) de polynômes ; vérifier que $\deg \mathrm{H}_n = n$ pour tout $n \ge 0$, et déterminer H_n pour $n \le 3$. Montrer que les fonctions $f_n(x) = \mathrm{H}_n(x)\,\mathrm{e}^{-x^2/4}$ sont deux à deux orthogonales pour le produit scalaire $\int_{-\infty}^{+\infty} f(x)g(x)\,dx$.

Dual d'un espace euclidien

Théorème 2.1.2. *Soit* E *un espace euclidien ; pour toute forme linéaire* ℓ *sur* E *il existe un vecteur* $y \in \mathrm{E}$ *(unique) tel que*

$$\forall x \in \mathrm{E}, \quad \ell(x) = x\,.\,y.$$

Démonstration. Si $\ell = 0$ il suffit de prendre $y = 0$. Sinon soit $x_0 \in \mathrm{E}$ tel que $\ell(x_0) = 1$ et soit $\mathrm{F} = \ker \ell$; soient $y_0 = \mathrm{P_F}(x_0)$ et $z_0 = x_0 - y_0$; on a $\ell(z_0) = 1$, et de plus $z_0 \perp \mathrm{F}$. Soit $x \in \mathrm{E}$; on peut écrire $x = (x - \ell(x)z_0) + \ell(x)z_0$. On vérifie que $x - \ell(x)z_0 \in \mathrm{F}$, donc $z_0 \perp (x - \ell(x)z_0)$ et $x\,.\,z_0 = \ell(x)(z_0\,.\,z_0)$ pour tout $x \in \mathrm{E}$, donc en posant $y = z_0/(z_0\,.\,z_0)$ on obtient bien

$$\forall x \in \mathrm{E}, \quad \ell(x) = x\,.\,y.$$

Autre démonstration. On définit une application linéaire i_E de E dans E^* en posant pour tout $y \in \mathrm{E}$,

$$i_\mathrm{E}(y)(x) = x\,.\,y \quad \text{pour tout } x \in \mathrm{E}.$$

Si $i_\mathrm{E}(y) = 0_{\mathrm{E}^*}$, on a en particulier $0 = i_\mathrm{E}(y)(y) = y\,.\,y$, donc $y = 0_\mathrm{E}$. L'application i_E est donc injective, donc surjective puisque $\dim \mathrm{E} = \dim \mathrm{E}^*$.

On notera que si $\ell = i_\mathrm{E}(y)$, alors

$$\forall x \in \mathrm{E}, \quad x\,.\,i_\mathrm{E}^{-1}(\ell) = \ell(x).$$

2.2. Endomorphismes des espaces euclidiens

Adjointe d'une application linéaire

Proposition 2.2.1. *Soient* E *et* F *deux espaces euclidiens ; pour toute application linéaire* $u \in \mathcal{L}(\mathrm{E}, \mathrm{F})$*, il existe une application linéaire unique* $u^* \in \mathcal{L}(\mathrm{F}, \mathrm{E})$ *telle que*

$$u(x) \,.\, y = x \,.\, u^*(y) \ \ \textit{pour tous } x \in \mathrm{E},\ y \in \mathrm{F}.$$

Démonstration. Montrons d'abord l'unicité de u^*. Choisissons une base orthonormée $\mathbf{e} = (e_1, \ldots, e_n)$ de E, soit $\mathbf{f} = (f_1, \ldots, f_m)$ une base orthonormée de F et soit $\mathrm{A} = (a_{i,j}) = \mathrm{mat}(u, \mathbf{e}, \mathbf{f})$; supposons que v soit une application linéaire de F dans E telle que $u(x) \,.\, y = x \,.\, v(y)$ pour tous $x \in \mathrm{E}$ et $y \in \mathrm{F}$. On aura pour la matrice $\mathrm{B} = \mathrm{mat}(v, \mathbf{f}, \mathbf{e})$ de l'application linéaire v

$$b_{i,j} = v(f_j) \,.\, e_i = e_i \,.\, v(f_j) = u(e_i) \,.\, f_j = a_{j,i},$$

donc la matrice B est complètement déterminée, et l'endomorphisme v aussi. Inversement, posons $a^*_{i,j} = a_{j,i}$ pour tous $i = 1, \ldots, n$ et $j = 1, \ldots, m$. On définit ainsi une matrice A^*. Si on définit u^* par

$$\forall j = 1, \ldots, m, \quad u^*(f_j) = \sum_{i=1}^{n} a^*_{i,j} e_i$$

on vérifie que u^* convient.

Autre démonstration. On utilise la transposée ${}^t u : \mathrm{F}^* \to \mathrm{E}^*$ et l'identification du dual donnée par le théorème 2.1.2, qui donne deux isomorphismes $i_\mathrm{E} : \mathrm{E} \to \mathrm{E}^*$ et $i_\mathrm{F} : \mathrm{F} \to \mathrm{F}^*$. On pose alors $u^* = (i_\mathrm{E})^{-1} \circ {}^t u \circ i_\mathrm{F}$. En effet, avec cette définition

$$x \,.\, u^*(y) = x \,.\, i_\mathrm{E}^{-1}({}^t u(i_\mathrm{F}(y))) = {}^t u(i_\mathrm{F}(y))(x) = i_\mathrm{F}(y)(u(x)) = u(x) \,.\, y$$

pour tous $x \in \mathrm{E}$ et $y \in \mathrm{F}$.

Matrice de l'adjoint par rapport à deux bases orthonormées

Supposons donnée une base orthonormée $\mathbf{e}$ de E et une base orthonormée $\mathbf{f}$ de F. Soit $\mathrm{A} = (a_{i,j}) = \mathrm{mat}(u, \mathbf{e}, \mathbf{f})$; on vient de voir dans la démonstration précédente que la matrice $\mathrm{B} = \mathrm{mat}(u^*, \mathbf{f}, \mathbf{e})$ de u^* est égale à la matrice transposée de A, c'est à dire que pour tous $i = 1, \ldots, n$ et $j = 1, \ldots, m$ on a

$$b_{i,j} = a_{j,i}.$$

Ce qui a été dit s'applique en particulier aux endomorphismes d'un espace euclidien E (lorsque F = E). Dans ce cas, pour tout endomorphisme $u \in \mathcal{L}(\mathrm{E})$, l'application u^* est aussi un endomorphisme de E. Quand E = F, on choisit en général de calculer la matrice de u (et celle de u^*) en prenant deux fois la même base pour E.

Opérations sur l'adjoint

Donnons quelques propriétés faciles à vérifier. Il est clair que $u^{**} = u$ et $(\lambda u)^* = \lambda u^*$ (pour tout réel λ). Pour tous $u, v \in \mathcal{L}(\mathrm{E}, \mathrm{F})$, on a $(u + v)^* = u^* + v^*$. Si w est une application linéaire de F dans un troisième espace euclidien G, $(wu)^* = u^* w^*$. Dans le cas particulier des endomorphismes, on peut ajouter la remarque $\mathrm{Id}_\mathrm{E}^* = \mathrm{Id}_\mathrm{E}$.

Endomorphismes symétriques (ou autoadjoints)

Définition 2.2.1. Soient E un espace euclidien et u un endomorphisme de E ; on dit que u est un *endomorphisme symétrique* si $u^* = u$, ce qui équivaut à dire que

$$\forall x, y \in \mathrm{E},\ u(x) \,.\, y = x \,.\, u(y).$$

La matrice de u dans toute base orthonormée de E est alors symétrique, c'est à dire $\mathrm{A} = {}^t\mathrm{A}$. Inversement, si la matrice de u dans *une* base orthonormée est symétrique, l'endomorphisme u est symétrique.

Une somme d'endomorphismes symétriques est un endomorphisme symétrique. Si λ est réel et u symétrique, λu est symétrique. Si u est symétrique, ses puissances sont symétriques. Si P est un polynôme à coefficients réels, $\mathrm{P}(u)$ est donc symétrique.

Soit E un espace euclidien ; on dit que $u \in \mathcal{L}(\mathrm{E})$ est *anti-symétrique* si $u^* = -u$, c'est à dire si

$$u(x) \,.\, y = -\, x \,.\, u(y) \ \text{ pour tous } x, y \in \mathrm{E}.$$

On note qu'alors $u(x) \,.\, x = 0$ pour tout $x \in \mathrm{E}$. On voit que u est anti-symétrique si et seulement si la matrice de u dans une base orthonormée est anti-symétrique.

Lemme. *Si $u \in \mathcal{L}(\mathrm{E})$ est symétrique, et si $u(\mathrm{F}) \subset \mathrm{F}$, la restriction v de u à F est un endomorphisme symétrique de F. De même, si $u \in \mathcal{L}(\mathrm{E})$ est anti-symétrique, et si $u(\mathrm{F}) \subset \mathrm{F}$, la restriction v de u à F est un endomorphisme anti-symétrique de F.*

Démonstration. Vérifions le premier cas. Supposons u symétrique et posons $v = u_{|\mathrm{F}}$. On aura alors

$$\forall x, y \in \mathrm{F},\ v(x) \,.\, y = u(x) \,.\, y = x \,.\, u(y) = x \,.\, v(y),$$

ce qui montre que v est symétrique. La démonstration du cas anti-symétrique est analogue.

Proposition 2.2.2. *Si $u \in \mathcal{L}(\mathrm{E})$ est symétrique ou bien anti-symétrique, et si $u(\mathrm{F}) \subset \mathrm{F}$, l'orthogonal $\mathrm{F}^\perp$ de F est stable par u.*

Démonstration. Supposons que $y \in \mathrm{F}^\perp$ et montrons que $u(y) \in \mathrm{F}^\perp$. Soit $x \in \mathrm{F}$; on a $u(x) \in \mathrm{F}$, donc $u(x) \,.\, y = 0$, ce qui entraîne

$$0 = u(x) \,.\, y = \pm\, x \,.\, u(y),$$

donc $u(y)$ est orthogonal à tout vecteur $x \in \mathrm{F}$.

Diagonalisation des endomorphismes symétriques

Pour parler de diagonalisation, il est plus agréable d'admettre les valeurs complexes.

Définition 2.2.2. Soit A une matrice carrée à coefficients complexes ; on dit que A est une *matrice hermitienne* si ${}^t\overline{\mathrm{A}} = \mathrm{A}$, c'est à dire si

$$a_{i,j} = \overline{a_{j,i}}$$

pour tout couple (i, j) d'indices.

Exemples.

1. La matrice

$$\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$

est hermitienne.

2. On remarque que les matrices symétriques réelles sont un cas particulier des matrices hermitiennes.

Pour travailler dans le cas complexe, il est plus commode d'utiliser le *produit scalaire complexe* sur $\mathbb{C}^n$, défini par

$$\langle \mathrm{X}, \mathrm{Y} \rangle = \sum_{i=1}^{n} x_i \overline{y}_i = {}^t\mathrm{X}\overline{\mathrm{Y}} = {}^t\overline{\mathrm{Y}}\mathrm{X},$$

si $\mathrm{X}, \mathrm{Y} \in \mathbb{C}^n$ ont pour coordonnées respectives $x_1, \ldots, x_n \in \mathbb{C}$ et $y_1, \ldots, y_n \in \mathbb{C}$. On note que $\langle \mathrm{X}, \mathrm{X} \rangle = \sum_{i=1}^{n} |x_i|^2$ est un réel ≥ 0, qui est nul si et seulement si $\mathrm{X} = 0$. On remarque aussi les propriétés importantes suivantes

$$\langle \mathrm{Y}, \mathrm{X} \rangle = \overline{\langle \mathrm{X}, \mathrm{Y} \rangle}; \quad \langle \lambda \mathrm{X}, \mathrm{Y} \rangle = \lambda \langle \mathrm{X}, \mathrm{Y} \rangle; \quad \langle \mathrm{X}, \lambda \mathrm{Y} \rangle = \overline{\lambda} \langle \mathrm{X}, \mathrm{Y} \rangle$$

pour tous $\mathrm{X}, \mathrm{Y} \in \mathbb{C}^n$ et $\lambda \in \mathbb{C}$. On dit que $\mathrm{X}, \mathrm{Y} \in \mathbb{C}^n$ sont $\mathbb{C}$-orthogonaux si $\langle \mathrm{X}, \mathrm{Y} \rangle = 0$. Cette relation d'orthogonalité est symétrique (bien que le produit scalaire complexe ne soit pas symétrique !).

Lemme. *Soit* A *une matrice carrée hermitienne ; les valeurs propres de* A *sont réelles. En particulier, si* A *est réelle symétrique, toutes ses valeurs propres sont réelles.*

Démonstration. Si A est une matrice hermitienne, on voit que pour tous les vecteurs colonne X et $\mathrm{Y} \in \mathbb{C}^n$, on a

$$\langle \mathrm{AX}, \mathrm{Y} \rangle = \langle \mathrm{X}, \mathrm{AY} \rangle.$$

En effet, $\langle \mathrm{AX}, \mathrm{Y} \rangle = {}^t\overline{\mathrm{Y}}\mathrm{AX} = {}^t\overline{\mathrm{Y}}\,{}^t\overline{\mathrm{A}}\mathrm{X} = {}^t\overline{(\mathrm{AY})}\,\mathrm{X} = \langle \mathrm{X}, \mathrm{AY} \rangle$. Il en résulte que $\langle \mathrm{AX}, \mathrm{X} \rangle$ est un nombre réel pour tout $\mathrm{X} \in \mathbb{C}^n$, puisque $\langle \mathrm{AX}, \mathrm{X} \rangle = \langle \mathrm{X}, \mathrm{AX} \rangle = \overline{\langle \mathrm{AX}, \mathrm{X} \rangle}$. Supposons que λ soit une valeur propre de A. Il existe alors un vecteur colonne $\mathrm{Z} \in \mathbb{C}^n$ non nul tel que $\mathrm{AZ} = \lambda \mathrm{Z}$. On a alors

$$\langle \mathrm{AZ}, \mathrm{Z} \rangle = \lambda \langle \mathrm{Z}, \mathrm{Z} \rangle,$$

ce qui entraîne que λ est réel puisque $\langle \mathrm{AZ}, \mathrm{Z} \rangle$ est réel et $\langle \mathrm{Z}, \mathrm{Z} \rangle$ réel > 0.

Exercice. Diagonaliser la matrice

$$\begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}.$$

Pour toute matrice complexe A, on définit la matrice A^* qui est la transposée de la conjuguée de A,

$$\mathrm{A}^* = {}^t\overline{\mathrm{A}}$$

et qu'on appelle la *matrice adjointe* de la matrice A. C'est la notion correcte d'adjoint dans le cas complexe, qui vérifie $\langle \mathrm{AX}, \mathrm{Y} \rangle = \langle \mathrm{X}, \mathrm{A}^*\mathrm{Y} \rangle$ pour tous $\mathrm{X}, \mathrm{Y} \in \mathbb{C}^n$. Une matrice carrée complexe A est donc hermitienne si et seulement si $\mathrm{A}^* = \mathrm{A}$.

Revenons à la diagonalisation des endomorphismes symétriques.

Théorème 2.2.1. Diagonalisation des endomorphismes symétriques. *Soit* E *un espace euclidien ; si $u \in \mathcal{L}(\mathrm{E})$ est symétrique, u est diagonalisable (en particulier, les racines du polynôme caractéristique de u sont réelles) et les sous-espaces propres de u sont deux à deux orthogonaux. Il en résulte qu'on peut trouver une base orthonormée de* E *formée de vecteurs propres de u.*

Démonstration. On montre d'abord que les racines du polynôme caractéristique sont réelles. On choisit une base orthonormée de E. Soit A la matrice de u dans cette base ; la matrice A est réelle et symétrique, donc ses valeurs propres sont réelles, ce qui donne notre premier point. On montre ensuite que les sous-espaces propres sont deux à deux orthogonaux. Si λ et μ sont deux valeurs propres distinctes, si $u(x) = \lambda x$ et $u(y) = \mu y$, on a

$$\lambda\,(x\,.\,y) = u(x)\,.\,y = x\,.\,u(y) = \mu\,(x\,.\,y),$$

donc $x\,.\,y = 0$ puisque $\lambda \neq \mu$.

On montre maintenant que u est diagonalisable. On applique la proposition 2.2.2 à la somme des espaces propres, $\mathrm{F} = \mathrm{E}_1 \oplus \cdots \oplus \mathrm{E}_k$. Il est clair que $u(\mathrm{F}) \subset \mathrm{F}$, donc l'orthogonal $\mathrm{G} = \mathrm{F}^\perp$ de F est stable par u et la restriction de u à G est un endomorphisme symétrique, qui aura un vecteur propre si $\mathrm{G} \neq \{0\}$ (d'après la première partie de la démonstration). Mais ceci est impossible puisque tous les vecteurs propres de u sont dans F. On a donc $\mathrm{F} = \mathrm{E}$ et u est diagonalisable.

Indiquons une autre démonstration de la diagonalisabilité de u : puisque toutes les racines μ du polynôme caractéristique sont réelles, il reste à montrer que pour chaque racine μ on a $\ker(u - \mu\,\mathrm{Id}_\mathrm{E}) = \ker(u - \mu\,\mathrm{Id}_\mathrm{E})^2$: l'endomorphisme $v = u - \mu\,\mathrm{Id}_\mathrm{E}$ est symétrique ; si $v^2(x) = 0$, on aura $v(x)\,.\,v(x) = v^2(x)\,.\,x = 0$, donc $v(x) = 0$.

Si A est une matrice hermitienne, elle permet de définir un endomorphisme $a = \mathrm{A}^\mathbb{C}$ de $\mathbb{C}^n$ en posant $a(x) = \mathrm{AX}$ pour tout élément x de $\mathbb{C}^n$, représenté par la matrice colonne X. Cet endomorphisme vérifie alors

$$\langle a(x), y\rangle = \langle x, a(y)\rangle$$

pour tous $x, y \in \mathbb{C}^n$. Il est utile d'énoncer le résultat de diagonalisation pour un endomorphisme d'un sous-espace de $\mathbb{C}^n$, plutôt que de se restreindre au seul langage matriciel. Si E est un $\mathbb{C}$-sous-espace vectoriel de $\mathbb{C}^n$, on dira qu'un endomorphisme $u \in \mathcal{L}(\mathrm{E})$ est un *endomorphisme hermitien* si $\langle u(x), y\rangle = \langle x, u(y)\rangle$ pour tous $x, y \in \mathrm{E}$.

Théorème 2.2.2. Diagonalisation des endomorphismes hermitiens. *Soient* E *un* $\mathbb{C}$*-sous-espace vectoriel de* $\mathbb{C}^n$ *et* u *un endomorphisme hermitien de* E *; alors*

– *les valeurs propres de u sont réelles*

– *les sous-espaces propres de u sont deux à deux $\mathbb{C}$-orthogonaux ;*

– *de plus u est diagonalisable, et on peut trouver une base $\mathbb{C}$-orthonormée de* E *formée de vecteurs propres de u.*

Démonstration. On peut donner une démonstration presque identique à celle du théorème 2.2.1. Rappelons que les valeurs propres de u sont réelles : si $u(x) = \lambda x$, on aura en supposant $\langle x, x\rangle = 1$ que $\lambda = \langle \lambda x, x\rangle = \langle u(x), x\rangle = \langle x, u(x)\rangle = \overline{\lambda}$. On montre ensuite que les sous-espaces propres de u sont deux à deux $\mathbb{C}$-orthogonaux. Si λ et μ sont deux valeurs propres distinctes, si $u(x) = \lambda x$ et $u(y) = \mu y$, on a

$$\lambda\,\langle x, y\rangle = \langle u(x), y\rangle = \langle x, u(y)\rangle = \overline{\mu}\,\langle x, y\rangle = \mu\,\langle x, y\rangle,$$

(parce que μ est réel) donc $\langle x, y\rangle = 0$. Pour finir on montre que u est diagonalisable par récurrence sur la dimension k de E. C'est évident si $k = 1$. Supposons $k > 1$ et supposons le résultat vrai lorsque $v \in \mathcal{L}(\mathrm{F})$ est hermitien, pour tout sous-espace $\mathrm{F} \subset \mathbb{C}^n$ tel que $\dim \mathrm{F} < k$. Soit maintenant E un sous-espace de dimension k de $\mathbb{C}^n$ et soit u un endomorphisme hermitien de E ; puisqu'on travaille sur $\mathbb{C}$, on peut toujours trouver un vecteur propre x de u, qui vérifie donc que $x \neq 0_{\mathrm{E}}$ et $u(x) = \lambda x$ (avec λ réel d'après ce qui précède). On peut supposer $\|x\| = 1$. Alors

$$\mathrm{F} = \{y \in \mathrm{E} : y \perp x\}$$

est un sous-espace de dimension $k-1$, qui est stable par u (en effet, pour tout $y \in \mathrm{F}$, on a $\langle u(y), x\rangle = \langle y, u(x)\rangle = \lambda\langle y, x\rangle = 0$, donc $u(y) \in \mathrm{F}$). On applique alors l'hypothèse de récurrence à la restriction v de u à F : on trouve une base $\mathbb{C}$-orthonormée $(e_2, \ldots, e_k)$ de F formée de vecteurs propres de v, donc de u ; pour finir, $(x, e_2, \ldots, e_k)$ est une base $\mathbb{C}$-orthonormée de E formée de vecteurs propres de u.

Cas de deux endomorphismes symétriques qui commutent

Proposition 2.2.3. *Si u et v sont deux endomorphismes symétriques d'un espace euclidien* E *tels que $uv = vu$, il existe une base orthonormée de* E *dont les vecteurs sont à la fois vecteurs propres de u et de v (avec des valeurs propres qui peuvent bien sûr être différentes).*

Démonstration. Soient u et v ces deux endomorphismes symétriques ; on décompose d'abord $\mathrm{E} = \mathrm{E}_1 \oplus \cdots \oplus \mathrm{E}_k$ en somme directe de sous-espaces propres de u. Considérons par exemple le sous-espace propre $\mathrm{E}_1 = \{x \in \mathrm{E} : ux = \mu_1 x\}$. Puisque u et v commutent, on sait que E_1 est stable par v. On peut donc considérer la restriction $v_1 \in \mathcal{L}(\mathrm{E}_1)$ de v à E_1. C'est un endomorphisme symétrique de l'espace euclidien E_1, donc on peut trouver une base orthonormée de E_1 formée de vecteurs propres de v (qui sont aussi des vecteurs propres pour u), et on forme une base orthonormée de E tout entier en rassemblant de telles bases pour chaque E_i, $i = 1, \ldots, k$.

La démonstration s'adapte immédiatement au cas de deux matrices hermitiennes A et B qui commutent, ou plus généralement au cas de deux endomorphismes hermitiens a et b d'un sous-espace E de $\mathbb{C}^n$, tels que $ab = ba$. On trouve une base $\mathbb{C}$-orthogonale de $\mathbb{C}^n$ formée de vecteurs propres communs aux deux matrices, ou bien une base $\mathbb{C}$-orthogonale de E formée de vecteurs propres communs aux deux endomorphismes.

Proposition 2.2.4. *Si a et b sont deux endomorphismes hermitiens d'un $\mathbb{C}$-sous-espace vectoriel* E *de $\mathbb{C}^n$ et si $ab = ba$, il existe un vecteur $z \neq 0_{\mathrm{E}}$ dans* E *tel que $a(z) = \lambda z$ et $b(z) = \mu z$, avec λ, μ réels.*

Vérification. Si a et b commutent, les sous-espaces propres de a sont stables par b, donc il suffit de prendre pour z un vecteur propre de la restriction de b à un sous-espace propre de a. Les valeurs propres seront réelles par les théorèmes précédents.

On peut poursuivre cette ligne de raisonnement pour trouver une diagonalisation commune des endomorphismes a et b.

Endomorphismes symétriques positifs

Soit u un endomorphisme symétrique d'un espace euclidien E ; on dit que u est un endomorphisme *positif* si

$$u(x) \,.\, x \geq 0 \quad \text{pour tout } x \in \mathrm{E}.$$

Cela équivaut évidemment à dire que la forme quadratique Q définie sur E par la formule $\mathrm{Q}(x) = u(x) \,.\, x$ est positive.

Il est aussi équivalent de dire que toutes les valeurs propres de u sont ≥ 0 (mais ceci est un petit exercice : si on suppose que toutes les valeurs propres de u sont positives ou nulles, on exprimera $u(x) \, . \, x$ en utilisant une base orthonormée formée de vecteurs propres de u ; l'autre direction est très facile).

Dans le même ordre d'idées, on pourra vérifier que $\mathrm{Q}(x) = u(x) \, . \, x$ est définie positive sur l'espace E si et seulement si toutes les valeurs propres de u sont > 0.

Théorème 2.2.3. Racine carrée d'un endomorphisme symétrique positif. *Si* E *est un espace euclidien et si* $u \in \mathcal{L}(\mathrm{E})$ *est symétrique positif, il existe un unique endomorphisme symétrique positif* $v \in \mathcal{L}(\mathrm{E})$ *tel que* $v^2 = u$.

Démonstration. Pour trouver une solution, diagonalisons l'endomorphisme u dans une base orthonormée $(e_1, \ldots, e_n)$ de E ; on a alors $u(e_i) = \lambda_i e_i$, avec $\lambda_i \geq 0$ pour $i = 1, \ldots, n$. Il suffit de définir v en posant $v(e_i) = \sqrt{\lambda_i} \, e_i$ pour $i = 1, \ldots, n$.

Montrons l'unicité de la solution. Si w est une autre solution, c'est à dire si w est symétrique positif et $w^2 = u$, on aura $wu = w^3 = uw$ donc les sous-espaces propres de u sont invariants par w. Si $\mathrm{F} = \mathrm{E}_\mu = \ker(u - \mu \, \mathrm{Id}_\mathrm{E})$ est l'un quelconque des sous-espaces propres de u, nous allons montrer que la restriction w_1 de w à F est égale à $\sqrt{\mu} \, \mathrm{Id}_\mathrm{F}$, donc uniquement déterminée, ce qui entraîne que w est uniquement déterminé puisque E est la somme des sous-espaces propres de u ; on voit d'abord que w_1^2 est la restriction de $w^2 = u$ à F, donc $w_1^2 = \mu \, \mathrm{Id}_\mathrm{F}$. Si $\mu = 0$ on obtient que $w_1^2 = 0$, donc pour tout $x \in \mathrm{F}$

$$\|w_1(x)\|^2 = w_1(x) \, . \, w_1(x) = w_1^2(x) \, . \, x = 0$$

puisque w_1 est symétrique, donc $w_1 = 0 = \sqrt{\mu} \, \mathrm{Id}_\mathrm{F}$ dans ce cas. Si $\mu > 0$ on écrit en posant $\rho = \sqrt{\mu}$

$$0 = w_1^2 - \mu \, \mathrm{Id}_\mathrm{F} = (w_1 + \rho \, \mathrm{Id}_\mathrm{F})(w_1 - \rho \, \mathrm{Id}_\mathrm{F})$$

et $w_1 + \rho \, \mathrm{Id}_\mathrm{F}$ est inversible (ses valeurs propres sont > 0, donc non nulles, parce que w_1 est positif et $\rho > 0$), donc $w_1 - \rho \, \mathrm{Id}_\mathrm{F} = 0$, ce qu'il fallait démontrer.

Réduction des endomorphismes anti-symétriques

Théorème 2.2.4. Réduction des endomorphismes anti-symétriques. *Soit* E *un espace euclidien ; si* $u \in \mathcal{L}(\mathrm{E})$ *est anti-symétrique, les racines du polynôme caractéristique sont de la forme* $\mu = ic$, *avec* c *réel. Pour chaque racine* $\mu = ic$ *du polynôme caractéristique de* u, *l'opposé* $-ic$ *est aussi racine. On peut trouver une base orthonormée de* E *formée d'une famille* $(e_1, \ldots, e_{2k})$ *suivie d'une base orthonormée* $(e_{2k+1}, \ldots, e_n)$ *du noyau de* u, *telle que la partie de la matrice de* u *correspondant aux vecteurs* $(e_1, \ldots, e_{2k})$ *soit diagonale par blocs* 2×2 *de la forme*

$$\begin{pmatrix} 0 & -c \\ c & 0 \end{pmatrix}$$

où c *est réel non nul.*

Si u n'est pas injective, on sait que 0 est valeur propre, et c'est alors la seule valeur propre de u, puisque toutes les racines non nulles du polynôme caractéristique sont complexes non réelles. Si u est un isomorphisme, u n'a aucune valeur propre (c'est le cas par exemple pour $u = r_{\pi/2}$, la rotation d'angle $\pi/2$ dans le plan $\mathbb{R}^2$).

Démonstration. Considérons le noyau $\mathrm{F} = \ker u$; c'est un sous-espace stable pour u, donc son orthogonal $\mathrm{G} = \mathrm{F}^\perp$ est lui aussi stable par u d'après la proposition 2.2.2. On travaille maintenant sur la restriction v de u à G. C'est un endomorphisme anti-symétrique de G, et v est inversible puisqu'il est injectif : si $y \in \mathrm{G}$, l'hypothèse $v(y) = 0$ implique $u(y) = v(y) = 0$, donc $y \in \ker u \cap \mathrm{G} = \{0_\mathrm{E}\}$, c'est à dire $y = 0_\mathrm{E}$. Choisissons une base orthonormée $\mathbf{g} = (g_1, \ldots, g_m)$ de G et soit B la matrice de v dans cette base ; c'est une matrice réelle anti-symétrique, donc la matrice complexe $\mathrm{A} = i\mathrm{B}$ est hermitienne, elle a par conséquent des valeurs propres réelles d'après le théorème 2.2.2, qui sont non nulles puisque B est inversible. Il en résulte que les valeurs propres de la matrice B sont toutes de la forme ic, avec c réel non nul. Soit $\mathrm{Z} = \mathrm{X} + i\mathrm{Y} \in \mathbb{C}^m$ un vecteur complexe non nul tel que $\mathrm{BZ} = ic\mathrm{Z}$, avec X, Y vecteurs réels ; on a alors $\mathrm{AZ} = -c\mathrm{Z}$, donc

$$\mathrm{A}\overline{\mathrm{Z}} = \overline{\overline{\mathrm{A}}\,\mathrm{Z}} = -\overline{\mathrm{AZ}} = c\overline{\mathrm{Z}},$$

(noter que $\overline{\mathrm{A}} = -\mathrm{A}$) donc $\overline{\mathrm{Z}}$ est vecteur propre de A pour la valeur propre $c \neq -c$. Il en résulte que

$$0 = \langle \mathrm{Z}, \overline{\mathrm{Z}} \rangle = {}^t\mathrm{ZZ} = ({}^t\mathrm{XX} - {}^t\mathrm{YY}) + 2i{}^t\mathrm{XY}.$$

On voit donc que X et Y sont orthogonaux et $\|\mathrm{X}\| = \|\mathrm{Y}\| \neq 0$ (sinon Z serait nul) ; on peut alors supposer que $\|\mathrm{X}\| = \|\mathrm{Y}\| = 1$. D'autre part $\mathrm{B}(\mathrm{X} + i\mathrm{Y}) = \mathrm{BX} + i\mathrm{BY} = -c\mathrm{Y} + ic\mathrm{X}$. En séparant parties réelles et imaginaires, on obtient

$$\mathrm{BX} = -c\mathrm{Y}, \quad \mathrm{BY} = c\mathrm{X}.$$

Revenons à l'espace G. A partir de X et Y, de coordonnées $(x_1, \ldots, x_m)$ et $(y_1, \ldots, y_m)$, on trouve deux vecteurs $x, y \in \mathrm{G}$ tels que $v(x) = -cy$ et $v(y) = cx$, en posant

$$x = \sum_{i=1}^m x_i g_i, \quad y = \sum_{i=1}^m y_i g_i.$$

On a alors

$$x \,.\, y = {}^t\mathrm{XY} = 0\,; \quad \|x\| = \|\mathrm{X}\| = \|\mathrm{Y}\| = \|y\| = 1.$$

Posons $e_1 = y$ et $e_2 = x$; ces deux vecteurs sont de norme 1 et orthogonaux. Le sous espace $\mathrm{G}_1 = [e_1, e_2]$ est stable par v, et la matrice de la restriction de v à ce sous-espace est de la forme annoncée. De plus, l'orthogonal de G_1 est stable par v et on peut continuer la décomposition jusqu'à épuisement de l'espace G.

Remarque. La dimension de G est paire.

Endomorphismes normaux

Définition 2.2.3. Soit E un espace euclidien et soit $u \in \mathcal{L}(\mathrm{E})$; on dit que u est **normal** si $uu^* = u^*u$.

Exemple. Rotations dans $\mathbb{R}^2$. Si $u = r_\theta$ désigne la rotation d'angle θ dans $\mathbb{R}^2$, on constate que $u^* = r_{-\theta}$, donc $u^*u = \mathrm{Id}_{\mathbb{R}^2} = uu^*$.

Théorème 2.2.5. *Supposons que u soit un endomorphisme normal d'un espace euclidien* E. *Il existe une base orthonormée de* E *dans laquelle la matrice de u est diagonale par blocs, avec des blocs diagonaux de taille un ou deux, de façon que les blocs de taille* 1×1 *correspondent à des valeurs propres réelles de u et que les blocs de taille* 2×2 *soient de la forme*

$$\begin{pmatrix} \lambda & -b \\ b & \lambda \end{pmatrix}.$$

avec λ réel et b réel non nul.

On remarque que ces matrices 2×2 sont des multiples de matrices de rotation. En effet, si on pose $r = \sqrt{\lambda^2 + b^2} > 0$, et si on écrit $\lambda = r\mu$ et $b = rc$, on aura $\mu^2 + c^2 = 1$, donc on peut trouver un angle θ tel que $\mu = \cos\theta$ et $c = \sin\theta$, et alors

$$\begin{pmatrix} \lambda & -b \\ b & \lambda \end{pmatrix} = r \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Démonstration. Supposons que u soit un endomorphisme normal de E. On voit que $v = (u + u^*)/2$ est un endomorphisme symétrique et $w = (u - u^*)/2$ est anti-symétrique. On a $u = v + w$, et ce qui est très important est que $vw = wv$ dans le cas où $uu^* = u^*u$. On sait que l'endomorphisme symétrique v est diagonalisable. Soit F un sous-espace propre de v, pour la valeur propre λ (on sait que $\lambda \in \mathbb{R}$). Alors $w(\mathrm{F}) \subset \mathrm{F}$ parce que v et w commutent. La restriction w' de w à F est un endomorphisme anti-symétrique de F, on peut donc lui appliquer la réduction déjà vue. On trouve alors une base orthonormée de F, avec certains blocs de taille 1 (correspondant au noyau de w') et les autres de taille 2×2. En ajoutant v pour reformer u, on trouve des blocs 2×2 de la forme

$$\begin{pmatrix} \lambda & -b \\ b & \lambda \end{pmatrix}.$$

Le cas complexe

Si A est une matrice carrée complexe, on dira que A est normale si $A^*A = AA^*$.

Exemple. Si A est une matrice diagonale complexe, elle est normale.

Théorème 2.2.6. Diagonalisation des matrices normales complexes. *Si* N *est une matrice normale, il existe une base* $\mathbb{C}$*-orthonormée de* $\mathbb{C}^n$ *formée de vecteurs propres de* N.

Démonstration. On peut décomposer N sous la forme $N = A + iB$, où A et B sont hermitiennes et commutent. Il suffit de poser $A = (N + N^*)/2$ et $B = i(N^* - N)/2$. Il existe alors une base $\mathbb{C}$-orthonormée de $\mathbb{C}^n$ formée de vecteurs propres communs à A et B. On peut donc diagonaliser N dans cette base $\mathbb{C}$-orthonormée.

Attention. Il s'agit d'un théorème sur $\mathbb{C}$. Comme on l'a vu, un endomorphisme normal d'un espace réel n'est pas nécessairement diagonalisable. Par ailleurs, les valeurs propres d'une matrice normale sont en général complexes (ne pas confondre avec le cas des matrices hermitiennes dont les valeurs propres sont réelles).

Remarque. Soit u un endomorphisme qui est diagonal dans une base $\mathbf{e}$,

$$u(\sum_{i=1}^n c_i e_i) = \sum_{i=1}^n \lambda_i c_i e_i.$$

Alors l'ensemble des valeurs propres est l'ensemble $\{\lambda_1, \ldots, \lambda_n\}$ (qui n'a pas nécessairement n éléments distincts !), et pour chaque λ de cet ensemble,

$$E_\lambda = [e_i : \lambda_i = \lambda]$$

(sous-espace vectoriel engendré par les vecteurs e_i tels que $\lambda_i = \lambda$). En effet, si $x = \sum_i c_i e_i \in E_\lambda$, on a $(u - \lambda\,\mathrm{Id}_E)(x) = 0_E$, donc $\sum_{i=1}^n (\lambda_i - \lambda) c_i e_i = 0_E$. Puisque (e_i) est une base, cela implique que $(\lambda_i - \lambda)c_i = 0$ pour tout $i = 1, \ldots, n$. Puisque $x \neq 0_E$, il existe des $c_i \neq 0$, et pour chacun d'entre eux, on aura nécessairement $\lambda_i = \lambda$.

Conséquence.

Les sous espaces propres d'une matrice normale sont $\mathbb{C}$*-orthogonaux.*

2.3. Isométries d'un espace euclidien

Soit T une application (qu'on ne suppose pas linéaire *a priori*) d'un espace euclidien ou préhilbertien E dans un autre espace euclidien ou préhilbertien F ; on dit que T *conserve les distances* si

$$d(T(A), T(B)) = d(A, B)$$

pour tous "points" A, B de E (les éléments de E sont aussi des vecteurs, et on peut écrire la conservation de la distance sous la forme $\|T(B) - T(A)\| = \|B - A\|$ pour tous $A, B \in E$).

Proposition 2.3.1. *Si* T *conserve les distances et si* $T(0) = 0$*, c'est une application linéaire de* E *dans* F.

Idée de démonstration (le lecteur ne se laissera pas troubler par le mélange pas très joli entre le point de vue des "points" et celui des vecteurs). En utilisant la relation du parallélogramme, on vérifie que l'image du milieu de deux points doit être le milieu des images. En considérant les deux points 0 et A, on en déduit que $T(\frac{1}{2}A) = \frac{1}{2}A$, ce qui donne aussi $T(2A) = 2\,T(A)$; on voit aussi que $T(-A) = -T(A)$, en raisonnant sur A et $-A$, dont le milieu est 0. On itère ces remarques pour montrer que

$$T\big(\frac{k}{2^n}A\big) = \frac{k}{2^n}A$$

pour tout entier $n \geq 0$ et pour tout $k \in \mathbb{Z}$. On passe ensuite à la limite pour voir que $\mathrm{T}(\lambda \mathrm{A}) = \lambda \mathrm{A}$ pour tout λ réel, et on vérifie que $\mathrm{T}(\mathrm{A}+\mathrm{B}) = \mathrm{T}(\mathrm{A})+\mathrm{T}(\mathrm{B})$ avec la propriété des milieux et l'homogénéité.

Supposons donc que T conserve les distances et que $\mathrm{T}(0) = 0$. On vient de voir que T est linéaire. De plus il est clair qu'une application qui préserve les distances est injective. Si $\mathrm{E} = \mathrm{F}$ est de dimension finie, il en résulte que T est surjective.

Corollaire 2.3.1. *Toute application* T *d'un espace euclidien* E *dans lui-même qui conserve les distances est affine et bijective. Si de plus* $\mathrm{T}(0) = 0$*, l'application* T *est un isomorphisme linéaire de* E *sur lui-même.*

Démonstration. Posons $\mathrm{S}(x) = \mathrm{T}(x) - \mathrm{T}(0)$. L'application S conserve les distances et $\mathrm{S}(0) = 0$, donc S est un isomorphisme linéaire. L'application T est obtenue en composant S avec la translation de vecteur $-\mathrm{T}(0)$, donc T est affine et bijective.

Remarque. Cette propriété de surjectivité n'est pas vraie en dimension infinie. Considérons sur l'espace ℓ_2 des suites infinies $x = (x_0, \ldots, x_n, \ldots)$ telles que $\sum x_n^2 < +\infty$, muni de la norme $\|x\| = (\sum_{n=0}^{+\infty} x_n^2)^{1/2}$, l'application linéaire S définie par

$$\mathrm{S}(x_0, x_1, \ldots) = (0, x_0, x_1, \ldots).$$

On voit facilement que S est isométrique, mais pas surjective. En effet, le vecteur

$$(1, 0, 0, \ldots)$$

n'est pas dans l'image de S.

Etude linéaire

Soient E un espace euclidien ou préhilbertien, et u un endomorphisme de E ; on dit que u est une *isométrie* si $\|u(x)\| = \|x\|$ pour tout $x \in \mathrm{E}$. Puisque u est linéaire, il en résulte que pour tous $x, y \in \mathrm{E}$,

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

donc u conserve les distances. La relation $\|u(x)\| = \|x\|$ pour tout $x \in \mathrm{E}$ signifie que $x \,.\, x = u(x) \,.\, u(x)$ pour tout $x \in \mathrm{E}$. On obtient alors par polarisation

$$y \,.\, x = u(y) \,.\, u(x)$$

pour tous $x, y \in \mathrm{E}$, ce qui montre la propriété importante suivante : toute application isométrique conserve le produit scalaire. En particulier, des vecteurs orthogonaux ont des images orthogonales. Si $u(x) = \pm x$, il en résulte que l'orthogonal de x est stable par l'isométrie u.

Supposons maintenant E euclidien et soit u une isométrie de E dans E ; on a vu que u est un isomorphisme. Il est clair que l'application u^{-1} est aussi une isométrie. Si v est une autre isométrie, il est immédiat que la composition $v \circ u$ est une isométrie. Il résulte de ces remarques que l'ensemble des isométries de E, muni de l'opération de composition des applications, est un groupe. On le note O(E), et on l'appelle le *groupe orthogonal* de l'espace euclidien E.

On peut traduire la relation précédente $y \,.\, x = u(y) \,.\, u(x)$ par

$$y \,.\, x = y \,.\, u^* u(x)$$

pour tout $x, y \in \mathrm{E}$, ce qui implique que $u^* u = \mathrm{Id}_\mathrm{E}$. Comme E est de dimension finie on a aussi $uu^* = \mathrm{Id}_\mathrm{E}$, donc $u^{-1} = u^*$. Réciproquement, si u^* est l'inverse de u, on aura pour tout $x \in \mathrm{E}$

$$\|u(x)\|^2 = u(x) \,.\, u(x) = u^* u(x) \,.\, x = x \,.\, x = \|x\|^2.$$

Définition 2.3.1. Soit u un endomorphisme d'un espace euclidien E ; on dit que u est un *endomorphisme orthogonal* si $uu^* = u^*u = \mathrm{Id}_\mathrm{E}$. On dit qu'une matrice carrée réelle U est orthogonale si ${}^t\mathrm{U}$ est l'inverse de U, c'est à dire si ${}^t\mathrm{U}\,\mathrm{U} = \mathrm{U}\,{}^t\mathrm{U} = \mathrm{I}_n$, où n est la taille de U.

Puisque E est de dimension finie, il suffit de savoir que $uu^* = \mathrm{Id}_\mathrm{E}$, ou bien que $u^*u = \mathrm{Id}_\mathrm{E}$, pour en déduire que $u^* = u^{-1}$. La même remarque s'applique aux matrices orthogonales. La matrice $\mathrm{U} = \mathrm{mat}_{\mathbf{e}}(u)$ d'un endomorphisme orthogonal u dans une base orthonormée $\mathbf{e}$ quelconque de E est une matrice orthogonale. Si U est une matrice orthogonale, on aura aussi

$$\langle \mathrm{UZ}, \mathrm{UZ}\rangle = \langle \mathrm{Z}, \mathrm{Z}\rangle$$

pour tout $\mathrm{Z} \in \mathbb{C}^n$, puisque $\langle \mathrm{UZ}, \mathrm{UZ}\rangle = {}^t(\overline{\mathrm{UZ}})\,\mathrm{UZ} = {}^t\overline{\mathrm{Z}}{}^t\overline{\mathrm{U}}\mathrm{UZ} = {}^t\overline{\mathrm{Z}}\,({}^t\mathrm{U}\,\mathrm{U})\,\mathrm{Z}$. Il en résulte que les racines du polynôme caractéristique de U sont toutes de module 1 (si un vecteur complexe Z non nul vérifie $\mathrm{UZ} = \lambda\mathrm{Z}$, on voit que $\lambda\overline{\lambda}\,\langle \mathrm{Z}, \mathrm{Z}\rangle = \langle \mathrm{UZ}, \mathrm{UZ}\rangle = \langle \mathrm{Z}, \mathrm{Z}\rangle$ ce qui donne $\lambda\overline{\lambda} = 1$). En particulier, les seules possibilités pour les valeurs propres d'un endomorphisme orthogonal u sont ± 1.

Exemple. Pour une matrice de rotation d'angle θ, les racines du polynômes caractéristique sont les complexes de module un $\lambda = \mathrm{e}^{\pm i\theta}$.

Matrices orthogonales et changement de base orthonormée

Soient $\mathbf{e}$ et $\mathbf{f}$ deux bases orthonormées du même espace euclidien E ; la matrice de passage P est une matrice orthogonale. En effet, la colonne j de cette matrice est formée des coordonnées dans la base $\mathbf{e}$ du $j$ième vecteur f_j de la base $\mathbf{f}$; si on considère l'endomorphisme u de E défini par $u(f_j) = e_j$ pour $j = 1, \ldots, n$, on vérifie facilement que u est une isométrie puisque pour tout $x = \sum_{j=1}^n c_j f_j \in \mathrm{E}$,

$$\|u(x)\|^2 = \|\sum_{j=1}^n c_j e_j\|^2 = \sum_{j=1}^n c_j^2 = \|x\|^2.$$

La matrice de u dans la base $\mathbf{f}$ est donc orthogonale. Mais on voit que $\mathrm{P} = \mathrm{mat}_{\mathbf{f}}(u)$, donc la matrice de passage P est une matrice orthogonale.

On peut voir directement par le calcul que U est une matrice orthogonale si et seulement si

– les colonnes de la matrice U forment une base orthonormée de $\mathbb{R}^n$

ou bien si et seulement si

– les lignes de la matrice U forment une base orthonormée de $\mathbb{R}^n$.

En effet, si les colonnes $\mathrm{X}_1, \ldots, \mathrm{X}_n$ d'une matrice U forment une base orthonormée de $\mathbb{R}^n$, on voit que les coefficients de la matrice produit ${}^t\mathrm{U}\,\mathrm{U}$ sont égaux aux produits scalaires $\mathrm{X}_i \,.\, \mathrm{X}_j = {}^t\mathrm{X}_i\mathrm{X}_j = \delta_{i,j}$, donc ${}^t\mathrm{U}\,\mathrm{U} = \mathrm{I}_n$ (et réciproquement, ...). Si les lignes de la matrice U forment une base orthonormée, on utilise la relation $\mathrm{U}\,{}^t\mathrm{U} = \mathrm{I}_n$.

Les matrices orthogonales de taille $n \times n$ forment un groupe pour la multiplication des matrices. On le note $\mathrm{O}(n)$, et on l'appelle le *groupe orthogonal*. Si U est orthogonale, on a $(\det \mathrm{U})^2 = \det {}^t\mathrm{U} \det \mathrm{U} = \det \mathrm{I}_n = 1$, et puisque U est réelle, ceci entraîne que $\det \mathrm{U} = \pm 1$. Le déterminant d'un endomorphisme orthogonal, ou bien d'une matrice orthogonale, est donc égal à ± 1. Les matrices $\mathrm{U} \in \mathrm{O}(n)$ telles que $\det \mathrm{U} = 1$ forment un sous-groupe de $\mathrm{O}(n)$, appelé le *groupe spécial orthogonal*, et noté $\mathrm{SO}(n)$.

Exemples.

1. Rotations de $\mathbb{R}^2$.

2. Matrice de Walsh. On part de la matrice

$$\mathrm{W}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

et on définit par récurrence des matrices de taille $2^n \times 2^n$ dont l'écriture par blocs est donnée par

$$\mathrm{W}_n = \frac{1}{\sqrt{2}} \begin{pmatrix} \mathrm{W}_{n-1} & -\mathrm{W}_{n-1} \\ \mathrm{W}_{n-1} & \mathrm{W}_{n-1} \end{pmatrix}.$$

Pour tout entier $n \geq 1$, la matrice W_n est une matrice orthogonale de taille $2^n \times 2^n$, dont tous les coefficients ont le même module $2^{-n/2}$.

Exercice. Si B est une matrice anti-symétrique, montrer que la matrice $\mathrm{e}^{-t\mathrm{B}}$ est une isométrie pour tout $t \in \mathbb{R}$. Examiner le cas de la matrice

$$\mathrm{B} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Matrices complexes unitaires

Soit A une matrice carrée complexe ; on dit que A est *unitaire* si $\mathrm{A}^* = \mathrm{A}^{-1}$. Les matrices unitaires de taille $n \times n$ forment un groupe $\mathrm{U}(n)$, le *groupe unitaire*.

Exercices.

1. Si A est une matrice hermitienne, montrer que $\mathrm{U} = \mathrm{e}^{it\mathrm{A}}$ est une matrice unitaire pour tout $t \in \mathbb{R}$.
2. Pour tout entier $n \geq 1$ on définit une matrice $\mathrm{F} = \mathrm{F}_n = (f_{j,k})$ par

$$f_{j,k} = \mathrm{e}^{2i\pi(j-1)(k-1)/n} = \xi^{(j-1)(k-1)}$$

pour $j, k = 1, \ldots, n$, et $\xi = \mathrm{e}^{2i\pi/n}$. Montrer que la matrice F est unitaire.

Soit T une isométrie (linéaire) d'un espace euclidien E ; désignons par $\mathrm{E}_1(\mathrm{T})$ le sous-espace (peut-être réduit à $\{0\}$) formé des vecteurs fixes par T, c'est à dire tels que $\mathrm{T}(x) = x$. Désignons de même par $\mathrm{E}_{-1}(\mathrm{T})$ le sous-espace formé des vecteurs x tels que $\mathrm{T}(x) = -x$. On vérifie que

– les sous-espaces $\mathrm{E}_1(\mathrm{T})$ et $\mathrm{E}_{-1}(\mathrm{T})$ sont orthogonaux.
– pour tout sous-espace F de $\mathrm{E}_1(\mathrm{T})$, l'orthogonal $\mathrm{F}^\perp$ est stable par T.

Réduction d'une isométrie

Lemme. *Soit u une isométrie d'un espace euclidien* E *; si un sous-espace* F *de* E *est stable par u^*, son orthogonal* $\mathrm{F}^\perp$ *est stable par u.*

Démonstration. Supposons que y est orthogonal à F, et montrons que $u(y)$ reste orthogonal à F ; prenons un vecteur quelconque $x \in \mathrm{F}$, et calculons $x \,.\, u(y)$. D'après l'hypothèse, on a $u^*(x) \in \mathrm{F}$, donc

$$x \,.\, u(y) = u^*(x) \,.\, y = 0.$$

Bien entendu, si F est à la fois stable par u et par u^*, son orthogonal $\mathrm{F}^\perp$ est stable par u^* et u, ce qui donne une situation plus symétrique.

Lemme. *Soit* A *une matrice carrée réelle, et soit* $Z \in \mathbb{C}^n$ *un vecteur complexe non nul tel que* $AZ = \alpha Z$*, avec* $\alpha = \lambda + i\mu$*,* λ, μ *réels et* $\mu \neq 0$ *; si on pose* $Z = X + iY \neq 0$*,* $X, Y \in \mathbb{R}^n$*, alors* X *et* Y *sont indépendants et le plan* $[X, Y]$ *est stable par* A.

Démonstration. Si on avait $Y = \beta X$, on aurait $Z = (1 + i\beta)X \neq 0$, d'où $AX = \alpha X$ ce qui est impossible puisque A et X sont réels et αX non réel. En identifiant parties réelles et imaginaires dans AZ et αZ, on trouve $AX = \lambda X - \mu Y$ et $AY = \mu X + \lambda Y$, ce qui montre bien que le plan $[X, Y]$ est stable par A.

Théorème 2.3.1. *Soit* u *une isométrie d'un espace euclidien* E *; il existe une décomposition de* E *en somme directe orthogonale de sous-espaces de dimension* 1 *ou* 2 *stables par l'endomorphisme* u.

On voit que u est normal, donc on peut décomposer E en sous-espaces stables de dimension 1 et 2, deux à deux orthogonaux. Les blocs 2×2 de la décomposition des normaux sont nécessairement ici des rotations, et les seules valeurs propres possibles pour un endomorphisme orthogonal sont 1 et -1.

Démonstration. La démonstration se fait par récurrence sur $n = \dim E$. Le résultat est évident pour $n = 1$. Prenons maintenant $n > 1$, et supposons le résultat démontré pour toute isométrie d'un espace euclidien de dimension $< n$. Soit u une isométrie d'un espace euclidien E de dimension n ; si u a une valeur propre λ (forcément ± 1) on prend un vecteur propre $x \in E$, on remarque que $u^*(x) = \lambda^{-1} x$, donc $\mathbb{R}x$ est stable par u et par u^*, ce qui implique que $F = (\mathbb{R}x)^\perp$ est stable par u, et on peut appliquer l'hypothèse de récurrence : la restriction v de u à F est une isométrie de l'espace euclidien F, et $\dim F < n$, donc F se décompose en somme directe orthogonale $F_1 \oplus \cdots \oplus F_k$ de sous-espaces de dimension 1 ou 2 stables par v, donc par u, et $E = \mathbb{R}x \oplus F_1 \oplus \cdots \oplus F_k$ donne la décomposition voulue en sous-espaces stables par u deux à deux orthogonaux.

Sinon, toutes les racines du polynôme caractéristique de u sont imaginaires non réelles ; on se ramène alors à $\mathbb{R}^n$ et à une matrice orthogonale U en prenant une base orthonormée $\mathbf{e}$ de E et $U = \mathrm{mat}_{\mathbf{e}}(u)$. On pose $A = (U + {}^tU)/2$ et $B = -i(U - {}^tU)/2$; on remarque que A et B sont hermitiennes et commutent (parce que $U\,{}^tU = {}^tU U = I_n$), et que $U = A + iB$, ${}^tU = A - iB$. D'après la proposition 2.2.4, il existe un vecteur non nul $Z \in \mathbb{C}^n$ tel que $AZ = \lambda Z$ et $BZ = \mu Z$, avec λ, μ réels. On en déduit que $UZ = (\lambda + i\mu)Z$; si on pose $\alpha = \lambda + i\mu$, on voit que α est racine du polynôme caractéristique de U (c'est à dire de u). D'après notre division en cas, α n'est pas réelle, donc on a ici $\mu \neq 0$; on a ${}^tUZ = AZ - iBZ = (\lambda - i\mu)Z$. Posons $Z = X + iY$, avec $X, Y \in \mathbb{R}^n$; d'après le lemme précédent $[X, Y]$ est stable par ${}^tU$ (et par U). Ceci entraîne que $F = [X, Y]^\perp$ est stable par U, donc on peut lui appliquer l'hypothèse de récurrence et terminer comme dans le premier cas.

Réflexions, symétries orthogonales

Soit E un espace euclidien ; supposons que $E = F \oplus G$, avec F et G orthogonaux (c'est à dire que $G = F^\perp$). Tout vecteur $x \in E$ admet une décomposition unique $x = f + g$, avec $f \in F$ et $g \in G$. On note que $f = P_F(x)$ et $g = P_G(x)$. La *symétrie orthogonale* autour de F est l'application linéaire définie par $S_F(x) = f - g$ lorsque $x = f + g$, ou encore

$$S_F = P_F - P_G = 2P_F - \mathrm{Id}_E = \mathrm{Id}_E - 2P_G.$$

On voit que si $x = f + g$, on a à cause de l'orthogonalité de f et g

$$\|S_F(x)\|^2 = \|f\|^2 + \|g\|^2 = \|f + g\|^2 = \|x\|^2,$$

donc S_F est une isométrie. On voit que $S_F(f) = f$ pour tout $f \in F$ et $S_F(g) = -g$ pour tout vecteur $g \in G$. Dans une base orthogonale de E formée d'une base $(f_1, \ldots, f_k)$ de F suivie d'une base $(g_{k+1}, \ldots, g_n)$ de G, la matrice de S_F est diagonale, avec des 1 pour la partie F et des -1 pour la partie G. Le déterminant est donc égal à $(-1)^{\dim G}$. On note que $S_F^2 = \mathrm{Id}_E$.

On appelle *réflexion* toute symétrie orthogonale par rapport à un hyperplan F d'un espace euclidien E (la symétrie S_F est donc appelée réflexion dans le cas où $G = F^\perp$ est de dimension un). Désignons par w un vecteur non nul orthogonal à F ; on a alors $G = \mathbb{R}w$. La projection orthogonale P_G sur la droite G est donnée par

$$P_G(x) = \frac{x \,.\, w}{w \,.\, w}\, w$$

pour tout $x \in E$. On obtient la symétrie orthogonale S_F par rapport à l'hyperplan F en prenant $S_F = \mathrm{Id}_E - 2P_G$. On appellera R_w la réflexion ainsi obtenue, autour de l'hyperplan $F = [w]^\perp$ orthogonal au vecteur w ; elle est donc définie par la formule

$$\forall x \in E, \quad R_w(x) = x - 2\,\frac{x \,.\, w}{w \,.\, w}\, w$$

qui nous donne une isométrie, pour laquelle $R_w(w) = -w$, et $R_w(x) = x$ pour tout vecteur x orthogonal à w (les vérifications directes de ces affirmations, indépendamment de la discussion précédente, sont faciles ou immédiates). On montrera plus loin que toute isométrie est produit de réflexions.

Posons $z = w/\|w\|$; la matrice P de P_G dans une base orthonormée de E est donnée par $P_{i,j} = z_i z_j$, où $z_1, \ldots, z_n$ sont les coordonnées de z dans ladite base. On obtient la matrice de S_F en prenant $I_n - 2P$. Les coefficients de la matrice de R_w dans la base précédente sont donc $(\delta_{i,j} - 2\, z_i z_j)$.

Isométries de $\mathbb{R}^2$ et de $\mathbb{R}^3$

Prenons d'abord un espace euclidien E de dimension 2 et u une isométrie de E. On sait que le déterminant de u est égal à 1 ou à -1. Etudions d'abord le cas où $\det u = -1$. Le polynôme caractéristique de u est réel de degré deux. Si le produit des racines est -1, il y a deux racines réelles. Les racines réelles sont des valeurs propres, nécessairement égales à 1 ou à -1. Si le produit est égal à -1, il y a une racine 1 et une racine -1, donc u est diagonalisable. Il existe donc une base de vecteurs propres (e_1, e_2) de E telle que

$$u(e_1) = e_1, \quad u(e_2) = -e_2.$$

On voit que les vecteurs e_1 et e_2 sont orthogonaux puisque $-e_1 \,.\, e_2 = u(e_1) \,.\, u(e_2) = e_1 \,.\, e_2$. On voit aussi que $u^2 = \mathrm{Id}_E$. L'application u est la symétrie orthogonale par rapport à la droite $[e_1]$, c'est à dire la réflexion R_{e_2}.

Dans la base canonique, écrivons une réflexion de $\mathbb{R}^2$. Soit $v = (a, b)$ un vecteur fixe pour u de norme 1 ; la projection orthogonale sur $[v]$ est donnée par

$$P(x) = (x \,.\, v)\, v,$$

de matrice

$$\begin{pmatrix} a^2 & ab \\ ab & b^2 \end{pmatrix}$$

et la symétrie est $\mathrm{S} = 2\mathrm{P} - \mathrm{Id}$ de matrice

$$\begin{pmatrix} 2a^2-1 & 2ab \\ 2ab & 2b^2-1 \end{pmatrix} = \begin{pmatrix} \cos(2\theta) & \sin(2\theta) \\ \sin(2\theta) & -\cos(2\theta) \end{pmatrix}$$

si on a posé $a = \cos\theta$ et $b = \sin\theta$.

Supposons maintenant que $\det u = 1$. L'image du vecteur $(1,0)$ est un vecteur de norme 1, qu'on peut écrire $(\cos\theta, \sin\theta)$. L'image de $(0,1)$ est un vecteur de norme 1 orthogonal au précédent, donc égal à $\pm(-\sin\theta, \cos\theta)$. Puisque le déterminant est égal à 1 on a nécessairement ici la matrice de la rotation d'angle θ.

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Le groupe $\mathrm{SO}(2)$ est donc le groupe des rotations de $\mathbb{R}^2$. Il est isomorphe (en tant que groupe bien sûr) au groupe des nombres complexes de module 1, muni de la multiplication. Passons maintenant à la question de la décomposition des isométries en composition d'opérations simples ; en dimension deux, nous avons :

soit ρ une rotation de E ; il existe deux réflexions R_1 et R_2 telles que $\rho = \mathrm{R}_1\mathrm{R}_2$.

C'est un cas particulier du résultat général qui sera démontré plus loin. Cela revient à dire que $\rho\mathrm{R}_2 = \mathrm{R}_1$. Si on prend n'importe quelle réflexion R_2, le produit $\rho\mathrm{R}_2$ est une isométrie de déterminant -1, donc c'est une réflexion R_1, ce qui démontre l'affirmation ci-dessus. Mais on aimerait avoir une construction plus explicite. Soient x un vecteur de norme 1, et $y = \rho(x)$; déterminons une réflexion R_2 telle que $\mathrm{R}_2(y) = x$. Puisque x et y ont la même norme, la droite Δ passant par 0 et par le milieu $m = (x+y)/2$ de x et y est orthogonale au segment $[x,y]$: c'est la médiatrice du segment $[x,y]$. Soit R_2 la réflexion autour de Δ ; on voit que $\mathrm{R}_2(y) = x$, donc $\rho\mathrm{R}_2(y) = y$ et $\rho\mathrm{R}_2$ est une isométrie de déterminant -1, donc c'est une réflexion R_1. Puisque y reste fixe, R_1 est la réflexion autour de la droite $[y]$.

Le cas de $\mathbb{R}^3$. Soit S une isométrie de $\mathbb{R}^3$; supposons encore que $\det \mathrm{S} = 1$. On vérifie qu'alors 1 est nécessairement racine du polynôme caractéristique de S : s'il y a une racine λ non réelle (de module un), les racines seront $\mu \in \mathbb{R}$, λ et $\overline{\lambda}$, donc $1 = \det \mathrm{S} = \mu\,|\lambda|^2 = \mu$; d'un autre côté, si les trois racines sont réelles, elles valent ± 1 et le produit 1, il y a donc au moins un 1 dans la liste. Il existe donc un vecteur e_3 (de norme un si on veut) tel que $\mathrm{S}(e_3) = e_3$. Alors l'orthogonal F de $\mathbb{R}e_3$ est un plan stable par S, et la restriction S' de S à F est une isométrie en dimension 2, dont on vérifie que le déterminant est encore égal à 1 (écrire la matrice de S dans une base de $\mathbb{R}^3$ formée d'une base de F et du vecteur e_3), donc S' est une rotation du plan F. Finalement, S est une rotation de $\mathbb{R}^3$ autour de l'axe $\mathbb{R}e_3$.

Si $\det \mathrm{S} = -1$, on peut considérer $-\mathrm{S}$ pour déduire qu'il existe maintenant e_3 tel que $\mathrm{S}(e_3) = -e_3$. Si T désigne la symétrie autour de $\mathrm{F} = e_3^{\perp}$, on voit d'après ce qui précède que $\mathrm{R} = \mathrm{ST} = \mathrm{TS}$ est une rotation autour de la droite $\mathbb{R}e_3$, et $\mathrm{S} = \mathrm{RT} = \mathrm{TR}$. En résumé, une isométrie S de $\mathbb{R}^3$ est

– ou bien une rotation autour d'un axe (si la rotation est d'angle nul, on obtient simplement l'application identique) ; ce cas se produit lorsque $\det \mathrm{S} = 1$.

– ou bien une rotation autour d'un axe de vecteur directeur z suivie (ou précédée) de la symétrie autour du plan $z^{\perp}$ orthogonal à l'axe de rotation ; ce cas se produit lorsque $\det \mathrm{S} = -1$.

Il résulte de l'étude que les éléments de SO(3) sont exactement les rotations autour d'un axe.

Théorème 2.3.2. *Tout endomorphisme orthogonal d'un espace euclidien est produit de réflexions.*

Démonstration. On pose $n = \dim \mathrm{E} > 0$ et on travaille, pour n fixé, par récurrence descendante sur la dimension $k = \dim \mathrm{E}_1(\mathrm{T})$, avec $k = n, n-1, \ldots, 0$, du sous-espace des vecteurs fixes de l'isométrie T. Si $k = n$, T est l'identité, que l'on peut considérer comme produit de 0 réflexions (ou produit de deux réflexions si on préfère, en écrivant $\mathrm{Id}_\mathrm{E} = \mathrm{R}^2$ pour une réflexion R quelconque). Supposons $0 \le k < n$ et supposons le résultat établi pour toute isométrie S de E telle que $\dim \mathrm{E}_1(\mathrm{S}) > k$; soit T une isométrie de E telle que $\dim \mathrm{E}_1(\mathrm{T}) = k$, et posons $\mathrm{F} = \mathrm{E}_1(\mathrm{T})^\perp$; on vérifie que F est stable par T ; on a $\dim \mathrm{F} = n - k > 0$, donc on peut sélectionner un vecteur non nul $y \in \mathrm{F}$. Puisque $y \notin \mathrm{E}_1(\mathrm{T})$ on a $y \neq \mathrm{T}(y)$; posons $z = \mathrm{T}(y)$ et $w = z - y$. On a $w \neq 0$, $z \in \mathrm{F}$ (car F est stable) donc $w \in \mathrm{F}$, et d'autre part

$$(z+y)\,.\,w = (z+y)\,.\,(z-y) = z\,.\,z - y\,.\,y = 0$$

($\|z\|^2 = z\,.\,z = y\,.\,y$ puisque z est l'image de y par une isométrie T). Considérons la réflexion R_w. Pour cette réflexion on a $\mathrm{R}_w(w) = -w$ c'est à dire $\mathrm{R}_w(z-y) = y - z$ et $\mathrm{R}_w(z+y) = z+y$ puisque $z+y$ est orthogonal à w ; on a donc $\mathrm{R}_w(z) = y$ en additionnant ; d'autre part $\mathrm{R}_w(x) = x$ pour tout $x \in \mathrm{E}_1(\mathrm{T})$, parce que $x \perp w$. Il en résulte que $\mathrm{R}_w(\mathrm{T}(y)) = y$ et $\mathrm{R}_w(\mathrm{T}(x)) = \mathrm{R}_w(x) = x$ pour tout $x \in \mathrm{E}_1(\mathrm{T})$; le sous-espace $\mathrm{E}_1(\mathrm{R}_w\mathrm{T})$ des vecteurs fixes du produit $\mathrm{R}_w\mathrm{T}$, qui contient à la fois $\mathrm{E}_1(\mathrm{T})$ et y, est donc de dimension $> k$; par l'hypothèse de récurrence, il en résulte que $\mathrm{S} = \mathrm{R}_w\mathrm{T}$ est produit de réflexions. Il en résulte que T aussi est produit de réflexions, puisque $\mathrm{T} = \mathrm{R}_w\mathrm{S}$.

Exercices. Les groupes $\mathrm{O}(1,1)$ et $\mathrm{O}(2,1)$

1. Montrer que les matrices

$$\mathrm{A} = \begin{pmatrix} \operatorname{ch}\theta & \operatorname{sh}\theta \\ \operatorname{sh}\theta & \operatorname{ch}\theta \end{pmatrix}$$

où θ varie dans $\mathbb{R}$, forment un groupe pour la multiplication des matrices, et qu'elles préservent la forme quadratique Q_2 sur $\mathbb{R}^2$ égale à $\mathrm{Q}_2(x,y) = x^2 - y^2$, c'est à dire que $\mathrm{Q}_2(\mathrm{AX},\mathrm{AX}) = \mathrm{Q}_2(\mathrm{X},\mathrm{X})$ pour tout $\mathrm{X} \in \mathbb{R}^2$. Déterminer toutes les matrices qui préservent Q_2 (elles forment le groupe $\mathrm{O}(1,1)$, le groupe orthogonal de la forme quadratique Q_2).

2. On considère sur $\mathbb{R}^3$ la forme quadratique $\mathrm{Q}_3(x,y,z) = x^2 + y^2 - z^2$, on désigne par φ la forme polaire et on considère la matrice

$$\mathrm{J} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

– 2a. Montrer qu'une matrice A de taille 3×3 préserve la forme quadratique Q_3 si et seulement si ${}^t\mathrm{AJA} = \mathrm{J}$. Montrer que le déterminant de A est alors égal à ± 1. Montrer que ces matrices forment un groupe (le groupe $\mathrm{O}(2,1)$).

– 2b. Essayer de généraliser le théorème 2.3.2 au cas actuel.

– 2c. Vérifier qu'il n'existe pas de sous-espace de dimension deux de $\mathbb{R}^3$ formé de vecteurs isotropes de Q_3. Si $\mathrm{A} \in \mathrm{O}(2,1)$ et si λ est une valeur propre réelle de A telle que $\lambda^2 \neq 1$, montrer que l'espace propre est de dimension 1 et engendré par un vecteur isotrope.

– 2d. Si A admet un vecteur propre $x \in \mathbb{R}^3$ qui est isotrope, montrer que l'orthogonal de x (par rapport à φ : ici et dans la suite) est stable, et montrer que toutes les valeurs propres de A sont alors réelles. Si $\mathrm{A}x = \lambda x$ avec $\lambda^2 \neq 1$, montrer qu'il existe y isotrope non nul tel que $\mathrm{A}y = \lambda^{-1}y$ et un troisième vecteur z non nul, non isotrope, tel que $\mathrm{A}z = \pm z$.

– 2e. Si $\mathrm{A}x = \pm x$ pour un vecteur non nul, non isotrope, vérifier que $\mathrm{V} = [x]^\perp$ est stable par A. Montrer que l'étude de la restriction de A à V ramène à $\mathrm{O}(2)$ ou à $\mathrm{O}(1,1)$.

– 2f. Si $\mathrm{A}x = x$ pour un vecteur non nul et isotrope, l'espace $\mathrm{V} = [x]^\perp$ est toujours stable mais il contient x ; si la deuxième valeur propre de $\mathrm{A}_{|\mathrm{V}}$ est $\neq 1$, montrer que A est diagonalisable.

– 2g. Etudier la matrice

$$\frac{1}{4}\begin{pmatrix} 3 & 2\sqrt{2} & 1 \\ -2\sqrt{2} & 4 & 2\sqrt{2} \\ -1 & 2\sqrt{2} & 5 \end{pmatrix}.$$

Donner une condition qui garantisse qu'une matrice $\mathrm{A} \in \mathrm{O}(2,1)$ soit diagonalisable sur $\mathbb{C}$.

Chapitre 3. Introduction au calcul différentiel

3.1. Préliminaires

Soient E un espace vectoriel réel de dimension finie, a un point de E et f, g deux fonctions réelles définies au voisinage de a ; on dit que f appartient à la classe $\mathrm{O}_a(g)$ s'il existe un voisinage W de a, contenu dans les ensembles de définition de f et de g, et une constante M tels que

$$\forall x \in \mathrm{W}, \quad |f(x)| \le \mathrm{M}\,|g(x)|.$$

On dit que f appartient à la classe $o_a(g)$ si pour tout $\varepsilon > 0$, il existe un voisinage W_ε de a (contenu dans les ensembles de définition de f et de g) tel que

$$\forall x \in \mathrm{W}_\varepsilon, \quad |f(x)| \le \varepsilon\,|g(x)|.$$

Dire que $f \in o_a(g)$ signifie exactement qu'il existe une fonction $x \to \varepsilon(x)$ définie au voisinage de a telle que $\lim_{x\to a} \varepsilon(x) = 0$ et $f(x) = g(x)\varepsilon(x)$ au voisinage de a. Il est évident que la classe $o_a(g)$ est contenue dans la classe $\mathrm{O}_a(g)$ (prendre $\varepsilon = 1$ par exemple dans la définition de $o_a(g)$, puis $\mathrm{M} = \varepsilon = 1$ dans celle de $\mathrm{O}_a(g)$).

Les deux notions précédentes se définissent aussi pour des fonctions définies sur un voisinage épointé de a, c'est à dire un ensemble de la forme $\mathrm{V} \setminus \{a\}$, où V est un voisinage de a dans E.

Presque tout le temps dans ce poly on prendra $a = 0_\mathrm{E}$. Si on considère une norme $x \to \|x\|$ sur E et si on pose $g(x) = \|x\|$ nous utiliserons la notation peu correcte mais traditionnelle "$f(x) = o(\|x\|)$ lorsque $x \to 0_\mathrm{E}$" ou bien "$f(x) = \mathrm{O}(\|x\|)$ lorsque $x \to 0_\mathrm{E}$", au lieu des notations $f \in o_0(g)$ ou bien $f \in \mathrm{O}_0(g)$. A l'occasion dans ce qui suit, on parlera aussi des classes "$o(\|x\|)$ lorsque $x \to 0_\mathrm{E}$" et "$\mathrm{O}(\|x\|)$ lorsque $x \to 0_\mathrm{E}$".

Exemples.

1. Sur $\mathbb{R}^2$, les deux fonctions coordonnées $x \to x_1$, $x \to x_2$ sont $\mathrm{O}(\|x\|)$ lorsque $x \to 0$. On en déduit que x_1x_2 est $\mathrm{O}(\|x\|^2)$, donc $\sqrt{|x_1x_2|}$ est $\mathrm{O}(\|x\|)$.
2. Pour tout $\beta > 1$, on a $\mathrm{O}(\|x\|^\beta) \subset o(\|x\|)$ lorsque $x \to 0$.
3. Etudions au voisinage de $(0,0)$ la fonction f de deux variables définie par la formule $f(x_1, x_2) = \mathrm{e}^{x_1+2x_2}$. On a

$$e^u = 1 + u + \mathrm{O}(u^2)$$

lorsque $u \to 0$, donc

$$f(x) = 1 + (x_1 + 2x_2) + \mathrm{O}(\|x\|^2)$$

lorsque $x \to 0$.

Puisque E est un espace vectoriel réel de dimension finie, on sait que toutes les normes sur E sont équivalentes ; on voit donc que la classe $\mathrm{O}(\|x\|)$ lorsque $x \to 0_\mathrm{E}$ ne dépend pas de la norme choisie sur E ; on dira donc que $f(x) = \mathrm{O}(x)$ lorsque $x \to 0_\mathrm{E}$, si $f(x) = \mathrm{O}(\|x\|)$ pour une norme quelconque sur E. On dira que $f(x) = o(x)$ lorsque $x \to 0_\mathrm{E}$, si $f(x) = o(\|x\|)$ pour une norme quelconque sur E ; ici encore, la notion ne dépend pas de la norme choisie.

Supposons maintenant que F soit un deuxième espace vectoriel réel de dimension finie, et que f soit maintenant à valeurs dans F (et soit définie sur un voisinage de 0_E dans E). Si on choisit une norme sur F, on pourra dire que $f(h) = \mathrm{O}(h)$ lorsque $h \to 0_\mathrm{E}$

si $\|f(h)\|_{\mathrm{F}} = \mathrm{O}(h)$ lorsque $h \to 0_{\mathrm{E}}$ (on notera parfois $f(h) = \overrightarrow{\mathrm{O}}(h)$ quand on voudra insister sur le caractère vectoriel des valeurs de f). La notion ne dépend pas de la norme choisie sur F. On définira de même la notion $f(h) = \overrightarrow{o}(h)$ lorsque $h \to 0_{\mathrm{E}}$.

Proposition 3.1.1. *Les fonctions f ou g qui suivent sont supposées définies dans un voisinage du zéro d'un espace vectoriel réel* E *de dimension finie, à valeurs réelles ou vectorielles.*

1. Si $f(x) = o(x)$ lorsque $x \to 0_{\mathrm{E}}$, alors $f(x) = \mathrm{O}(x)$; si $g(x) = \mathrm{O}(x)$ lorsque $x \to 0_{\mathrm{E}}$, alors $\lim_{x \to 0} g(x) = 0$.

2. Si $f_1(x) = o(x)$ et $f_2(x) = o(x)$ lorsque $x \to 0_{\mathrm{E}}$, on a $f_1(x) + f_2(x) = o(x)$; on a un résultat analogue d'addition pour les classes $\mathrm{O}(x)$.

3. Soient V *un voisinage de* 0_{E} *dans* E, W *un voisinage de* 0_{F} *dans* F *; soient d'autre part $f : \mathrm{V} \to \mathrm{W}$ et $g : \mathrm{W} \to \mathrm{G}$ deux applications telles que $f(x) = \mathrm{O}(x)$ lorsque $x \to 0_{\mathrm{E}}$ et $g(y) = o(y)$ lorsque $y \to 0_{\mathrm{F}}$. La composée $g \circ f$ est alors $o(x)$ lorsque $x \to 0_{\mathrm{E}}$; on a le même résultat lorsque $f(x) = o(x)$ et $g(y) = \mathrm{O}(y)$.*

4. Si u est une application linéaire entre deux espaces vectoriels de dimension finie, on a $u(x) = \mathrm{O}(x)$ lorsque $x \to 0_{\mathrm{E}}$.

Démonstration. Les deux premiers points sont évidents. Montrons le troisième ; on peut écrire $g(y) = \|y\| \, \varepsilon(y)$, où ε est une fonction de F dans G telle que $\lim_{y \to 0_{\mathrm{F}}} \varepsilon(y) = 0_{\mathrm{G}}$. D'autre part, il existe une constante M telle que $\|f(x)\| \le \mathrm{M}\|x\|$ pour x assez petit. On aura donc pour $x \ne 0_{\mathrm{E}}$

$$g(f(x)) = \|f(x)\| \, \varepsilon(f(x)) = \|x\| \left(\frac{\|f(x)\|}{\|x\|} \, \varepsilon(f(x)) \right) = \|x\| \, \varepsilon_1(x),$$

où $\varepsilon_1(x) = (\|f(x)\| / \|x\|) \, \varepsilon(f(x))$ tend vers 0_{G} lorsque $x \to 0_{\mathrm{E}}$: en effet, le quotient $\|f(x)\| / \|x\|$ reste borné au voisinage de 0_{E}, $f(x)$ tend vers 0 lorsque $x \to 0_{\mathrm{E}}$ puisque $f(x) = \mathrm{O}(x)$, donc $\varepsilon(f(x))$ tend vers 0_{G} lorsque $x \to 0_{\mathrm{E}}$.

Etant donnés deux espaces vectoriels normés E et F, on a défini une norme sur l'espace $\mathcal{L}(\mathrm{E}, \mathrm{F})$ des applications linéaires de la façon suivante : si u est une application linéaire de E dans F, on a posé

$$\|u\| = \sup\{\|u(x)\| : x \in \mathrm{E}, \ \|x\| \le 1\}.$$

On a alors l'inégalité

$$\|u(x)\| \le \|u\| \, \|x\|$$

pour tout vecteur $x \in \mathrm{E}$. On dit parfois que cette norme sur $\mathcal{L}(\mathrm{E}, \mathrm{F})$ est *subordonnée* aux deux normes données sur E et F. Cette norme a la propriété importante suivante : si v et u sont composables, on a $\|v \circ u\| \le \|v\| \, \|u\|$. En particulier, dans le cas d'un endomorphisme u on a $\|u^k\| \le \|u\|^k$ pour tout entier $k \ge 0$.

Bien que très utile pour développer la théorie, cette norme "subordonnée" est rarement calculable explicitement. Par exemple, si on se donne une matrice réelle 3×3, qui définit un endomorphisme u de $\mathbb{R}^3$, muni de la norme euclidienne, on n'aura en général pas de formule explicite donnant la norme de u en fonction des neuf coefficients de la matrice.

Cette définition de norme s'applique en particulier aux formes linéaires sur E : étant donnée une norme sur E, on en déduit une norme sur l'espace dual $E^* = \mathcal{L}(E, \mathbb{R})$, qu'on appelle la *norme duale*, et qui est donc définie par

$$\|x^*\| = \sup\{|x^*(x)| : x \in E,\ \|x\| \leq 1\}.$$

On a encore l'inégalité $|x^*(x)| \leq \|u\|\,\|x\|$ pour tous $x^* \in E^*$, $x \in E$. Examinons le cas particulier des formes linéaires sur un espace *euclidien* E. On a vu (théorème 2.1.2) que toute forme linéaire $\ell \in E^*$ provient d'un unique élément $x_\ell \in E$ au moyen de la formule

$$\forall y \in E, \quad \ell(y) = y \,.\, x_\ell$$

On a alors $|\ell(y)| \leq \|y\|\,\|x_\ell\|$ par Cauchy-Schwarz, donc $\|\ell\|_{E^*} \leq \|x_\ell\|$. Inversement, $\|x_\ell\|^2 = x_\ell \,.\, x_\ell = \ell(x_\ell) \leq \|\ell\|\,\|x_\ell\|$ montre que $\|\ell\| = \|x_\ell\|$.

On sait que les applications linéaires entre espaces vectoriels réels de dimension finie sont lipschitziennes, donc continues. Les formes quadratiques sur un espace de dimension finie sont également continues : supposons que Q soit une forme quadratique sur un espace réel E de dimension finie, avec φ comme forme polaire ; si on choisit une base $\mathbf{e}$ de E, et si on désigne par Φ la matrice de Q dans cette base, on obtiendra pour tous vecteurs x et y de E, dont les composantes donnent les vecteurs colonne X et Y de $\mathbb{R}^n$ la majoration

$$|\varphi(x, y)| = \left|{}^t X \Phi Y\right| = \left|X \,.\, (\Phi Y)\right| \leq \|X\|\,\|\Phi\|_2\,\|Y\| = \|\Phi\|_2\,\|x\|\,\|y\|.$$

En particulier, il existe une constante M telle que pour tout $x \in E$, on puisse écrire $|Q(x)| \leq M\|x\|^2$. Il en résulte facilement que la fonction Q est continue. En effet, on a $Q(x+h) - Q(x) = 2\varphi(x, h) + Q(h)$, et on peut majorer la valeur absolue de la différence par $2M\|x\|\,\|h\| + M\|h\|^2$, quantité que l'on peut rendre petite, pour tout $x \in E$ fixé, en choisissant $\|h\|$ assez petit (dépendant de x).

Il est facile de montrer suivant le même principe la continuité des applications bilinéaires entre espaces de dimension finie.

3.2. La différentielle (cas des valeurs réelles)

Commençons par expliquer la différence de point de vue entre dérivée et différentielle dans le cas de la dimension un. Si f est une fonction réelle d'une seule variable, dérivable au point a, on peut considérer deux objets très voisins : le nombre $f'(a)$, qui est la dérivée, et l'application $(df)_a : h \in \mathbb{R} \to f'(a)\,h$, application linéaire de $\mathbb{R}$ dans $\mathbb{R}$, qui apparaît par exemple dans l'expression de la dérivabilité par un développement limité d'ordre 1,

$$f(a+h) = f(a) + (df)_a(h) + o(h).$$

Si on fait l'opération un peu absurde de considérer la matrice de l'application linéaire $(df)_a$ dans la base canonique de $\mathbb{R}$, on obtient bien sûr la matrice 1×1 dont le seul coefficient est le nombre $f'(a)$. On pourrait continuer à privilégier le point de vue *nombre dérivé* plutôt que la notion d'application linéaire dans le cas de fonctions numériques définies sur $\mathbb{R}^n$, en mettant en avant la matrice des dérivées partielles. Ce point de vue n'est pas très satisfaisant pour plusieurs raisons ; tout d'abord, il y a des cas où on a besoin de considérer des fonctions réelles définies sur un espace vectoriel qui n'a pas de base naturelle. Dans une telle situation, le point de vue “application linéaire” sera mieux adapté. Ensuite, certaines notions telles que la différentielle d'une composition d'applications sont bien plus agréables à exprimer dans le langage abstrait des applications linéaires que dans celui des matrices de dérivées partielles.

Effet de zoom et différentielle

Soient a un point de $\mathbb{R}^2$, U un disque ouvert de centre a et de rayon $r > 0$ et soit $f : \mathrm{U} \to \mathbb{R}$ une fonction réelle ; on veut étudier la fonction f au voisinage de a. A cet effet on commence par recentrer le problème en posant $\varphi(h) = f(a+h) - f(h)$, de façon à remplacer l'étude de la fonction f au voisinage de a par l'étude de φ au voisinage de 0, avec aussi $\varphi(0) = 0$. On considère le graphe $\mathrm{G} \subset \mathbb{R}^2 \times \mathbb{R}$ de la fonction φ,

$$\mathrm{G} = \{(h, \varphi(h)) : h \in \mathrm{V}\},$$

où V est le disque ouvert centré en $0_{\mathbb{R}^2}$ et de rayon r. Pour "regarder de plus près" ce qui se passe au voisinage du point $(0,0) \in \mathrm{G}$, effectuons un "zoom" en multipliant l'ensemble G par 2 (homothétie dans $\mathbb{R}^2 \times \mathbb{R}$). On obtient ainsi un ensemble G_2 qui est le graphe d'une nouvelle fonction φ_2 : si $(k, y) \in \mathrm{G}_2$, on veut écrire $y = \varphi_2(k)$; on a $(k, y) = 2(h, \varphi(h))$ pour un certain $h \in \mathrm{V}$, donc $k = 2h$ et $y = 2\varphi(h)$; on voit donc que φ_2 est définie par la formule $\varphi_2(k) = y = 2\varphi(k/2)$. On peut plus généralement poser pour tout entier $n \ge 2$ et pour tout $h \in \mathrm{V}$

$$\varphi_n(h) = 2^n \varphi(\frac{h}{2^n}) = 2^n\big(f(a + \frac{h}{2^n}) - f(a)\big).$$

La situation est sympathique quand le résultat obtenu tend vers une fonction linéaire de h quand $n \to \infty$ (vu de près, le graphe est "plat"). La fonction linéaire limite, si elle existe, sera la *différentielle* de la fonction f au point a.

Exemples.

1. Fonction affine. On suppose que la fonction f est de la forme

$$\forall x \in \mathrm{E}, \quad f(x) = c + \ell(x),$$

où c est une constante réelle et où ℓ est une forme linéaire sur E. Le graphe étant déjà plat, on doit s'attendre à ce que le zoom ne fasse aucun effet. En effet, $\varphi(h) = f(a+h) - f(a) = \ell(h)$, et $\varphi_n(h) = 2^n \ell(2^{-n} h) = \ell(h)$ pour tout entier $n \ge 1$ (il y a bien convergence de la suite (φ_n) vers une fonction linéaire ! La "limite" est la fonction linéaire ℓ).

2. Posons $f(x_1, x_2) = x_1^2 + x_1 x_2$, avec $a = (1, 2)$. Alors, en posant $x = (1 + h_1, 2 + h_2)$ on aura

$$\varphi(h_1, h_2) = 2h_1 + h_1^2 + h_1(2 + h_2) + h_2 = 4h_1 + h_2 + h_1^2 + h_1 h_2.$$

Alors

$$\varphi_n(h_1, h_2) = 4h_1 + h_2 + 2^{-n}(h_1^2 + h_1 h_2)$$

converge vers la fonction linéaire $\ell(h_1, h_2) = 4h_1 + h_2$.

3. Définissons une fonction f sur $\mathbb{R}^2$ par

$$f(x, y) = \frac{x^2 y}{x^2 + y^2} \quad \text{si} \quad (x, y) \ne (0, 0), \quad f(0, 0) = 0$$

et posons $a = (0, 0)$. On vérifie que $|f(x, y)| \le |x|$, donc f est continue au point $(0, 0) = a$. On voit aussi que f est positivement homogène, c'est à dire que $f(\lambda x, \lambda y) = \lambda f(x, y)$ si $\lambda > 0$. Il en résulte que $\varphi_n = f$ pour tout n. La fonction limite de la suite (φ_n) existe donc, mais cette fonction limite n'est pas linéaire : la fonction f ne sera pas déclarée différentiable au point $a = (0, 0)$.

Définition 3.2.1. Soient E un espace vectoriel de dimension finie, a un point de E, V un voisinage de a et f une application de V dans $\mathbb{R}$; on dit que f est *différentiable au point a* s'il existe une forme linéaire $\ell \in \mathrm{E}^*$ telle que

$$f(a + h) - f(a) - \ell(h) = o(h)$$

quand $h \to 0_{\mathrm{E}}$.

Si on choisit une norme sur E, on peut encore dire que f est différentiable au point a si la fonction $\varepsilon(h)$ définie pour tout $h \in \mathrm{E}$ tel que $h \ne 0$ et $a + h \in \mathrm{V}$ par la formule

$$\varepsilon(h) = \|h\|^{-1}\big(f(a + h) - f(a) - \ell(h)\big)$$

tend vers 0 lorsque h tend vers 0 dans E. Autrement dit, f est différentiable au point a si et seulement si on peut écrire un développement limité d'ordre 1,

$$f(a+h) = f(a) + \ell(h) + \|h\|\, \varepsilon(h),$$

où $\varepsilon(h) \in \mathbb{R}$ tend vers 0 lorsque h tend vers 0 dans E.

Quand f est différentiable au point a, l'application linéaire ℓ est uniquement déterminée ; en effet, pour tout vecteur $u \in$ E, on a $\ell(u) = \lim_{t\to 0} t^{-1}\big(f(a+tu) - f(a)\big)$, ce qui montre que ℓ est complètement connue. Quand la limite

$$\lim_{t\to 0} \frac{f(a+tu) - f(a)}{t}$$

existe pour un vecteur $u \in$ E, on l'appelle la *dérivée de f au point a dans la direction du vecteur u*.

Il est possible que f admette au point a une dérivée $\mathrm{D}(u)$ dans toute direction u, mais que l'application $u \to \mathrm{D}(u)$ ne soit pas linéaire. C'est ce qui se passe dans l'exemple 3 qui précède. Dans un tel cas, la fonction f n'est pas différentiable au point a.

Notation. On notera $(df)_a$ l'application linéaire ℓ ci-dessus (on note aussi souvent df_a –sans parenthèses– et parfois $d_a f$) ; dans le cas où l'espace E est muni d'une structure euclidienne, on peut aussi représenter cette forme linéaire au moyen du produit scalaire avec un vecteur en utilisant le théorème 2.1.2. On appelle dans ce cas *gradient* de f au point a, noté $\nabla f(a)$, ou $(\nabla f)_a$, l'unique vecteur de l'espace euclidien E tel que

$$(df)_a(x) = \nabla f(a) \,.\, x$$

pour tout $x \in$ E.

Remarque. On a vu que pour tout vecteur $u \in$ E,

$$(df)_a(u) = \lim_{t\to 0} \frac{f(a+tu) - f(a)}{t}.$$

Exemples de différentielles.

1. Si $f : \mathrm{U} \to \mathbb{R}$ est constante, $f(x) = c$, alors $(df)_a = 0_{\mathrm{E}^*}$ pour tout $a \in$ U.

2. Si f est affine, disons $f(x) = c + \ell(x)$, avec ℓ forme linéaire, on voit que $(df)_a = \ell$ pour tout a.

3. Posons $f(x) = x_1^2 + x_2^2 + x_3^2$ pour tout $x \in \mathbb{R}^3$. Pour tous $a = (a_1, a_2, a_3)$ et $h = (h_1, h_2, h_3) \in \mathbb{R}^3$, on calcule $f(a+h) = f(a) + 2\,a\,.\,h + h\,.\,h$. Puisque $h\,.\,h = o(h)$ lorsque $h \to 0$, on voit que

$$(df)_a(h) = 2\,a\,.\,h = 2a_1h_1 + 2a_2h_2 + 2a_3h_3\,;$$

si on calcule la matrice de cette forme linéaire dans la base canonique, on obtient la matrice ligne $\mathrm{A} = \mathrm{mat}((df)_a) = (2a_1 \ 2a_2 \ 2a_2)$. Le *vecteur gradient* $(\nabla f)_a$ est simplement le vecteur $2a$, représenté par la matrice colonne ${}^t\mathrm{A}$ transposée de la matrice ligne précédente.

Plus généralement, si $f(x) = \mathrm{Q}(x)$ est quadratique sur un espace vectoriel réel E, de forme polaire φ, on voit que

$$f(a+h) = \mathrm{Q}(a+h) = \mathrm{Q}(a) + 2\varphi(a,h) + \mathrm{Q}(h).$$

La fonction $\ell(h) = 2\varphi(a,h)$ est une forme linéaire, et on sait qu'il existe une constante M telle que $|\mathrm{Q}(h)| \le \mathrm{M}\,\|h\|^2$, ce qui permet d'écrire $\mathrm{Q}(h) = o(h)$ quand $h \to 0$. On a donc $(df)_a(h) = 2\,\varphi(h,a)$ pour tout $a \in \mathrm{E}$. Si $\mathrm{Q}_0(x) = \|x\|^2$ (norme euclidienne), on obtiendra $(df)_a(h) = 2a\,.\,h$, ce qui permet d'identifier tout de suite le gradient, $\nabla \mathrm{Q}_0(a) = 2a$. Si on effectue une translation de vecteur x_1, on obtient une nouvelle fonction $f_1(x) = \|x - x_1\|^2$, et on trouve que $(\nabla f_1)_a = 2(a - x_1)$ pour tout a.

4. Soit φ une forme bilinéaire sur $\mathrm{F} \times \mathrm{G}$; posons $\mathrm{E} = \mathrm{F} \times \mathrm{G}$ et définissons une fonction f sur E en posant $f(x) = \varphi(y,z)$ si $x = (y,z) \in \mathrm{F} \times \mathrm{G}$. On voit que f est différentiable en tout point $a = (b,c) \in \mathrm{E}$, et

$$(df)_a(k,l) = \varphi(k,c) + \varphi(b,l)$$

pour tout $h = (k,l) \in \mathrm{E}$.

5. Pour la fonction f définie sur $\mathbb{R}^2$ par $f(x) = \mathrm{e}^{x_1 + 2x_2}$, on a vu que

$$f(h) = f(0+h) = 1 + (h_1 + 2h_2) + o(\|h\|)$$

lorsque $h \to 0$, donc $(df)_0(h) = h_1 + 2h_2$. Si on calcule la matrice de $(df)_0$ dans la base canonique de $\mathbb{R}^2$, on obtiendra la matrice ligne $\mathrm{mat}((df)_0) = (1\ 2)$.

Dérivées partielles

Soient $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, U un voisinage de a et $f : \mathrm{U} \to \mathbb{R}$ une fonction réelle différentiable au point a ; l'application $(df)_a$ est une forme linéaire sur $\mathbb{R}^n$, dont on peut calculer la matrice (de taille $1 \times n$) par rapport à la base canonique $\mathbf{e}$ de $\mathbb{R}^n$ (définie par $e_1 = (1, 0, \ldots)$, $\ldots$). Cette matrice contient dans chaque colonne l'image par $(df)_a$ du vecteur de base correspondant. Ainsi, dans la première colonne on a

$$(df)_a(e_1) = \lim_{t \to 0} \frac{f(a_1 + t, a_2, \ldots, a_n) - f(a_1, \ldots, a_n)}{t} \in \mathbb{R}.$$

Il s'agit donc de la dérivée au point $a_1 \in \mathbb{R}$, au sens habituel, de la fonction d'une seule variable $s \to f(s, a_2, \ldots, a_n)$, fonction qui est définie pour s voisin de a_1, et qui est obtenue en fixant les variables $a_2, \ldots, a_n$ et en ne laissant varier que la première coordonnée ; on appelle cette fonction la première *fonction partielle* associée à la fonction donnée f. On appelle la quantité

$$\lim_{t \to 0} \frac{f(a_1 + t, a_2, \ldots, a_n) - f(a_1, \ldots, a_n)}{t}$$

la *première dérivée partielle* de f au point a, que nous noterons $\dfrac{\partial f}{\partial x_1}(a)$. On définit de même les autres dérivées partielles $\dfrac{\partial f}{\partial x_j}(a)$, pour $j = 2, \ldots, n$. Il est parfois commode d'employer une autre notation, $\mathrm{D}_j f(a)$,

$$\mathrm{D}_j f(a) = \frac{\partial f}{\partial x_j}(a).$$

La matrice (ligne) de l'application linéaire $(df)_a$ est donc égale (dans la base canonique de $\mathbb{R}^n$) à

$$(\,\mathrm{D}_1 f(a) \quad \ldots \quad \mathrm{D}_n f(a)\,)$$

et l'image d'un vecteur $h = (h_1, \ldots, h_n)$ s'obtient par produit matriciel usuel ligne-colonne

$$(df)_a(h) = \mathrm{D}_1 f(a)\, h_1 + \cdots + \mathrm{D}_n f(a)\, h_n,$$

de sorte que

$$f(a+h) = f(a) + \mathrm{D}_1 f(a)\, h_1 + \cdots + \mathrm{D}_n f(a)\, h_n + o(h).$$

Considérons maintenant $\mathbb{R}^n$ comme un espace euclidien (avec son produit scalaire usuel). Le vecteur gradient est le vecteur (habituellement représenté comme vecteur colonne) qui a les mêmes composantes, mais dans la base de $\mathbb{R}^n$, au lieu de la base duale (nuance subtile !!!).

Exemple. Revenons sur le cas des formes quadratiques en considérant les coordonnées. Si $f(x) = \mathrm{Q}(x)$ est quadratique sur $\mathbb{R}^n$, de forme polaire φ, et de matrice Φ dans la base canonique, on a en utilisant les coordonnées, si $x = (x_1, \ldots, x_n)$

$$f(x) = \sum_{i,j=1}^{n} \varphi_{i,j} x_i x_j,$$

et si on calcule la dérivée partielle dans la variable x_{i_0} on aura pour chaque $j \neq i_0$ le terme $2\varphi_{i_0,j} x_{i_0} x_j$, qui se dérive en $2\varphi_{i_0,j}\, x_j$, plus le terme $\varphi_{i_0,i_0} x_{i_0}^2$ qui donne $2\varphi_{i_0,i_0} x_{i_0}$, donc finalement

$$\mathrm{D}_{i_0} f(x) = 2 \sum_{j=1}^{n} \varphi_{i_0,j}\, x_j,$$

qui est égal à deux fois la ligne i_0 du produit matriciel $\Phi\mathrm{X}$, donc le gradient de Q au point x a pour représentation en vecteur colonne $2\Phi\mathrm{X}$, où X est la matrice colonne des coordonnées de x. Dans le cas particulier où $\mathrm{Q}(x) = \|x\|^2$, la matrice est I_n et on retrouve que le gradient au point x est égal à $2x$.

On vient de dire que la différentiabilité entraîne l'existence des dérivées partielles. La réciproque n'est pas vraie : on a vu que la fonction f définie sur $\mathbb{R}^2$ par

$$f(x,y) = \frac{x^2 y}{x^2 + y^2}, \quad f(0,0) = 0,$$

n'est pas différentiable au point $a = (0,0)$, pourtant elle a deux dérivées partielles (nulles) au point $a = (0,0)$ et elle a même une dérivée au point a dans toute direction u. On verra un peu plus loin qu'il existe une réciproque lorsque les dérivées partielles sont continues au point a.

On va maintenant donner quelques propriétés de base des fonctions différentiables et des différentielles.

Proposition 3.2.1. *Si f est différentiable au point a, alors f est continue au point a.*

Démonstration. On écrit

$$f(a+h) - f(a) = (df)_a(h) + h\,\varepsilon(h).$$

D'après la proposition 3.1.1 (point 4, puis points 1 et 2), on sait que $(df)_a(h)$ est un $\mathrm{O}(h)$ quand $h \to 0_{\mathrm{E}}$, que $h\,\varepsilon(h) = o(h)$ est aussi un $\mathrm{O}(h)$ et que la somme de deux $\mathrm{O}(h)$ est un $\mathrm{O}(h)$; ce qui permet facilement de conclure que $f(a+h) - f(a) = \mathrm{O}(h)$, donc $f(a+h) - f(a)$ tend vers 0 lorsque $h \to 0_{\mathrm{E}}$.

Proposition 3.2.2. *Si f et g sont deux fonctions réelles différentiables au point a, alors*

1. Pour tous $\lambda, \mu \in \mathbb{R}$, la fonction $\lambda f + \mu g$ est différentiable au point a et

$$\big(d(\lambda f + \mu g)\big)_a = \lambda (df)_a + \mu (dg)_a.$$

2. La fonction produit fg est différentiable au point a, et on a

$$\big(d(fg)\big)_a = (df)_a \, g(a) + f(a) \, (dg)_a.$$

3. Si φ est une fonction à valeurs réelles, définie sur un voisinage de $f(a)$ dans $\mathbb{R}$, et si $\varphi'(f(a))$ existe, la fonction composée $x \in \mathrm{V} \to \varphi(f(x))$ est différentiable au point a, et

$$\big(d(\varphi \circ f)\big)_a = \varphi'(f(a)) \, (df)_a.$$

Démonstration. Le premier point est très facile et laissé au lecteur. Pour les deux autres, il s'agit d'un exercice de DL. Prenons le point 3 par exemple. On va écrire en posant $b = f(a) \in \mathbb{R}$

$$f(a+h) - f(a) - (df)_a(h) = o_1(h), \quad \varphi(b+s) - \varphi(b) - \varphi'(b) \, s = o_2(s).$$

On sait que $s(h) = f(a+h) - f(a)$ est $\mathrm{O}(h)$, donc par composition avec $o_2(s)$ on obtient (proposition 3.1.1) $\varphi(b + s(h)) - \varphi(b) - \varphi'(b) \, s(h) = o(h)$, c'est à dire

$$\varphi(f(a+h)) - \varphi(f(a)) - \varphi'(b) \, s(h) = o(h).$$

En ajoutant $\varphi'(b)(s(h) - (df)_a(h))$ qui est $o(h)$, on obtient finalement

$$\varphi(f(a+h)) - \varphi(f(a)) - \varphi'(b)(df)_a(h) = o(h).$$

Exemple de composition : posons $f(x) = \|x - x_0\|$. Cette fonction n'est pas différentiable au point x_0, mais pour $a \neq x_0$ on peut calculer la différentielle (ou le gradient) en composant la fonction $x \to \|x - x_0\|^2$ d'un exemple précédent avec la fonction φ définie par $\varphi(t) = \sqrt{t}$, définie et dérivable au voisinage de $b = f(a) = \|a - x_0\|^2 > 0$. On obtient ainsi

$$(\nabla f)_a = \frac{a - x_0}{\|a - x_0\|}.$$

Exercice : retrouver le résultat en calculant les dérivées partielles.

3.3. La classe C^1

On a vu plus haut que la différentiabilité en un point a entraîne l'existence des dérivées partielles au point a, mais que la réciproque est fausse. Il est intéressant d'avoir un résultat qui fournisse un passage en sens inverse, conduisant d'hypothèses sur les dérivées partielles à la différentiabilité. Bien entendu, il faut faire des hypothèses plus fortes que la simple existence des dérivées partielles au point a.

Théorème 3.3.1. *Soit V un voisinage d'un point $a \in \mathbb{R}^n$ et soit $f : \mathrm{V} \to \mathbb{R}$; si f admet en tout point b du voisinage V des dérivées partielles $\mathrm{D}_j f(b)$, pour tout $j = 1, \ldots, n$ et si toutes ces fonctions dérivées partielles $b \to \mathrm{D}_j f(b)$ sont continues au point a, la fonction f est différentiable au point a.*

Démonstration. Ecrivons la démonstration dans le cas $n = 2$ pour simplifier. Posons $a = (a_1, a_2) \in \mathbb{R}^2$. Soit $\varepsilon > 0$; d'après l'hypothèse de continuité on peut trouver $\eta > 0$ tel que pour tout vecteur $k \in \mathbb{R}^2$ la condition $\|k\| < \eta$ implique

$$(*) \qquad |\mathrm{D}_i f(a+k) - \mathrm{D}_i f(a)| < \varepsilon, \ i = 1, 2$$

(la norme est la norme usuelle ; on suppose aussi que η est assez petit pour que la condition $\|k\| < \eta$ garantisse que $a + k \in \mathrm{V}$). Fixons un vecteur $h = (h_1, h_2)$ quelconque tel que $\|h\| < \eta$; pour aller de a à $a+h$, on va parcourir un chemin formé de deux segments parallèles aux axes : d'abord un segment horizontal de $a = (a_1, a_2)$ à $a' = (a_1 + h_1, a_2)$, puis un segment vertical de $a' = (a_1 + h_1, a_2)$ à $(a_1 + h_1, a_2 + h_2) = a + h$. On évaluera $f(a + h) - f(a)$ en deux parties, $f(a') - f(a)$ et $f(a + h) - f(a')$ (en dimension n, il faudrait utiliser n segments). Pour effectuer ces deux évaluations, on va appliquer le théorème des accroissements finis (usuel, celui des fonctions réelles de variable réelle) aux deux fonctions auxiliaires d'une seule variable $g_1 : s \to f(a_1 + s, a_2) - \mathrm{D}_1 f(a)\, s$ (où s variera de 0 à h_1, ce qui correspond à un parcours du segment horizontal de a à a') et $g_2 : t \to f(a_1 + h_1, a_2 + t) - \mathrm{D}_2 f(a)\, t$ (où t variera de 0 à h_2, ce qui correspond au segment vertical de a' à $a + h$) ; les dérivées de g_1 et g_2 sont égales à

$$g_1'(s) = \mathrm{D}_1 f(a_1 + s, a_2) - \mathrm{D}_1 f(a) ; \quad g_2'(t) = \mathrm{D}_2 f(a_1 + h_1, a_2 + t) - \mathrm{D}_2 f(a).$$

Si s est entre 0 et h_1 et si on pose $k = (s, 0)$ on aura $\|k\| = \|(s, 0)\| \le \|h\| < \eta$, donc $|g_1'(s)| = |\mathrm{D}_1 f(a + k) - \mathrm{D}_1 f(a)| < \varepsilon$ d'après la condition $(*)$, et de même si t est entre 0 et h_2 on aura $\|(h_1, t)\| \le \|h\| < \eta$ donc $|g_2'(t)| < \varepsilon$ d'après la condition $(*)$. On obtient alors par la majoration des accroissements finis

$$|g_1(h_1) - g_1(0)| = |f(a_1 + h_1, a_2) - f(a) - \mathrm{D}_1 f(a)\, h_1| \le |h_1|\, \varepsilon,$$

$$|g_2(h_2) - g_2(0)| = |f(a_1 + h_1, a_2 + h_2) - f(a_1 + h_1, a_2) - \mathrm{D}_2 f(a)\, h_2| \le |h_2|\, \varepsilon,$$

ce qui montre la différentiabilité de f au point a en additionnant,

$$|f(a + h) - f(a) - \mathrm{D}_1 f(a)\, h_1 - \mathrm{D}_2 f(a)\, h_2| \le (|h_1| + |h_2|)\, \varepsilon \le 2\|h\|\, \varepsilon.$$

Définition 3.3.1. Soit U un ouvert d'un espace vectoriel de dimension finie E ; on dit qu'une fonction f définie sur U, à valeurs réelles est *de classe* C^1 *sur* U si elle est différentiable en tout point de U, et si de plus l'application df, définie sur U, à valeurs dans E^*, qui associe à chaque $a \in \mathrm{U}$ la forme linéaire $(df)_a \in \mathrm{E}^*$ est continue de U dans E^* (dans le cas où on adopte le point de vue euclidien, il revient au même de dire que l'application gradient $a \to (\nabla f)_a$ est continue de U dans E).

Corollaire 3.3.1. *Soit* U *un ouvert de* $\mathbb{R}^n$ *et soit* $f : \mathrm{U} \to \mathbb{R}$ *; la fonction* f *est de classe* C^1 *sur* U *si et seulement si* f *admet en tout point de* U *des dérivées partielles* $\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}$*, et si toutes ces dérivées partielles sont continues dans* U.

Démonstration. Si f a des dérivées partielles continues dans U, f sera différentiable dans l'ouvert U d'après le théorème 3.3.1. De plus, l'application vectorielle df, dont l'expression en coordonnées est

$$a \in \mathrm{U} \to (df)_a = \Big(\frac{\partial f}{\partial x_1}(a), \ldots, \frac{\partial f}{\partial x_n}(a)\Big)$$

est continue si et seulement si les n fonctions coordonnées $a \to \frac{\partial f}{\partial x_j}(a)$ sont continues, donc la continuité des dérivées partielles implique la continuité de la différentielle.

Inversement, si on sait que f est de classe C^1 sur U, elle est différentiable en tout point de $\mathrm{U} \subset \mathbb{R}^n$, les dérivées partielles existent donc en tout point de U, et elles sont continues en tant que fonctions coordonnées de la fonction vectorielle continue $a \to (df)_a$.

Condition nécessaire pour un extremum local

Définition 3.3.2. Soit E un espace vectoriel réel de dimension finie et soit f une fonction à valeurs réelles définie dans un voisinage du point $a \in \mathrm{E}$; on dit que f admet un *maximum local au point* a s'il existe un voisinage W de a dans lequel f est majorée par la valeur $f(a)$,

$$\forall x \in \mathrm{W}, \quad f(x) \leq f(a).$$

On définit de même la notion de *minimum local*. On parle d'*extremum local* quand on a soit un maximum local, soit un minimum local au point a.

Proposition 3.3.1. *Si une fonction réelle f, définie sur un voisinage $\mathrm{V} \subset \mathbb{R}^n$ d'un point $a \in \mathbb{R}^n$, admet des dérivées partielles au point a et si f admet un extremum local au point a, alors*

$$\mathrm{D}_1 f(a) = \mathrm{D}_2 f(a) = \cdots = \mathrm{D}_n f(a) = 0.$$

Si f est différentiable au point $a \in \mathrm{E}$ et admet un extremum local au point a, on a

$$(df)_a = 0_{\mathrm{E}^*}.$$

Démonstration. Traitons le cas d'un maximum local. Soit e_i l'un des vecteurs de la base canonique de $\mathbb{R}^n$, $1 \leq i \leq n$; pour t assez petit, le point $a+te_i$ se trouve dans le voisinage W où f est majorée par $f(a)$, donc la fonction d'une variable réelle $t \to f(a+te_i)$ admet un maximum local pour $t = 0$. Il en résulte que sa dérivée s'annule pour $t = 0$, ce qui donne

$$0 = \lim_{t \to 0} \frac{f(a+te_i) - f(a)}{t} = \mathrm{D}_i f(a).$$

Traitons ensuite le cas “abstrait” où f est définie sur un voisinage V d'un point a d'un espace vectoriel E de dimension finie, et admet un maximum local au point a. Soit h un vecteur quelconque de E ; pour t assez petit, le point $a + th$ se trouve dans le voisinage W où f est majorée par $f(a)$, donc la fonction $t \to f(a + th)$ admet un maximum local pour $t = 0$. Il en résulte que sa dérivée s'annule pour $t = 0$, ce qui donne

$$0 = \lim_{t \to 0} \frac{f(a+th) - f(a)}{t} = (df)_a(h).$$

Ceci étant vrai pour tout vecteur $h \in \mathrm{E}$, on a bien $(df)_a = 0_{\mathrm{E}^*}$.

Attention !! La fonction f définie par $f(x) = x$ atteint son maximum sur $[0, 1]$ au point 1, mais bien sûr la dérivée n'a aucune raison d'être nulle, et effectivement elle n'est pas nulle. Le même phénomène se produit en plusieurs variables : une fonction définie sur un disque *fermé* peut très bien atteindre son maximum (ou son minimum) sur le disque fermé en un point du bord du disque sans que la différentielle s'annule en ce point.

Recherche d'extrema pour les fonctions réelles de plusieurs variables

On suppose donnée une fonction réelle f définie sur un fermé F, et on cherche à trouver son maximum (absolu) ou bien son minimum sur F. On utilise à cet effet deux ingrédients principaux.

Si une fonction réelle f définie sur un voisinage de $a \in \mathbb{R}^n$ admet une dérivée partielle $\mathrm{D}_i f(a)$ au point a et si f admet un extremum local au point a, on a

$$\mathrm{D}_i f(a) = 0.$$

Le second ingrédient se trouve dans le poly de MT241.

Si f est une fonction réelle continue définie sur un compact K *non vide de* $\mathbb{R}^n$*, elle atteint son minimum sur* K*. Si une fonction réelle continue f définie sur un fermé* F *non vide et non borné de* $\mathbb{R}^n$ *tend vers* $+\infty$ *lorsque* $\|x\| \to +\infty$ *(par valeurs* $x \in \mathrm{F}$*), alors f atteint son minimum sur* F*.*

Quand on a de la chance ces deux principes suffisent pour identifier le minimum (ou maximum) de certaines fonctions de plusieurs variables. En effet, si le deuxième principe permet de savoir que f atteint par exemple son minimum sur le fermé F, et si un argument auxiliaire permet de montrer que le minimum n'est pas atteint sur le bord de F, le minimum sera alors atteint en un point a intérieur à F, c'est à dire un point a tel que F soit un voisinage de a. Mais si a est minimum absolu de f, c'est alors un minimum local et les dérivées partielles doivent s'annuler au point a. On cherchera donc les candidats parmi les *points critiques* de f, c'est à dire les points où la différentielle s'annule. S'il n'y a qu'un petit nombre de points critiques, il suffira de calculer les valeurs de f en ces points pour trouver le point de F où le minimum de f est atteint.

Exemples et exercices.

1. On considère $\alpha, \beta, \gamma > 0$ tels que $\alpha + \beta + \gamma = 1$, et un nombre $c > 1$. Maximiser la fonction
$$f(x, y, z) = x^\alpha y^\beta z^\gamma - \alpha x^c - \beta y^c - \gamma z^c$$
sur le fermé $\mathrm{F} = \{(x, y, z) : x \geq 0, y \geq 0, z \geq 0\}$ de $\mathbb{R}^3$.

On constate que f tend vers $-\infty$ à l'infini sur F, donc le maximum existe, que $f \leq 0$ "au bord" de F, alors qu'il existe des points de F où $f > 0$, donc ce maximum est atteint à l'intérieur de F (en un point où $x > 0$, $y > 0$ et $z > 0$), et peut par conséquent être trouvé en annulant les dérivées partielles. Par chance, on trouve un seul candidat qui annule les dérivées partielles, $x = y = z = c^{1/(1-c)}$. Ce point réalise donc le maximum de la fonction f sur F.

2. Soient a, b, c trois points de $\mathbb{R}^2$, et α, β, γ trois nombres réels > 0 ; minimiser l'expression
$$f(x) = \alpha \|x - a\|^2 + \beta \|x - b\|^2 + \gamma \|x - c\|^2.$$
Ici il est plus agréable de raisonner globalement, par exemple en annulant le gradient de la fonction f.

Plus généralement soient $a^{(1)}, \ldots, a^{(k)}$ des points d'un espace euclidien E, et $c_1, \ldots, c_k$ des nombres > 0 ; minimiser l'expression
$$f(x) = \sum_{i=1}^{k} c_i \, \|x - a^{(i)}\|^2.$$

3. Exemple de recherche de minimum. Soient $x^{(1)}, x^{(2)}$ et $x^{(3)}$ trois points du plan ; minimiser l'expression
$$f(x) = \sum_{i=1}^{3} \|x - x^{(i)}\|.$$
Ce problème consiste à chercher le réseau de longueur minimale qui permette de relier les trois points donnés. Cette fonction f est différentiable en tout point a différent de $x^{(1)}$, $x^{(2)}$

ou $x^{(3)}$. Si le minimum de f est atteint en un point $a \notin \{x^{(1)}, x^{(2)}, x^{(3)}\}$, on aura $(df)_a = 0$ ou bien $(\nabla f)_a = 0$, ce qui donne

$$0 = \sum_{i=1}^{3} \frac{a - x^{(i)}}{\|a - x^{(i)}\|},$$

et une somme de trois vecteurs de norme un ne peut être nulle que s'ils forment des angles de $2\pi/3$. Cette information aide à résoudre le problème. Si le triangle ayant ces trois points pour sommets a un des angles plus grand que $2\pi/3$, le réseau minimal consiste à prendre pour a celui des trois sommets pour lequel l'angle est $> 2\pi/3$. Le "réseau minimal" est formé des deux côtés adjacents à ce sommet. Si tous les angles sont $< 2\pi/3$, le réseau minimal est obtenu en trouvant l'unique point a intérieur au triangle tel que les trois vecteurs $x^{(i)} - a$ fassent des angles égaux à $2\pi/3$ (il existe une construction géométrique très simple de ce point, proposée au 17ème siècle par Torricelli et Cavalieri).

Essayer de généraliser au problème du réseau de longueur minimale, par exemple pour les sommets d'un carré.

4. Méthode de pénalités. On suppose $f \geq 0$ sur $\mathbb{R}^2$, différentiable en tout point, et telle que les deux dérivées partielles soient continues. On veut minimiser f sur la boule unité fermée $\mathrm{B}_1 = \{x : \|x\| \leq 1\}$ de $\mathbb{R}^2$ (puisque f est continue sur un fermé borné, on sait que le minimum sur B_1 est atteint, mais on ne peut pas dire à coup sûr que la différentielle y sera nulle, si le minimum est atteint sur le cercle unité). Pour tout entier $n > 1$ on introduit le problème

$$\min\{f(x) + n\mathrm{D}(x)^2 : x \in \mathbb{R}^2\}$$

où $\mathrm{D}(x) = \min\{\|x\| - 1, 0\}$ est la distance de x au disque unité. Ce problème a une solution optimale $a^{(n)} \in \mathbb{R}^2$ par Bolzano-Weierstrass. Montrer que la suite $(a^{(n)})$ admet une sous-suite qui converge vers un point $a \in \mathrm{B}_1$ où f atteint son minimum sur B_1. Si le minimum est atteint en un point a du cercle unité, montrer que $(\nabla f)_a$ est normal au cercle, et dirigé vers l'extérieur du disque.

5. Maximiser sur $\mathbb{R}^3$ la fonction

$$(x, y, z) \to ux + vy + wz - \frac{1}{p}(|x|^p + |y|^p + |z|^p),$$

où $p > 1$. En déduire l'inégalité de Hölder (dans un cas particulier)

$$|ux + vy + wz| \leq (|x|^p + |y|^p + |z|^p)^{1/p}(|u|^q + |v|^q + |w|^q)^{1/q},$$

où q est défini par $1/p + 1/q = 1$.

6. Inégalité arithmético-géométrique. Soient $x_1, \ldots, x_n$ des nombres réels ≥ 0, et $\alpha_1, \ldots, \alpha_n$ des nombres réels ≥ 0 tels que $\sum_{i=1}^n \alpha_i = 1$. Montrer que

$$\prod_{i=1}^{n} x_i^{\alpha_i} \leq \sum_{i=1}^{n} \alpha_i x_i.$$

Petite formule des accroissements finis

Si a et b sont deux points de E, on désigne par $[a, b]$ le segment de E qui joint ces deux points, c'est à dire l'ensemble des points de la droite affine passant par a et b, et qui sont compris entre a et b. On peut paramétrer tous ces points c sous la forme

$$c(t) = (1 - t)a + tb, \quad 0 \leq t \leq 1.$$

Pour $t = 0$, on a $c(0) = a$ et pour $t = 1$, on a $c(1) = b$. Le résultat suivant sera vu plus plus loin dans le cas général d'applications entre deux espaces vectoriels, mais il n'est peut être pas inutile de voir d'abord le cas plus simple des fonctions à valeurs réelles.

Théorème 3.3.2. *Soient* E *un espace vectoriel normé de dimension finie,* $a, b \in \mathrm{E}$ *et* U *un ouvert de* E *contenant le segment* $[a, b]$ *; soit* $f : \mathrm{U} \to \mathbb{R}$ *une fonction différentiable en tout point de* $[a, b]$ *et soit* M *un nombre réel tel que* $\|(df)_x\| \leq \mathrm{M}$ *pour tout* $x \in [a, b]$ *(la norme utilisée sur* E^* *est la norme duale de celle de* E*). On a alors l'inégalité*

$$|f(b) - f(a)| \leq \mathrm{M}\,\|b - a\|.$$

Démonstration. On étudie la variation de la fonction g d'une seule variable définie par

$$g(t) = f(a + t(b - a))$$

lorsque t varie de 0 à 1. Sa dérivée en un point $t \in]0, 1[$ est égale à $(df)_{a+t(b-a)}\,(b - a)$, que l'on majore par $\|(df)_{a+t(b-a)}\|\,\|b-a\| \leq \mathrm{M}\,\|b-a\|$. La majoration des accroissements finis dans le cas réel, appliquée à la fonction g sur l'intervalle $[0, 1]$ donne ensuite

$$|f(b) - f(a)| = |g(1) - g(0)| \leq \mathrm{M}\,\|b - a\|,$$

ce qui est le résultat cherché.

On dit qu'un sous-ensemble C de E est *convexe* si pour tous points $a, b \in \mathrm{C}$, le segment $[a, b]$ est contenu dans C.

Corollaire 3.3.2. *Soient* U *un ouvert convexe de* E *et* $f : \mathrm{U} \to \mathbb{R}$ *telle que* $\|(df)_x\| \leq \mathrm{M}$ *pour tout* $x \in \mathrm{U}$ *; alors* f *vérifie la propriété suivante :*

$$\forall x, y \in \mathrm{U}, \quad |f(y) - f(x)| \leq \mathrm{M}\,\|y - x\|$$

(c'est à dire que f *est lipschitzienne de constante* M *dans* U*)*

Démonstration. C'est une application immédiate du résultat précédent : si x, y sont deux points de U, le segment $[x, y]$ tout entier est contenu dans U puisque U est convexe.

Corollaire 3.3.3. *Soient* U *un ouvert convexe d'un espace vectoriel* E *de dimension finie et* $f : \mathrm{U} \to \mathbb{R}$ *telle que* $(df)_x = 0$ *pour tout* $x \in \mathrm{U}$ *; alors* f *est constante sur* U.

Démonstration. C'est clair à partir du corollaire précédent.

Le résultat précédent est évidemment faux si l'ouvert de définition est formé de deux morceaux disjoints : la fonction peut alors être égale à 1 sur un morceau et à 0 sur l'autre. On introduira plus tard la notion *d'ouvert connexe*, c'est à dire en gros qui n'a qu'un seul bout, et le résultat précédent est valable pour les ouverts connexes.

3.4. La différentielle dans le cas général

On va maintenant étudier les applications f définies sur une partie U d'un espace vectoriel réel de dimension finie E, et à valeurs dans un deuxième espace vectoriel réel de dimension finie F. Il n'y a pas de véritable nouveauté conceptuelle dans ce paragraphe, mais il faut introduire quelques adaptations des notions déjà introduites.

Le cas des valeurs vectorielles : premier contact

Supposons que $a \in \mathbb{R}^3$, que V soit un voisinage de a et f une fonction définie sur V et à valeurs dans $\mathbb{R}^2$. L'image $f(x_1, x_2, x_3)$ a donc deux composantes, que l'on écrira $f_1(x_1, x_2, x_3)$ et $f_2(x_1, x_2, x_3)$. Supposons que ces deux fonctions composantes soient différentiables au point a. On peut écrire deux DL, en posant $h = (h_1, h_2, h_3) \in \mathbb{R}^3$

$$\begin{aligned} f_1(a+h) = &\quad f_1(a_1,a_2,a_3) + \tfrac{\partial f_1}{\partial x_1}(a)h_1 \quad + \tfrac{\partial f_1}{\partial x_2}(a)h_2 \quad + \tfrac{\partial f_1}{\partial x_3}(a)h_3 \quad + o_1(h) \\ f_2(a+h) = &\quad f_2(a_1,a_2,a_3) + \tfrac{\partial f_2}{\partial x_1}(a)h_1 \quad + \tfrac{\partial f_2}{\partial x_2}(a)h_2 \quad + \tfrac{\partial f_2}{\partial x_3}(a)h_3 \quad + o_2(h) \end{aligned}$$

En regroupant les deux fonctions $o_j(h)$, $j = 1, 2$ dans une erreur vectorielle

$$\overrightarrow{o(h)} = (o_1(h), o_2(h)),$$

on peut écrire les relations ci-dessus sous forme vectorielle

$$f(a+h) = f(a) + \mathrm{MH} + \overrightarrow{o(h)},$$

où H est le vecteur colonne des trois coordonnées de h et M la matrice de taille 2×3 donnée par

$$\mathrm{M} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \frac{\partial f_1}{\partial x_2}(a) & \frac{\partial f_1}{\partial x_3}(a) \\ \frac{\partial f_2}{\partial x_1}(a) & \frac{\partial f_2}{\partial x_2}(a) & \frac{\partial f_2}{\partial x_3}(a) \end{pmatrix}.$$

L'application qui associe à chaque vecteur $h \in \mathbb{R}^3$ le vecteur $\mathrm{MH} \in \mathbb{R}^2$ est une application linéaire de $\mathbb{R}^3$ dans $\mathbb{R}^2$: c'est cette application qui sera appelée *différentielle au point a* dans le cas des valeurs vectorielles. Tout ceci nous conduit à la définition qui suit.

Définition 3.4.1. Soient E et F deux espaces vectoriels réels de dimension finie, a un point de E, V un voisinage de a et f une application de V dans F ; on dit que f est différentiable au point a s'il existe une application linéaire ℓ de E dans F telle que

$$f(a+h) - f(a) - \ell(h) = \overrightarrow{o}(h)$$

lorsque $h \to 0_{\mathrm{E}}$.

Si E est normé, cela revient à dire que f est différentiable au point a si et seulement si on peut écrire un développement limité d'ordre 1 de la forme

$$f(a+h) = f(a) + \ell(h) + \|h\|_{\mathrm{E}}\, \varepsilon(h),$$

où $\varepsilon(h) \in \mathrm{F}$ tend vers 0 (dans F) lorsque h tend vers 0 dans E.

Lorsque f est différentiable au point a, on appelle *différentielle de f au point a* l'application linéaire ℓ ci-dessus, et on la notera encore $(df)_a$. On remarquera bien que $(df)_a \in \mathcal{L}(\mathrm{E}, \mathrm{F})$ dans le cas présent. On a encore la relation qui permet de calculer $(df)_a(u)$ (qui est un élément de F) pour tout vecteur $u \in \mathrm{E}$,

$$(df)_a(u) = \lim_{t \to 0} \frac{f(a+tu) - f(a)}{t},$$

cette limite étant maintenant calculée dans l'espace F (au lieu de $\mathbb{R}$).

Supposons que $\mathrm{E} = \mathbb{R}$. Dans ce cas l'application $(df)_a$ définie ci-dessus est une application linéaire de $\mathbb{R}$ dans F. Alors en prenant $h = 1 \in \mathbb{R}$ on obtient

$$(df)_a(1) = \lim_{t \to 0} \frac{f(a+t) - f(a)}{t} \in \mathrm{F}$$

qui est le *vecteur dérivé* usuel $\overrightarrow{f'(a)}$ (dans cette situation il est plus habituel de s'intéresser au vecteur dérivé $\overrightarrow{v}$ plutôt qu'à l'application linéaire associée, $h \in \mathbb{R} \to h\overrightarrow{v}$; il est aussi traditionnel de parler de fonction dérivable, plutôt que différentiable, dans ce cas).

En termes pompeux, on pourra dire que $\mathcal{L}(\mathbb{R}, \mathrm{F})$ (l'espace qui contient la différentielle $(df)_a$) est canoniquement isomorphe à F (l'espace qui contient le vecteur dérivé) au moyen de l'application $\ell \to \ell(1)$.

Proposition 3.4.1. *Soit f une fonction définie au voisinage d'un point $a \in \mathrm{E}$ et à valeurs dans $\mathbb{R}^m$. Pour que f soit différentiable au point a, il faut et il suffit que toutes les fonctions coordonnées $f_1, \dots, f_m$ soient différentiables au point a.*

Si une fonction f est définie au voisinage de $a \in \mathbb{R}^n$ et à valeurs dans $\mathbb{R}^m$, elle est donc différentiable au point a si et seulement si toutes les fonctions coordonnées $f_1, \dots, f_m$ sont différentiables au point a, ce qui permet d'obtenir l'énoncé suivant en utilisant le théorème 3.3.1 : si toutes les dérivées partielles $\mathrm{D}_j f_i$, $i = 1, \dots, m$ et $j = 1, \dots, n$ existent dans un voisinage de a et sont continues au point a, alors l'application f est différentiable au point a.

Exemples de différentielles.

1. Application linéaire ou affine. Si $f(x) = c + \ell(x)$, où c est un vecteur fixé de F et ℓ une application linéaire de E dans F, on a $(df)_a = \ell$ pour tout point $a \in \mathrm{E}$.

2. Soit φ une application bilinéaire sur $\mathrm{F} \times \mathrm{G}$ à valeurs dans un troisième espace vectoriel L ; posons $\mathrm{E} = \mathrm{F} \times \mathrm{G}$ et définissons une application f de E dans L en posant $f(x) = \varphi(y, z)$ si $x = (y, z) \in \mathrm{F} \times \mathrm{G}$. On voit que f est différentiable en tout point $a = (b, c) \in \mathrm{E}$, et

$$(df)_a(u, v) = \varphi(u, c) + \varphi(b, v)$$

pour tout $h = (u, v) \in \mathrm{E}$.

Il faut parfois réfléchir un peu pour reconnaître la structure bilinéaire d'une fonction donnée. Si $f(x_1, x_2, x_3, x_4) = x_1x_4 + 2x_2x_4 - x_3x_4$, on voit que f peut s'écrire $\varphi(y, z)$, avec φ bilinéaire, en posant $y = (x_1, x_2, x_3) \in \mathrm{F} = \mathbb{R}^3$ et $z = x_4 \in \mathrm{G} = \mathbb{R}$. On trouve alors au point $a = (a_1, a_2, a_3, a_4) = (b, c)$, avec $b = (a_1, a_2, a_3)$ et $c = a_4$

$$(df)_a(h_1, h_2, h_3, h_4) = \varphi(b, v) + \varphi(u, c) = (a_1 + 2a_2 - a_3)h_4 + (h_1 + 2h_2 - h_3)a_4$$

où on a posé $h = (u, v)$, $u = (h_1, h_2, h_3)$ et $v = h_4$.

Matrice jacobienne dans le cas de $\mathbb{R}^n$ et $\mathbb{R}^m$

Soit f une application définie dans un voisinage de $a \in \mathbb{R}^n$, à valeurs dans $\mathbb{R}^m$; si on pose $f(x) = (f_1(x), \dots, f_m(x)) \in \mathbb{R}^m$, la *matrice jacobienne* au point a est la matrice de l'application linéaire $(df)_a$ (de $\mathbb{R}^n$ dans $\mathbb{R}^m$) par rapport aux bases canoniques,

$$\mathrm{J}_a(f) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \dots & \frac{\partial f_1}{\partial x_n}(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(a) & \dots & \frac{\partial f_m}{\partial x_n}(a) \end{pmatrix}.$$

Chaque ligne de la matrice jacobienne de f au point a correspond à la différentielle au point a de l'une des fonctions coordonnées $f_1, \ldots, f_m$.

Exercice. Calculons la différentielle de l'application f de $\mathrm{M}_2(\mathbb{R})$ dans $\mathrm{M}_2(\mathbb{R})$ définie par $\mathrm{A} \to \mathrm{A}^2$. On écrit

$$(\mathrm{A}+\mathrm{H})^2 = \mathrm{A}^2 + (\mathrm{AH}+\mathrm{HA}) + \mathrm{H}^2$$

ce qui donne immédiatement

$$(df)_{\mathrm{A}}(\mathrm{H}) = \mathrm{AH}+\mathrm{HA}.$$

Le lecteur pourra essayer d'écrire de la matrice jacobienne, histoire de voir qu'il y a des cas où il ne vaut mieux pas raisonner en coordonnées !

Exercice. Différentielle de l'application $u \to u^{-1}$.

Rappelons que si B est une matrice de taille $n \times n$ telle que $\|\mathrm{B}\| < 1$, la matrice $\mathrm{I}_n - \mathrm{B}$ est inversible et $(\mathrm{I}_n - \mathrm{B})^{-1} = \sum_{k=0}^{+\infty} \mathrm{B}^k$. En particulier, la relation

$$(\mathrm{I}_n - \mathrm{B})^{-1} - (\mathrm{I}_n + \mathrm{B}) = \sum_{k=2}^{+\infty} \mathrm{B}^k$$

montre que

$$\|(\mathrm{I}_n - \mathrm{B})^{-1} - (\mathrm{I}_n + \mathrm{B})\| \le \sum_{k=2}^{+\infty} \|\mathrm{B}\|^k = \frac{\|\mathrm{B}\|^2}{1-\|\mathrm{B}\|} = \mathrm{O}(\|\mathrm{B}\|^2)$$

lorsque $\mathrm{B} \to 0$. Soit A une matrice carrée de taille $n \times n$, inversible ; si K est une autre matrice de même taille, et telle que $\|\mathrm{K}\| < \|\mathrm{A}^{-1}\|^{-1}$, on vérifie que $\mathrm{A}+\mathrm{K}$ est inversible en écrivant $(\mathrm{A}+\mathrm{K}) = \mathrm{A}(\mathrm{I}_n + \mathrm{A}^{-1}\mathrm{K})$. En utilisant ce qui précède on trouve que

$$(\mathrm{A}+\mathrm{K})^{-1} = \mathrm{A}^{-1} - \mathrm{A}^{-1}\mathrm{K}\mathrm{A}^{-1} + \mathrm{O}(\|\mathrm{K}\|^2),$$

ce qui indique que la différentielle de l'application "matrice inverse" $\mathrm{M} \to \mathrm{M}^{-1}$, calculée au point A, est l'application linéaire $\mathrm{K} \to -\mathrm{A}^{-1}\mathrm{K}\mathrm{A}^{-1}$ (endomorphisme de $\mathrm{M}_n(\mathbb{R})$).

Théorème 3.4.1. Composition de deux applications différentiables. *Soient* V *un voisinage de* $a \in \mathrm{E}$, W *un voisinage de* $b \in \mathrm{F}$, $f : \mathrm{V} \to \mathrm{W}$ *une application différentiable au point* a. *Si* $f(a) = b$ *et si* $g : \mathrm{W} \to \mathrm{G}$ *est différentiable au point* b, *alors* $g \circ f$ *est différentiable au point* a *et*

$$\big(d(g \circ f)\big)_a = (dg)_b \circ (df)_a.$$

Démonstration. Au niveau conceptuel, la démonstration est exactement la même que celle qui a été donnée lorsque f est à valeurs réelles. On écrit

$$f(a+h) - f(a) - (df)_a(h) = o_1(h), \quad g(b+k) - g(b) - (dg)_b(k) = o_2(k)$$

lorsque $h \to 0_{\mathrm{E}}$ et $k \to 0_{\mathrm{F}}$. On sait que $(df)_a(h) = \mathrm{O}(h)$ donc $k(h) = f(a+h) - f(a)$ est $\mathrm{O}(h)$, donc par composition avec $o_2(k)$ on obtient (d'après la proposition 3.1.1) $g(b+k(h)) - g(b) - (dg)_b\,(k(h)) = o(h)$, c'est à dire

$$g(f(a+h)) - g(f(a)) - (dg)_b\,(k(h)) = o(h).$$

En ajoutant $(dg)_b\big(k(h) - (df)_a(h)\big)$ qui est $o(h)$, on obtient finalement

$$g(f(a+h)) - g(f(a)) - (dg)_b\big((df)_a(h)\big) = o(h).$$

Produit de matrices jacobiennes. Si f et g sont deux applications différentiables, de $\mathbb{R}^n$ dans $\mathbb{R}^p$ et de $\mathbb{R}^p$ dans $\mathbb{R}^q$ respectivement, on aura

$$\mathrm{J}_a(g \circ f) = \mathrm{J}_{f(a)}(g)\,\mathrm{J}_a(f).$$

Fonctions vectorielles d'une variable

Supposons que φ soit une application d'un intervalle ouvert J de $\mathbb{R}$, contenant un point t_0, et à valeurs dans un espace vectoriel réel E de dimension finie. On définit la dérivée de φ au point $t_0 \in \mathbb{R}$ par une formule voisine de celle qui définit la dérivée usuelle, en posant

$$\varphi'(t_0) = \lim_{s \to 0} \frac{\varphi(t_0 + s) - \varphi(t_0)}{s}.$$

On dit que $\varphi'(t_0) \in \mathrm{E}$ est le *vecteur dérivé* de φ au point t_0 (la limite est prise dans l'espace E).

Supposons que $\mathrm{E} = \mathbb{R}^2$ par exemple, et considérons les deux coordonnées du vecteur $\varphi(t) \in \mathbb{R}^2$, ce qui nous donne $\varphi(t) = (\varphi_1(t), \varphi_2(t))$. La limite vectorielle ci-dessus existe si et seulement si les deux coordonnées convergent vers les coordonnées correspondantes de la limite. On voit ainsi que φ est dérivable au point t_0 si et seulement si *les deux* fonctions coordonnées φ_1 et φ_2 sont dérivables au point t_0, et dans ce cas le vecteur dérivé a pour coordonnées les dérivées des deux coordonnées, $\varphi'(t) = (\varphi_1'(t_0), \varphi_2'(t_0))$. Il est facile de voir que φ est dérivable au point t_0 si et seulement si elle est différentiable au point t_0, et on a vu que ces deux notions sont reliées par la formule $\varphi'(t_0) = (d\varphi)_{t_0}(1)$.

Exemple. Posons $\varphi(t) = (\cos t, \sin t) \in \mathbb{R}^2$. On peut écrire un développement limité de tout ordre donné n des deux fonctions coordonnées au voisinage d'un point $t_0 \in \mathbb{R}$. On obtient ainsi par exemple à l'ordre deux

$$\begin{array}{rclllll} \cos(t_0 + s) & = & \cos(t_0) & -s\sin(t_0) & -\frac{s^2}{2}\cos(t_0) & +s^2\varepsilon_1(s) \\ \sin(t_0 + s) & = & \sin(t_0) & +s\cos(t_0) & -\frac{s^2}{2}\sin(t_0) & +s^2\varepsilon_2(s) \end{array}$$

qu'il est intéressant de regrouper vectoriellement

$$\mathrm{V}(t_0 + s) = \mathrm{V}(t_0) + s\mathrm{V}_1(t_0) + \frac{s^2}{2}\mathrm{V}_2(t_0) + s^2\overrightarrow{\varepsilon(s)},$$

et on note que le terme d'erreur $\overrightarrow{\varepsilon(s)}$ est *vectoriel*. En termes cinématiques, si on considère que $\varphi(t) = (\cos t, \sin t)$ décrit un mouvement sur une cercle de rayon un, la première dérivée $\varphi'(t) = (-\sin t, \cos t)$ est la vitesse ; elle indique la direction de la tangente à la courbe. La seconde dérivée $\varphi''(t) = -(\cos t, \sin t)$ est l'accélération. Dans cet exemple, elle est dirigée vers l'origine.

Composition de $\varphi : \mathbb{R} \to \mathrm{E}$ *et* $f : \mathrm{V} \to \mathrm{F}$

Proposition 3.4.2. *Soit* $\varphi : \mathrm{I} \to \mathrm{E}$ *une courbe tracée dans* E, *prenant ses valeurs dans un voisinage* V *du point* $a_0 = \varphi(t_0)$, I *étant un intervalle ouvert de* $\mathbb{R}$. *Si* φ *est dérivable au point* $t_0 \in \mathrm{I}$ *et* $f : \mathrm{V} \to \mathrm{F}$ *différentiable au point* $a_0 = \varphi(t_0)$, *la composée est dérivable en* t_0 *et*

$$(f \circ \varphi)'(t_0) = (df)_{\varphi(t_0)}(\varphi'(t_0)).$$

Si $\mathrm{E} = \mathbb{R}^2$ par exemple, on obtient avec les dérivées partielles,

$$(f \circ \varphi)'(t_0) = \frac{\partial f}{\partial x_1}(\varphi(t_0))\,\varphi_1'(t_0) + \frac{\partial f}{\partial x_2}(\varphi(t_0))\,\varphi_2'(t_0),$$

où $\varphi(t) = (\varphi_1(t), \varphi_2(t))$. Dans le cas où E est muni d'une structure euclidienne, on peut préférer exprimer le résultat avec le gradient,

$$(f \circ \varphi)'(t_0) = \nabla f_{\varphi(t_0)} \,.\, \varphi'(t_0).$$

Démonstration. Il s'agit simplement d'interpréter le théorème général sur la composition. La différentielle de la composée est $(df)_{a_0} \circ (d\varphi)_{t_0}$. Puisque la fonction $f \circ \varphi$ est définie sur un intervalle réel, on sait que son vecteur dérivé s'obtient à partir de la différentielle en écrivant

$$(f \circ \varphi)'(t_0) = \big((df)_{a_0} \circ (d\varphi)_{t_0}\big)(1) = (df)_{a_0}\big((d\varphi)_{t_0}(1)\big) = (df)_{a_0}(\varphi'(t_0)).$$

Il est important de savoir traiter le cas particulier où φ est affine. Supposons que $\varphi(t) = a_0 + th$, où h est un certain vecteur de E. On a alors $\varphi'(t) = h$ pour tout t, et la formule précédente nous redonne la dérivée dans la direction h

$$\frac{d}{dt} f(a_0 + th)_{|t=0} = (df)_{a_0}(h) = (\nabla f)_{a_0} \,.\, h.$$

Cette formule permet de comprendre la tendance de variation de f pour les points voisins de a_0. Si $(\nabla f)_{a_0} \,.\, h > 0$, la fonction f "augmente" dans la direction h, dans le sens précis suivant : il existe $t_0 > 0$ tel que pour tout t vérifiant $0 < t < t_0$,

$$f(a_0 + th) > f(a_0).$$

Cependant, cela n'entraîne pas en général que la fonction $t \to f(a + th)$ soit croissante dans aucun intervalle $]0, t_1[$ (pour $t_1 > 0$).

Lignes de niveau

Soit $x = (x_1, x_2) \in \mathbb{R}^2 \to f(x_1, x_2)$ une fonction réelle de deux variables, différentiable en tout point ; l'ensemble Γ_c des points $x \in \mathbb{R}^2$ tels que $f(x) = c$ s'appelle une *ligne de niveau* de la fonction f. *Dans les bons cas*, cette ligne est une brave courbe que l'on peut paramétrer par une fonction vectorielle $t \to \varphi(t) \in \Gamma_c$. On peut alors dire que puisque la fonction $t \to f(\varphi(t))$ est constante égale à c, sa dérivée est nulle,

$$\nabla f_{\varphi(t)} \,.\, \varphi'(t) = 0.$$

Toujours dans les bons cas, on aura $\varphi'(t) \neq 0$ et dans ce cas le vecteur $\varphi'(t)$ indique la direction de la tangente à la courbe de niveau au point $\varphi(t)$. Si $x \in \Gamma_c$, le vecteur ∇f_x est donc en général orthogonal à la tangente à la courbe, c'est donc un *vecteur normal* à la courbe de niveau au point x. D'une façon générale, on peut dire que le gradient est orthogonal aux lignes de niveau. Faisons un résumé du cas optimiste :

Proposition 3.4.3. *Supposons qu'il existe un voisinage* V *de a et une fonction $\varphi : \mathrm{I} \to \mathrm{V}$ telle que $\varphi(t_0) = a$ et qu'en posant $c = f(a)$ on ait*

$$\Big((x_1, x_2) \in \mathrm{V} \textit{ et } f(x_1, x_2) = c\Big) \Leftrightarrow \Big(\exists t \in \mathrm{I}, \quad \varphi(t) = (x_1, x_2)\Big).$$

On suppose de plus que φ est dérivable et que $\varphi'(t_0) \neq 0$. Dans ces conditions, $\varphi'(t_0)$ indique la direction de la tangente à la ligne de niveau et le gradient de f au point a est orthogonal à cette tangente.

Exemples.

1. Considérons la fonction f définie sur $\mathbb{R}^2$ par $f(x, y) = x^2 + y^2$. Pour $c = r^2 > 0$, la ligne de niveau est le cercle centré en 0 et de rayon r, et le gradient $\nabla f_x = 2x$ est bien orthogonal à la tangente au cercle de rayon r.

2. Soient P et Q deux points fixés dans le plan, par exemple $\mathrm{P} = (-1, 0)$ et $\mathrm{Q} = (1, 0)$; la fonction

$$f(\mathrm{M}) = d(\mathrm{M}, \mathrm{P}) + d(\mathrm{M}, \mathrm{Q})$$

admet pour lignes de niveau les ellipses de foyers P et Q. Dans cet exemple, il se passe un phénomène spécial. Si on considère la fonction

$$g(\mathrm{M}) = d(\mathrm{M}, \mathrm{P}) - d(\mathrm{M}, \mathrm{Q}),$$

les lignes de niveau de g sont les hyperboles de foyers P et Q, et on vérifie que les hyperboles de cette famille sont orthogonales à la famille d'ellipses. En effet pour tout point M différent de P et de Q, ∇f_{M} et ∇g_{M} sont non nuls et orthogonaux.

Plan tangent à une surface $z = f(x, y)$

On se donne un ouvert U du plan, une fonction réelle f, différentiable en tout point de U et on considère la surface S dans $\mathbb{R}^3$ "au dessus" de U définie par

$$\mathrm{S} = \{(x, y, z) \in \mathbb{R}^3 : (x, y) \in \mathrm{U},\ z = f(x, y)\}.$$

On dira que T est un *vecteur tangent* à S au point M $\in$ S s'il existe une courbe φ tracée sur la surface S telle que $\varphi(0) = \mathrm{M}$ et $\varphi'(0) = \mathrm{T}$. Pour tracer une courbe sur S, il revient au même de tracer une courbe ψ dans la base U et de poser ensuite $\varphi(t) = (\psi_1(t), \psi_2(t), f(\psi_1(t), \psi_2(t))$. Si T est un tel vecteur tangent, on vérifie que

$$\mathrm{T} = (\psi_1'(0), \psi_2'(0), \frac{\partial f}{\partial x}(\mathrm{M})\psi_1'(0) + \frac{\partial f}{\partial y}(\mathrm{M})\psi_2'(0)).$$

On voit donc que tout vecteur tangent $\mathrm{T} = (u, v, w)$ à la surface S au point M vérifie l'équation

$$w = \frac{\partial f}{\partial x}(\mathrm{M})\, u + \frac{\partial f}{\partial y}(\mathrm{M})\, v.$$

On obtient ainsi l'équation d'un plan vectoriel Π_0. Inversement on peut montrer que tout vecteur $\mathrm{T} \in \Pi_0$ est un vecteur tangent à S au point M. En géométrie, on appelle plutôt *plan tangent à la surface au point* M le plan affine Π parallèle au plan Π_0 et qui passe par le point M. Le plan Π admet le vecteur $\mathrm{N} = (-\frac{\partial f}{\partial x}(\mathrm{M}), -\frac{\partial f}{\partial y}(\mathrm{M}), 1)$ comme vecteur normal. On dit que N est le *vecteur normal à la surface* S au point M. Nous avons ici choisi le vecteur qui pointe vers le haut.

Stabilité des applications différentiables

On peut énoncer un petit nombre de propriétés qui permettent de vérifier la différentiabilité dans un grand nombre de situations concrètes. Tous les espaces considérés sont des espaces vectoriels réels de dimension finie.

– Si f est une application de $\mathrm{E} = \mathrm{E}_1 \times \cdots \times \mathrm{E}_k$ dans F, linéaire par rapport à chaque variable $x_j \in \mathrm{E}_j$, $j = 1, \ldots, k$, elle est différentiable en tout point de E.

– Si f_1 est différentiable au point a_1 de E_1 dans F_1 et si f_2 est différentiable au point a_2 de E_2 dans F_2, l'application $(x_1, x_2) \to (f_1(x_1), f_2(x_2))$ est différentiable au point $a = (a_1, a_2)$ de $\mathrm{E}_1 \times \mathrm{E}_2$ dans $\mathrm{F}_1 \times \mathrm{F}_2$.

– Si f est différentiable au point a et g différentiable au point $b = f(a)$, la composition $g \circ f$ est différentiable au point a.

Supposons par exemple que f_1 et f_2 soient deux fonctions réelles définies sur $\mathbb{R}^p$ et différentiables au point $a \in \mathbb{R}^p$. La combinaison linéaire $c_1 f_1 + c_2 f_2$ peut être obtenue comme composition de l'application linéaire $x \to (x, x)$, suivie de l'application $(x_1, x_2) \to (f_1(x_1), f_2(x_2)) \in \mathbb{R}^2$, et enfin de $(x_1, x_2) \to c_1 x_1 + c_2 x_2$, qui est linéaire de $\mathbb{R}^2$ dans $\mathbb{R}$. En remplaçant la troisième application par $(x, y) \to xy$ (bilinéaire), on montre que la fonction produit $f_1 f_2$ est différentiable au point a. Par itération, les puissances f^k sont différentiables. Si $f(a) \neq 0$, la fonction $x \to 1/f(x)$ est différentiable comme composition avec $t \to 1/t$ qui est dérivable au point $f(a)$. Les fonctions coordonnées sont différentiables parce que linéaires, les polynômes (de plusieurs variables) sont différentiables par application de tout ce qui précède, les quotients de polynômes, à condition que le dénominateur ne soit pas nul, etc. . .

Formule des accroissements finis

Supposons E et F munis de normes. Si la fonction f est différentiable au point a et M-lipschitzienne de E dans F au voisinage de a, on a $\|(df)_a\|_{\mathcal{L}(\mathrm{E},\mathrm{F})} \le \mathrm{M}$. En effet, pour tout vecteur $u \in \mathrm{E}$

$$\|(df)_a(u)\|_{\mathrm{F}} = \lim_{t \to 0} \left\| \frac{f(a+tu) - f(a)}{t} \right\|_{\mathrm{F}} \le \mathrm{M}\, \|u\|_{\mathrm{E}}.$$

On a une forme d'inverse (Théorème des accroissements finis, cas vectoriel).

Théorème 3.4.2. *Soient* E, F *deux espaces vectoriels normés de dimension finie,* U *un ouvert de* E *et* a, b *deux points de* U *tels que le segment* $[a, b]$ *soit contenu dans* U *; soit* $f : \mathrm{U} \to \mathrm{F}$ *une fonction continue sur* $[a, b]$ *et différentiable en tout point du segment ouvert* $]a, b[$. *On a*

$$\|f(b) - f(a)\|_{\mathrm{F}} \le \|b-a\|_{\mathrm{E}} \sup\{\|(df)_x\|_{\mathcal{L}(\mathrm{E},\mathrm{F})} : x \in]a, b[\}.$$

Démonstration. Considérons d'abord le cas plus simple où l'espace F est euclidien. Posons $\mathrm{M} = \sup\{\|(df)_x\| : x \in]a, b[\}$, soit $v = f(b) - f(a)$ et étudions la fonction réelle g

$$t \in [0, 1] \to (f(a + t(b-a)) - f(a)) \,.\, v$$

qui vaut 0 au point $t = 0$ et $\|f(b) - f(a)\|^2$ au point $t = 1$. Sa dérivée en un point $t \in]0, 1[$ est égale à

$$g'(t) = \big((df)_{a+t(b-a)}(b-a)\big) \,.\, v$$

que l'on majore par $\|(df)_{a+t(b-a)}\|\, \|b-a\|\, \|v\| \le \mathrm{M}\, \|b-a\|\, \|v\|$, ce qui donne

$$\|f(b) - f(a)\|^2 = |g(1) - g(0)| \le \sup |g'(t)| \le \mathrm{M}\, \|b-a\|\, \|f(b) - f(a)\|,$$

d'où le résultat en divisant par $\|f(b) - f(a)\|$.

La démonstration du cas général est plus délicate.

Lemme. *Soient* g *une fonction réelle continue sur* $[a, b]$ *et* C *une constante telles que pour tout* $x \in]a, b[$, *il existe* $h > 0$ *tel que* $x + h \le b$ *et* $g(x+h) - g(x) \le \mathrm{C}\, h$ *; alors*

$$g(b) \le g(a) + \mathrm{C}\,(b-a).$$

Démonstration. On commence par fixer $s_1 \in]a, b[$. On pose

$$\mathrm{F} = \{x \in [s_1, b] : g(x) \le g(s_1) + \mathrm{C}\,(x - s_1)\}.$$

On voit que F est fermé borné non vide, donc F admet un plus grand élément t. Cet élément t est nécessairement égal à b d'après l'hypothèse : sinon on pourrait trouver $h > 0$ tel que $y = t + h \le b$ et $g(t+h) \le g(t) + \mathrm{C}\, h$, donc puisque $t \in \mathrm{F}$,

$$g(y) = g(t+h) \le g(t) + \mathrm{C}\, h \le g(s_1) + \mathrm{C}\,(t - s_1) + \mathrm{C}\, h = g(s_1) + \mathrm{C}\,(y - s_1),$$

et on aurait ainsi trouvé un élément y de F plus grand que t. On a donc montré que $g(b) \le g(s_1) + \mathrm{C}\,(b - s_1)$, et ceci pour tout $s_1 \in]a, b]$. On en déduit $g(b) \le g(a) + \mathrm{C}\,(b-a)$ en faisant tendre s_1 vers a.

On se donne un espace vectoriel réel F de dimension finie, qu'on munit d'une norme notée $x \to \|x\|_{\mathrm{F}}$.

Lemme. *Soit* φ *continue sur* $[0, 1]$, *à valeurs dans* F, *dérivable sur l'intervalle ouvert* $]0, 1[$, *et telle que* $\|\varphi'(t)\|_{\mathrm{F}} \le \mathrm{M}$ *pour tout* $x \in]0, 1[$. *On a*

$$\|\varphi(1) - \varphi(0)\|_{\mathrm{F}} \le \mathrm{M}.$$

L'interprétation cinématique du résultat est évidente dans le cas où la norme est la norme euclidienne : si un mobile se déplace dans l'espace "usuel" avec une vitesse $\le \mathrm{M}$, il ne s'éloigne pas de plus de MT de son point de départ si le mouvement se déroule pendant un intervalle de temps égal à T.

Démonstration. Prenons $\mathrm{M}_1 > \mathrm{M}$ et posons $g(t) = \|\varphi(t) - \varphi(0)\|_\mathrm{F}$. On va montrer que la fonction g vérifie l'hypothèse du lemme précédent avec $\mathrm{C} = \mathrm{M}_1$. Fixons $t \in]0,1[$. La relation $\varphi(t+h)-\varphi(t) = h\varphi'(t)+o(h)$ donne $\|\varphi(t+h)-\varphi(t)\|_\mathrm{F} \le \mathrm{M}h+\|o(h)\|_\mathrm{F} = h(\mathrm{M}+\varepsilon(h)) < h\mathrm{M}_1$ pour h assez petit, ce qui permet de choisir $h > 0$ assez petit pour que $t + h \le 1$ et $\|\varphi(t + h) - \varphi(t)\|_\mathrm{F} \le \mathrm{M}_1\, h$. On aura ensuite

$$g(t + h) - g(t) = \|\varphi(t + h) - \varphi(0)\|_\mathrm{F} - \|\varphi(t) - \varphi(0)\|_\mathrm{F} \le \|\varphi(t + h) - \varphi(t)\|_\mathrm{F} \le \mathrm{M}_1\, h$$

d'après l'inégalité triangulaire. D'après le lemme précédent on obtiendra $g(0) \le g(1) + \mathrm{M}_1$, donc $\|\varphi(1) - \varphi(0)\|_\mathrm{F} \le \mathrm{M}_1$. Pour finir, on fait tendre M_1 vers M.

Démonstration du théorème dans le cas général. On étudie la variation de la fonction vectorielle φ d'une seule variable définie par

$$\varphi(t) = f(a + t(b - a))$$

lorsque t varie de 0 à 1. Sa dérivée en un point $t \in]0, 1[$ est égale à $(df)_{a+t(b-a)}(b-a)$, que l'on majore par $\mathrm{M}\,\|b - a\|_\mathrm{E}$. La majoration des accroissements finis du lemme précédent, appliquée à la fonction φ sur l'intervalle $[0, 1]$ donne ensuite

$$\|f(b) - f(a)\|_\mathrm{F} = \|g(1) - g(0)\|_\mathrm{F} \le \mathrm{M}\,\|b - a\|_\mathrm{E},$$

ce qui est le résultat cherché.

Corollaire 3.4.1. *Soient* U *un ouvert d'un espace vectoriel réel* E *de dimension finie,* C *un sous-ensemble convexe de* U *et* $f : \mathrm{U} \to \mathrm{F}$ *telle que* $\|(df)_x\|_{\mathcal{L}(\mathrm{E},\mathrm{F})} \le \mathrm{M}$ *pour tout* $x \in \mathrm{C}$ *; alors* f *est* M*-lipschitzienne dans* C.

Démonstration. C'est une application immédiate du résultat précédent : si x, y sont deux points de U, le segment $[x, y]$ tout entier est contenu dans U puisque U est convexe.

Corollaire 3.4.2. *Soient* U *un ouvert convexe de* E *et* $f : \mathrm{U} \to \mathrm{F}$ *telle que* $(df)_x = 0$ *pour tout* $x \in \mathrm{U}$ *; alors* f *est constante dans* U.

Démonstration. C'est clair à partir du corollaire précédent.

Le résultat précédent est évidemment faux si l'ouvert U de définition est formé de deux morceaux disjoints, mais il reste valable pour les ouverts connexes.

La classe C^1 *: cas des valeurs vectorielles*

Définition 3.4.2. Soit U un ouvert d'un espace vectoriel de dimension finie E ; on dit qu'une fonction f définie sur U, à valeurs dans F (espace vectoriel de dimension finie) est *de classe* C^1 *sur* U si elle est différentiable en tout point de U, et si de plus l'application df, définie sur U, à valeurs dans $\mathcal{L}(\mathrm{E}, \mathrm{F})$, qui associe à chaque $a \in \mathrm{U}$ l'application linéaire $(df)_a \in \mathcal{L}(\mathrm{E}, \mathrm{F})$ est continue de U dans $\mathcal{L}(\mathrm{E}, \mathrm{F})$.

Proposition 3.4.4. *Soit* U *un ouvert de* $\mathbb{R}^n$ *et soit* $f : \mathrm{U} \to \mathbb{R}^m$ *; la fonction* f *est de classe* C^1 *sur* U *si et seulement si on a les propriétés suivantes : chacune des fonctions coordonnées* f_i, $i = 1, \ldots, m$, *admet en tout point de* U *des dérivées partielles* $\frac{\partial f_i}{\partial x_1}, \ldots, \frac{\partial f_i}{\partial x_n}$, *et toutes ces dérivées partielles sont continues dans* U.

Démonstration. Si la fonction coordonnée f_i a des dérivées partielles continues dans U, f_i sera différentiable dans U d'après le théorème 3.3.1, donc f sera différentiable. De plus, l'application vectorielle df, dont l'expression en coordonnées est l'application qui associe à chaque $a \in \mathrm{U}$ la matrice jacobienne de f au point a,

$$a \in \mathrm{U} \to \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \dots & \frac{\partial f_1}{\partial x_n}(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(a) & \dots & \frac{\partial f_m}{\partial x_n}(a) \end{pmatrix}$$

est continue si et seulement si les $m \times n$ fonctions coordonnées $a \to \frac{\partial f_i}{\partial x_j}(a)$ sont continues, donc la continuité des dérivées partielles implique la continuité de la différentielle. Inversement, si on sait que f est différentiable en tout point de $\mathrm{U} \subset \mathbb{R}^n$, les dérivées partielles existent en tout point de U, et elles sont continues comme coordonnées d'une application vectorielle continue.

Théorème d'inversion locale

Ce théorème important ne sera pas démontré ici. On donnera une petite idée d'une démonstration à la fin du chapitre 5.

Théorème 3.4.3. Application ouverte (inversion locale). *Soient* E *et* F *deux espaces normés de dimension finie,* U *un voisinage de $a \in \mathrm{E}$ et f une fonction de classe* C^1 *de* U *dans* F *; si $(df)_a$ est inversible, il existe un ouvert* V *contenant a et un ouvert* W *contenant $f(a)$ tels que f soit une bijection de* V *sur* W*, dont l'inverse soit de classe* C^1 *sur* W*.*

Ce théorème dont le sens intuitif est assez clair (en gros, l'application f agit comme sa différentielle) implique le *théorème des fonctions implicites*, dont on n'énoncera ici que le cas le plus simple, que l'on peut démontrer directement sans passer par l'inversion locale. Supposons donnée une fonction réelle $f(x, y)$ de deux variables, et revenons à l'idée de courbe de niveau, c'est à dire à l'ensemble C des points x tels que $f(x, y) = c$ pour une certaine valeur de c. La question est la suivante : peut on, au voisinage d'un point $a_0 = (x_0, y_0)$ tel que $f(x_0, y_0) = c$, décrire les points (x, y) de C par une équation, ou bien de la forme $y = \varphi(x)$, ou bien de la forme $x = \psi(y)$? Pour se convaincre qu'il faut parfois renoncer à l'une des deux formes, le lecteur pensera à l'exemple simple du cercle unité, représenté par $f(x, y) = x^2 + y^2$ et $\mathrm{C} = \{(x, y) : x^2 + y^2 = 1\}$. Au voisinage du point $a_0 = (1, 0)$, on ne peut pas espérer représenter tous les points (x, y) du cercle C voisins de a_0 sous la forme $y = \varphi(x)$, parce que pour chaque valeur de $x < 1$ il y a deux solutions (et aucune si $x > 1$). Dans cet exemple, $\frac{\partial f}{\partial y}(a_0) = 0$, et c'est ça le problème.

Supposons donc $\frac{\partial f}{\partial y}(a_0) \neq 0$, et considérons l'application g de $\mathbb{R}^2$ dans $\mathbb{R}^2$ définie par $g(x, y) = (x, f(x, y))$. La matrice jacobienne de g au point a_0 est égale à

$$\begin{pmatrix} 1 & 0 \\ \frac{\partial f}{\partial x}(a_0) & \frac{\partial f}{\partial y}(a_0) \end{pmatrix}$$

qui est inversible. D'après le théorème d'inversion locale, il existe une fonction réciproque de g, définie dans un voisinage de $g(a_0) = (x_0, c)$. Si h est cette fonction, le point $h(x, c)$ doit être sur C pour x voisin de x_0 (réfléchir un peu) et on a trouvé un paramétrage de la courbe de niveau au voisinage de a_0, en écrivant

$$h(x, c) = (x, \varphi(x)).$$

Le lecteur comprendra vite comment généraliser, par exemple aux surfaces de $\mathbb{R}^3$ définies par une équation implicite $f(x, y, z) = c$.

3.5. Formules de Taylor à l'ordre deux

Dans cette section nous nous limiterons à des fonctions à valeurs réelles. Le lecteur déçu par cette restriction pourra imaginer comment traiter des situations à valeurs dans $\mathbb{R}^m$ en étudiant simultanément les différentes fonctions coordonnées.

Dérivées partielles secondes

Soit V un voisinage d'un point a de $\mathbb{R}^n$ et soit f une fonction réelle définie sur V, et admettant des dérivées partielles $\mathrm{D}_1 f(x), \ldots, \mathrm{D}_n f(x)$ en tout point $x \in \mathrm{V}$. On a alors pour chaque $j = 1, \ldots, n$ une fonction réelle $\mathrm{D}_j f : x \in \mathrm{V} \to \mathrm{D}_j f(x) \in \mathbb{R}$, définie sur V, pour laquelle on peut chercher si elle admet à son tour des dérivées partielles. Si $\mathrm{D}_j f$ admet une dérivée partielle D_i au point a, on dit que c'est la dérivée partielle seconde $\mathrm{D}_i(\mathrm{D}_j f)$ de f au point a. On note aussi pour $i \neq j$

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial}{\partial x_i}\Big(\frac{\partial f}{\partial x_j}\Big)(a) = \mathrm{D}_i(\mathrm{D}_j f)(a)$$

et pour $i = j$

$$\mathrm{D}_i^2 f(a) = \mathrm{D}_i(\mathrm{D}_i f)(a) = \frac{\partial^2 f}{\partial x_i^2}(a) = \frac{\partial}{\partial x_i}\Big(\frac{\partial f}{\partial x_i}\Big)(a).$$

Exemple. Calculer les dérivées secondes de $f(x_1, x_2) = \mathrm{e}^{x_1 x_2 - x_1^2/2}$. On remarque la symétrie des dérivées croisées, $\mathrm{D}_1\mathrm{D}_2 f = \mathrm{D}_2\mathrm{D}_1 f$.

Lemme de Schwarz. (Première manière) *Soit* V *un voisinage de* $a \in \mathbb{R}^n$ *; si* $f : \mathrm{V} \to \mathbb{R}$ *admet dans* V *des dérivées partielles secondes continues, on a pour tous* $i, j = 1, \ldots, n$

$$\mathrm{D}_i\mathrm{D}_j f(a) = \mathrm{D}_j\mathrm{D}_i f(a).$$

Démonstration. Faisons la démonstration pour $\mathbb{R}^2$, et avec $a = (0, 0)$. On pose

$$\varphi(t) = f(t, t) - f(t, 0) - f(0, t) + f(0, 0), \quad \psi(s) = f(s, t) - f(s, 0).$$

On voit que $\psi'(s) = \mathrm{D}_1 f(s, t) - \mathrm{D}_1 f(s, 0)$, donc par les accroissements finis

$$\varphi(t) = \psi(t) - \psi(0) = t\psi'(c_t) = t\big(\mathrm{D}_1 f(c_t, t) - \mathrm{D}_1 f(c_t, 0)\big)$$

où c_t est un point entre 0 et t, et à nouveau avec les accroissements finis

$$\varphi(t) = t^2\, \mathrm{D}_2\mathrm{D}_1 f(c_t, d_t)$$

avec d_t entre 0 et t ; quand t tend vers 0,

$$\mathrm{D}_2\mathrm{D}_1 f(0, 0) = \lim_{t \to 0} t^{-2}\varphi(t),$$

d'où le résultat en échangeant les rôles des variables x_1 et x_2 dans ce qui précède.

Considérons une fonction $f(x_1, x_2)$ différentiable dans un voisinage V d'un point $a = (a_1, a_2) \in \mathbb{R}^2$, vérifiant les hypothèses du lemme précédent, et considérons une direction $u = (u_1, u_2)$. Etudions la fonction f en des points $a + tu$ de la droite passant par a et de vecteur directeur u, avec t réel assez petit pour que $a + tu \in \mathrm{V}$. Posons donc

$$\varphi(t) = f(a_1 + tu_1, a_2 + tu_2).$$

On sait que

$$\varphi'(t) = \frac{\partial f}{\partial x_1}(a_1 + tu_1, a_2 + tu_2)\, u_1 + \frac{\partial f}{\partial x_2}(a_1 + tu_1, a_2 + tu_2)\, u_2,$$

ce qui nous donnera en dérivant une fois de plus, en posant $b_0 = a_0 + tu$

$$\varphi''(t_0) = \frac{\partial^2 f}{\partial x_1^2}(b_0)\, u_1^2 + \frac{\partial^2 f}{\partial x_2 \partial x_1}(b_0)\, u_1 u_2 + \frac{\partial^2 f}{\partial x_1 \partial x_2}(b_0)\, u_1 u_2 + \frac{\partial^2 f}{\partial x_2^2}(b_0)\, u_2^2 =$$

$$\frac{\partial^2 f}{\partial x_1^2}(b_0)\, u_1^2 + 2\frac{\partial^2 f}{\partial x_2 \partial x_1}(b_0)\, u_1 u_2 + \frac{\partial^2 f}{\partial x_2^2}(b_0)\, u_2^2.$$

On retiendra cette formule utile,

$$\frac{d^2}{dt^2} f(a+tu)_{\big|t=t_0} = \mathrm{H}_{a+t_0 u}(u,u)$$

en désignant par H_x la forme quadratique dont la matrice dans la base canonique est la matrice des dérivées partielles secondes de f au point x.

Différentielle seconde

Reprenons les mêmes questions de façon plus théorique. Soient E un espace vectoriel de dimension finie, a un point de E, V un voisinage de a et f une fonction à valeurs réelles définie sur V ; supposons que f soit différentiable en tout point de V. On a donc une application dérivée df définie de V dans E^*. On peut chercher si cette nouvelle application df est différentiable en un point $a \in \mathrm{U}$. Dans ce cas on obtiendra une différentielle qui sera un élément noté $(d^2 f)_a \in \mathcal{L}(\mathrm{E}, \mathrm{E}^*)$.

Attention ! Ne pas confondre $(df)_x$, qui est une *valeur* de l'application df, avec l'application df elle-même, qui associe à chaque point $x \in \mathrm{V}$ la différentielle $(df)_x$ de f au point x.

Définition 3.5.1. On dit que f est *deux fois différentiable au point a* si f est définie et différentiable dans un voisinage V de a et si l'application df, définie de V dans E^*, est différentiable au point a. On note

$$(d^2 f)_a = \big(d(df)\big)_a.$$

Pour tout vecteur $h \in \mathrm{E}$, l'élément $(d^2 f)_a(h)$ est une forme linéaire sur E ; on peut donc l'appliquer à un second vecteur $k \in \mathrm{E}$, pour obtenir

$$\varphi_a(h,k) = \big((d^2 f)_a(h)\big)(k) \in \mathbb{R}.$$

Il est clair que φ_a est bilinéaire. On montrera un peu plus loin qu'elle est symétrique. La forme linéaire $(d^2 f)_a(h)$ s'exprime par

$$\lim_{t\to 0} \frac{1}{t}\big((df)_{a+th} - (df)_a\big),$$

limite calculée dans E^* ; si $x^*(t) \in \mathrm{E}^*$ tend vers x^* dans E^* quand t tend vers 0, il en résulte que $x^*(t)(x)$ tend vers $x^*(x)$ pour tout vecteur $x \in \mathrm{E}$; l'existence de la limite précédente nous permet de conclure que

$$\big((d^2 f)_a(h)\big)(k) = \lim_{t\to 0} \frac{1}{t}\big((df)_{a+th} - (df)_a\big)(k) = \lim_{t\to 0} \frac{(df)_{a+th}(k) - (df)_a(k)}{t}.$$

Cas de $\mathbb{R}^n$. Appliquons la formule précédente lorsque $\mathrm{E} = \mathbb{R}^n$ et lorsque h est le vecteur e_i de la base canonique. On obtient en prenant les coordonnées, et avec $i = 1$ et $n = 2$ pour simplifier, que le vecteur $(d^2 f)_a(e_1)$ est le vecteur de $\mathbb{R}^2$ égal à

$$\Big(\lim_{t\to 0} \frac{1}{t}\big(\mathrm{D}_1 f(a_1 + t, a_2) - \mathrm{D}_1 f(a_1, a_2)\big), \lim_{t\to 0} \frac{1}{t}\big(\mathrm{D}_2 f(a_1 + t, a_2) - \mathrm{D}_2 f(a_1, a_2)\big)\Big).$$

Il s'agit des dérivées partielles par rapport à x_1 au point a des deux fonctions $x \to \mathrm{D}_j f(x) = \frac{\partial f}{\partial x_j}(x)$, $j = 1, 2$. On a donc obtenu l'expression

$$\big((d^2 f)_a(e_i)\big)(e_j) = \frac{\partial^2 f}{\partial x_i \partial x_j}(a)$$

donc la matrice de la forme bilinéaire φ_a dans la base canonique est la matrice des dérivées partielles secondes.

Exercice. Calculer les dérivées secondes de la fonction quadratique Q définie sur $\mathbb{R}^2$ par $\mathrm{Q}(x) = x_1^2 + 4x_1x_2 + 3x_2^2$.

On a montré la symétrie des dérivées partielles secondes sous l'hypothèse que toutes ces dérivées secondes soient positives. Le résultat qui suit donne un résultat plus précis, avec la “bonne” hypothèse, et qui s'exprime de façon plus abstraite par la symétrie de la forme bilinéaire déduite de la différentielle seconde.

Théorème 3.5.1. Symétrie (Schwarz). *Si f est deux fois différentiable au point $a \in \mathrm{E}$, on a*

$$\big((d^2 f)_a(h)\big)(k) = \big((d^2 f)_a(k)\big)(h)$$

pour tous $h, k \in \mathrm{E}$.

Démonstration. On étudie pour $h, k \in \mathrm{E}$ tels que $a + h, a + k \in \mathrm{V}$ la quantité

$$\psi(h, k) = f(a + h + k) - f(a + h) - f(a + k) + f(a).$$

Cette expression est symétrique en h, k et on va montrer que

$$(d^2 f)_a(h)(k) = \lim_{t\to 0} t^{-2}\, \psi(th, tk)$$

ce qui donnera la symétrie cherchée. La démonstration, un peu délicate, commence par une extension de la formule qui donne la dérivée dans une direction u d'une application différentiable g. Ce résultat sera appliqué avec $g = df$.

Lemme. *Soit g une application différentiable au point $a \in \mathrm{E}$, et soit $t \to a_t$ une application à valeurs dans E définie pour t réel voisin de 0 ; si la différence $a_t - a$ est $\mathrm{O}(t)$ quand $t \to 0$, on a pour tout vecteur $u \in \mathrm{E}$*

$$\lim_{t\to 0} \frac{g(a_t + tu) - g(a_t)}{t} = (dg)_a(u).$$

Démonstration. Supposons que $t \to h_t \in \mathrm{E}$ soit $\mathrm{O}(t)$ lorsque t réel tend vers 0 ; puisque $g(a+h) - g(a) - (dg)_a(h)$ est $o(h)$ quand $h \to 0_{\mathrm{E}}$, on obtient par composition (proposition 3.1.1) un résultat qui est $o(t)$, c'est à dire que

$$g(a + h_t) = g(a) + (dg)_a(h_t) + o(t).$$

Posons $a_t = a + v_t$; on a $v_t = \mathrm{O}(t)$. Appliquons la remarque précédente successivement avec $h_t = v_t + tu$ et avec $h_t = v_t$,

$$g(a + v_t + tu) = g(a) + (dg)_a(v_t + tu) + o(t),$$

$$g(a+v_t) = g(a) + (dg)_a(v_t) + o(t).$$

En prenant la différence et en divisant par $t \neq 0$ on obtient

$$\frac{g(a_t+tu)-g(a_t)}{t} - (dg)_a(u) = \frac{1}{t}\, o(t) \to 0$$

lorsque t tend vers 0.

Continuons la démonstration du théorème de symétrie de Schwarz. Soient h, k deux vecteurs de E, et choisissons $t_0 > 0$ assez petit pour que tous les points de la forme $a+sh+tk$ soient dans l'ensemble de définition de f et de df lorsque $|s|, |t| < t_0$. Fixons t tel que $|t| < t_0$, et considérons la fonction

$$\varphi(s) = f(a+sh+tk) - f(a+sh) - f(a+tk) + f(a),$$

définie pour $|s| < t_0$. Sa dérivée est donnée par

$$\varphi'(s) = (df)_{a+sh+tk}(h) - (df)_{a+sh}(h).$$

D'après le théorème des accoissements finis (en une dimension) il existe un nombre $c(t)$ entre 0 et t tel que

$$\varphi(t) = \varphi(t) - \varphi(0) = t\,\varphi'(c(t)).$$

On va maintenant faire tendre t vers 0, et appliquer le lemme précédent avec $a_t = a+c(t)h$ et $g = df$ (on a bien $a_t - a = \mathrm{O}(t)$ puisque $|c(t)| \leq |t|$). On aura ainsi pour $t \neq 0$

$$\frac{\varphi(t)}{t^2} = \frac{\varphi'(c(t))}{t} = \Big(\frac{(df)_{a_t+tk} - (df)_{a_t}}{t}\Big)(h)$$

qui tend d'après le lemme vers

$$\Big(\big(d(df)\big)_a(k)\Big)(h) = \Big((d^2f)_a(k)\Big)(h).$$

On pourra donc considérer $(d^2f)_a$ comme une forme bilinéaire symétrique, et on notera alors $(d^2f)_a(h,k)$. Dans le cas de $\mathbb{R}^n$ on a vu que la matrice de cette forme bilinéaire dans la base canonique de $\mathbb{R}^n$ est la matrice des dérivées partielles secondes :

Corollaire 3.5.1. *Pour une fonction $f : \mathrm{V} \to \mathbb{R}$ deux fois différentiable en un point $a \in \mathrm{V} \subset \mathbb{R}^n$, la matrice des dérivées partielles secondes au point a est symétrique,*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a),$$

pour tous $i, j = 1, \ldots, n$.

Développement limité à l'ordre 2 (Taylor-Young)

Considérons une fonction $f(x_1, x_2)$ différentiable dans un voisinage V d'un point $a = (a_1, a_2) \in \mathbb{R}^2$, et une direction $h = (h_1, h_2)$. Etudions la fonction f en des points $a+th$ de la droite passant par a et de vecteur directeur h, avec t réel assez petit pour que $a+th \in \mathrm{V}$ en posant

$$g(t) = f(a_1+th_1, a_2+th_2).$$

On a vu que

$$\frac{d^2}{dt^2} f(a+th)_{|t=t_0} = (d^2f)_{a+t_0h}(h,h).$$

D'après la formule de Taylor-Young pour les fonctions d'une variable, appliquée à la fonction g au voisinage de $t=0$, on obtiendra donc pour h donné

$$f(a+th) = f(a) + t(df)_a(h) + \frac{t^2}{2}(d^2f)_a(h,h) + t^2\varepsilon_h(t).$$

On va obtenir un peu mieux, en montrant que le terme d'erreur peut être majoré (quand h varie) en $\|h\|^2\varepsilon(h)$.

Proposition 3.5.1. *Soient* E *un espace vectoriel de dimension finie,* $a \in \mathrm{E}$, V *un voisinage de* a, $f : \mathrm{V} \to \mathbb{R}$ *une fonction différentiable dans* V *et deux fois différentiable au point* a. *On peut alors écrire*

$$f(a+h) = f(a) + (df)_a(h) + \frac{1}{2}(d^2 f)_a(h,h) + o(\|h\|^2).$$

Démonstration. Soit $\varepsilon > 0$ donné ; on choisit $\alpha > 0$ tel que la condition $\|k\| \leq \alpha$ implique $a + k \in \mathrm{V}$ et

$$\|(df)_{a+k} - (df)_a - (d^2 f)_a(k)\|_{\mathrm{E}^*} < \varepsilon \|k\|.$$

Fixons $h \in \mathbb{R}^n$ tel que $\|h\| < \alpha$ (ce qui donne $[a, a+h] \subset \mathrm{V}$) et posons pour $t \in [0,1]$

$$\varphi(t) = f(a+th) - f(a) - t(df)_a(h) - \frac{1}{2}t^2(d^2 f)_a(h,h).$$

Pour tout $0 \leq t \leq 1$ on a

$$\varphi'(t) = \big((df)_{a+th} - (df)_a - (d^2 f)_a(th)\big)(h)\,;$$

posons $k = th$; puisque $\|k\| = \|th\| \leq \|h\| < \alpha$ on obtient

$$|\varphi'(t)| \leq (\varepsilon\,\|h\|)\|h\| = \varepsilon\|h\|^2$$

pour tout $t \in [0,1]$, ce qui entraîne $|\varphi(1)| = |\varphi(1) - \varphi(0)| \leq \varepsilon\|h\|^2$ par la majoration des accroissements finis.

Il est relativement plus facile d'obtenir une formule de Taylor-Lagrange, parce qu'elle ne demande que le calcul de fonctions d'une seule variable :

si f *est deux fois différentiable dans un ouvert* U *contenant le segment* $[a,b]$, *il existe un point* c *de* $[a,b]$ *tel que*

$$f(b) = f(a) + (df)_a(b-a) + \frac{1}{2}(d^2 f)_c(b-a, b-a).$$

Conditions du second ordre pour un minimum local

Corollaire 3.5.2. *Soient* E *un espace vectoriel de dimension finie,* $a \in \mathrm{E}$, V *un voisinage de* a *et* $f : \mathrm{V} \to \mathbb{R}$ *différentiable dans* V. *On suppose que* f *est deux fois différentiable au point* a. *Si* $(df)_a = 0$ *et si* $(d^2 f)_a$ *vérifie*

$$\forall h \neq 0_{\mathrm{E}}, \quad (d^2 f)_a(h,h) > 0,$$

(c'est à dire que la forme bilinéaire $(d^2 f)_a$ *est définie positive sur* E*) la fonction* f *admet au point* a *un minimum local strict.*

Démonstration. Si $(d^2 f)_a$ est une forme bilinéaire symétrique définie positive, il existe un nombre $\delta > 0$ tel que $(d^2 f)_a(h,h) \geq \delta\|h\|^2$ pour tout h (décomposer dans une base orthonormée de vecteurs propres). Ce qui donne avec le DL, compte tenu de $(df)_a = 0$

$$f(a+h) - f(a) \geq \|h\|^2(\delta + \varepsilon(h)),$$

quantité > 0 pour $h \neq 0$ assez petit.

Si $\mathrm{Q}(x_1,x_2) = a\,x_1^2 + 2b\,x_1x_2 + d\,x_2^2$ est une forme quadratique réelle en deux variables, on voit immédiatement avec la décomposition de Gauss qu'elle est définie positive si et seulement si

$$a > 0, \quad ad - b^2 > 0.$$

Pour les fonctions de deux variables, on appliquera ce critère très simple à la matrice 2×2 des dérivées partielles secondes :

si $(df)_a = 0$ *et si* $\frac{\partial^2 f}{\partial x_1^2}(a) > 0$ *et*

$$\frac{\partial^2 f}{\partial x_1^2}(a)\,\frac{\partial^2 f}{\partial x_2^2}(a) - \Big(\frac{\partial^2 f}{\partial x_1 \partial x_2}(a)\Big)^2 > 0,$$

la fonction f *admet un minimum local au point* a.

Remarques. Supposons toujours que $(df)_a = 0$.

1. Si $(d^2 f)_a$ est de signature (p,q), avec $p \neq 0$, $q \neq 0$, on voit que la fonction $f - f(a)$ change de signe dans tout voisinage de a. Il n'y a donc pas d'extremum local au point a. Si $(d^2 f)_a$ est non dégénérée on peut même préciser la forme des régions où on a $f > f(a)$ et $f < f(a)$.

2. Si $(d^2 f)_a$ est dégénérée et de signature $(p,0)$ ou $(0,q)$, on ne peut pas conclure directement (on peut quand même dire que f n'a pas de maximum local en a si $p > 0$, et pas de minimum local en a si $q > 0$). Considérons par exemple les deux fonctions de deux variables

$$f(x,y) = x^2 + y^4\,; \quad g(x,y) = x^2 - y^4.$$

Ces deux fonctions ont la même dérivée seconde (dégénérée) au point $(0,0)$ de $\mathbb{R}^2$. Cependant, f admet un minimum au point $(0,0)$ alors que g change de signe dans tout voisinage de $(0,0)$.

Exercice. Rechercher les extrema locaux des fonctions f et g définies sur $\mathbb{R}^2$ par les formules

$$f(x,y) = \frac{x^2}{2} - y^2 - \frac{x^3}{3}; \quad g(x,y) = x^4 + y^4 - 2(x-y)^2.$$

Traitons en détail le cas de la fonction g. On commence par rechercher les points critiques de g. La fonction g étant paire, ces points critiques seront des paires de points symétriques. Le gradient de g en un point (x,y) est égal à

$$(4x^3 - 4(x-y), 4y^3 - 4(y-x)).$$

La nullité du gradient donne les relations $x^3 = x - y = -y^3$. On en déduit d'abord que $y = -x$, ensuite $x^3 = 2x$ donne les solutions $x = 0$ et $x = \pm\sqrt{2}$. On obtient trois points critiques, $(0,0)$ et $\pm(\sqrt{2}, -\sqrt{2})$. La matrice hessienne au point (x,y) quelconque est donnée par

$$\begin{pmatrix} 12x^2 - 4 & 4 \\ 4 & 12y^2 - 4 \end{pmatrix}.$$

Pour les points $\pm(\sqrt{2}, -\sqrt{2})$, on obtient

$$\begin{pmatrix} 20 & 4 \\ 4 & 20 \end{pmatrix}$$

qui est définie positive d'après le critère expliqué plus haut. Ces deux points sont des minima locaux (stricts). La valeur de g en ces points est $8 - 16 = -8$. Pour le point $(0,0)$, la matrice hessienne est

$$\begin{pmatrix} -4 & 4 \\ 4 & -4 \end{pmatrix}$$

est négative (mais pas définie négative). On a $g(x,x) = 2x^4 > 0$ si $x \neq 0$ mais $g(x,-x) \sim -8x^2 < 0$ pour x petit, donc le point $(0,0)$ n'est pas un extremum local de g.

On aurait pu remarquer que la fonction g tend vers $+\infty$ à l'infini (les puissances quatrièmes l'emportent sur les carrés), ce qui implique que le minimum de g sur $\mathbb{R}^2$ existe. On aurait pu en déduire immédiatement en comparant $g(0,0) = 0$ à $g(\sqrt{2}, -\sqrt{2}) = -8$ que ce dernier point est en fait le minimum absolu de la fonction g sur $\mathbb{R}^2$.

On va donner un critère un peu différent, qui demande que la différentielle seconde existe au voisinage du point a (et pas seulement au point a), mais qui demande seulement que la différentielle seconde soit une forme positive au sens large dans ce voisinage (au lieu de définie positive). Ce critère donne en fait la position du graphe de f par rapport à l'hyperplan tangent.

On dit qu'une fonction réelle f définie sur un ouvert convexe U est une *fonction convexe* si on a

$$f((1-t)a + tb) \leq (1-t)f(a) + tf(b)$$

pour tous $a, b \in \mathrm{U}$ et tout $t \in [0,1]$.

Proposition 3.5.2. *Soient* U *un ouvert convexe de* E *et* $f : \mathrm{U} \to \mathbb{R}$ *une fonction réelle deux fois différentiable en tout point de* U *; si*

$$\forall x \in \mathrm{U}, \forall h \in \mathrm{E}, \quad (d^2 f)_x(h,h) \geq 0,$$

la fonction f *est pour tout* $a \in \mathrm{U}$ *plus grande en tout point* $x \in \mathrm{U}$ *que la fonction affine* ℓ_a *tangente à* f *au point* a,

$$\forall x \in \mathrm{U}, \quad \ell_a(x) = f(a) + (df)_a(x-a) \leq f(x).$$

En particulier, f *est convexe sur* U. *Si* $(df)_a = 0$, *la fonction* f *atteint son minimum sur* U *au point* a *(minimum au sens large)*

Démonstration. La première partie résulte immédiatement de la formule de Taylor-Lagrange. En effet, pour tout point $x \in \mathrm{U}$ le segment $[a,x]$ est contenu dans U, et il existe un point c du segment (donc un point de U) tel que

$$f(x) = f(a) + (df)_a(x-a) + \frac{1}{2}(d^2 f)_c(x-a, x-a) \geq \ell_a(x)$$

puisque l'on a $(d^2 f)_c(x-a, x-a) \geq 0$ par hypothèse.

Pour montrer que la fonction f est convexe, considérons deux points $a, b \in \mathrm{U}$ et posons $c = (1-t)a + tb$, avec $t \in [0,1]$. On aura

$$f(c) = \ell_c(c) = \ell_c\big((1-t)a + tb\big) = (1-t)\ell_c(a) + t\ell_c(b) \leq (1-t)f(a) + tf(b).$$

La dernière affirmation est évidente à partir de ce qui précède.

Exercices.

1. Montrer que la fonction f définie sur $\mathbb{R}^3$ par

$$f(x,y,z) = \mathrm{e}^{ax+by+cz}$$

est convexe pour tout choix de a, b, c réels.

2. Montrer que pour la fonction g définie sur $\mathbb{R}^2$ par

$$g(x, y) = \mathrm{e}^{xy+x^2/2}$$

on a $(d^2g)_a$ positif en tout point $a \in \mathbb{R}^2$ (cette fonction g est donc également convexe).

Corollaire 3.5.3. *Soient* U *un ouvert de* E *et* $f : \mathrm{U} \to \mathbb{R}$ *deux fois différentiable en tout point de* U *; si* $(df)_a = 0$ *et s'il existe un voisinage* V *de* a *tel que*

$$\forall x \in \mathrm{V}, \forall h, \quad (d^2 f)_x(h, h) \geq 0,$$

la fonction f *admet un minimum local au point* a*.*

Démonstration. On peut trouver $r > 0$ tel que $\mathrm{B}(a, r) \subset \mathrm{V} \cap \mathrm{U}$. L'ouvert $\mathrm{W} = \mathrm{B}(a, r)$ est convexe, donc la proposition précédente donne le résultat.

Définition 3.5.2. On dit que f est de classe C^2 sur $\mathrm{U} \subset \mathrm{E}$ si $(d^2 f)_x$ est défini pour tout $x \in \mathrm{U}$ et si $d^2 f : x \to (d^2 f)_x \in \mathcal{L}(\mathrm{E}, \mathcal{L}(\mathrm{E}, \mathrm{F}))$ est continue sur U.

Si U est un ouvert de $\mathbb{R}^n$ et si $\mathrm{F} = \mathbb{R}$, l'application f est de classe C^2 si et seulement si toutes les dérivées partielles secondes de f sont continues dans U.

4. Intégrale de Riemann

Supposons que l'on veuille calculer l'intégrale

$$\int_0^1 \mathrm{e}^{-t^2}\ dt.$$

Très facile, direz-vous : il suffit de prendre une primitive F de la fonction $f : x \to \mathrm{e}^{-x^2}$ et de calculer $\mathrm{F}(1) - \mathrm{F}(0)$. Seulement voilà, il y a un problème : on ne sait pas exprimer une primitive de cette fonction f à partir des fonctions déjà connues.

Mais pourtant, direz-vous encore, on a appris que toute fonction continue admet des primitives ; il y a quelque chose qui cloche là-dedans.

Justement, c'est bien vrai qu'il existe une primitive, mais en gros la seule façon de la calculer est de poser

$$\mathrm{F}(x) = \int_0^x \mathrm{e}^{-t^2}\ dt$$

et on tourne en rond !

La moralité est la suivante : ce qu'on apprend en terminale sur le calcul d'intégrales à l'aide de primitives est très utile dans beaucoup de situations, chaque fois que l'on sait exprimer une primitive de la fonction à intégrer à partir de fonctions déjà connues ; mais pour l'essentiel, il s'agit d'une escroquerie : la notion fondamentale à définir est la notion d'intégrale (de Riemann), l'existence des primitives en sera une conséquence.

4.1. Intégrale de Riemann sur un intervalle fermé borné $[a, b]$.

Le premier objectif est d'introduire une classe de fonctions très simples, les *fonctions en escalier*, et de définir leur intégrale. La définition est plus commode si on met en place un minimum de terminologie au préalable : soient $a \le b$ deux nombres réels ; on appelle *subdivision* π de l'intervalle $[a, b]$ la donnée d'un entier $n \ge 1$ et d'une suite de points $x_0 = a < x_1 < \ldots < x_n = b$. On dira que les intervalles $]x_0, x_1[$, $]x_1, x_2[$,. . ., $]x_{n-1}, x_n[$ sont les intervalles (ouverts) de la subdivision, et on dira que $x_0, x_1, \ldots, x_n$ sont les *points de subdivision*.

On dit que la subdivision $\sigma = (z_0 = a < z_1 < \ldots < z_p = b)$ est *plus fine* que la subdivision $\pi = (x_0 = a < x_1 < \ldots < x_n = b)$ si chacun des points x_i, $i = 0, \ldots, n$ fait partie de l'ensemble $\{z_0, \ldots, z_p\}$ des points de subdivision de σ, c'est à dire que chaque intervalle $]x_{i-1}, x_i[$ de la subdivision π est "découpé" en intervalles plus petits $]z_{j-1}, z_j[$ dans la subdivision σ.

Etant données deux subdivisions π et ρ quelconques, on peut toujours trouver une subdivision σ plus fine que π et que ρ en même temps : il suffit de prendre un ensemble $z_0 < \ldots < z_p$ de points de subdivision qui contienne tous les points de subdivision de π **et** de ρ.

Définition 4.1.1. On dit qu'une fonction réelle φ définie sur $[a, b]$ est une *fonction en escalier* s'il existe une subdivision $\pi = (x_0 = a < x_1 < \ldots < x_n = b)$ de l'intervalle telle que φ soit constante sur chacun des intervalles ouverts $]x_{i-1}, x_i[$ de la subdivision, pour tout $i = 1, \ldots, n$.

Si cette condition est réalisée, on dira que π est une subdivision *adaptée* à la fonction en escalier φ. Bien entendu, toute subdivision σ plus fine que π sera encore adaptée à φ, puisque chaque intervalle de la subdivision plus fine est contenu dans un intervalle de π, sur lequel la fonction φ est constante. Si ψ est une deuxième fonction en escalier sur $[a,b]$ et ρ une subdivision adaptée à ψ, on peut trouver une subdivision σ plus fine que π et que ρ ; cette subdivision σ sera adaptée à la fois à φ et à ψ. Les deux fonctions φ et ψ seront donc constantes sur tous les intervalles ouverts de la subdivision σ, ce qui implique que toute fonction combinaison linéaire $\lambda\varphi+\mu\psi$ de ces deux fonctions en escalier sera elle aussi constante sur les intervalles de σ :

si φ et ψ sont en escalier sur $[a,b]$, toutes les combinaisons linéaires $\lambda\varphi+\mu\psi$ sont en escalier ; l'ensemble des fonctions réelles en escalier sur $[a,b]$ est un espace vectoriel réel.

Exercice. Pour tous c,d tels que $a \le c \le d \le b$ désignons par $x \to \chi_{[c,d]}(x)$ la fonction égale à 1 si $c \le x \le d$ et à 0 sinon. Montrer que l'espace des fonctions en escalier sur $[a,b]$ est l'espace vectoriel engendré par les fonctions $\chi_{(c,d]}$, lorsque c et d varient.

Plus précisément, si $\pi=(x_0=a<\ldots<x_n=b)$ est adaptée à la fonction en escalier φ, alors φ est combinaison linéaire des fonctions $\chi_{[x_i,x_j]}$, $0 \le i \le j \le n$.

Pour la même raison le maximum $\max(\varphi,\psi)$ — ou le minimum $\min(\varphi,\psi)$, ou plus généralement n'importe quelle fonction $x \to \mathrm{F}(\varphi(x),\psi(x))$ — de deux fonctions en escalier sur $[a,b]$ est une fonction en escalier sur $[a,b]$.

Soient φ une fonction en escalier sur $[a,b]$ et $\pi=(x_0=a<x_1<\ldots<x_n=b)$ une subdivision adaptée à φ ; choisissons un point quelconque $\xi_i \in]x_{i-1},x_i[$, pour tout $i=1,\ldots,n$. La fonction φ sera alors constamment égale à $c_i=\varphi(\xi_i)$ sur l'intervalle $]x_{i-1},x_i[$. Il est naturel d'essayer de définir l'intégrale de φ par l'expression

$$\mathrm{E}(\varphi,\pi)=\sum_{i=1}^{n}(x_i-x_{i-1})\,\varphi(\xi_i)=\sum_{i=1}^{n}(x_i-x_{i-1})\,c_i.$$

Pour être tout à fait rigoureux, il faut montrer que l'expression précédente ne dépend que de φ, et pas de la subdivision π adaptée à φ. C'est assez facile à comprendre mais terriblement pénible à écrire (et à lire !). Pour le faire on remarque que $\mathrm{E}(\varphi,\pi)$ ne change pas si on remplace π par une subdivision plus fine $\sigma=(z_0=a<z_1<\ldots<z_p=b)$. On peut passer de π à σ par étapes, en ajoutant à chaque étape un seul nouveau point de subdivision. Pour montrer que $\mathrm{E}(f,\pi)$ ne varie pas dans cette chaîne d'opérations, il suffit de le vérifier pour une étape. Supposons donc que l'on passe de π à σ en une seule étape, en ajoutant un unique nouveau point z de subdivision, et que $x_{i_0-1}<z<x_{i_0}$. La fonction φ est constante égale à c_{i_0} sur l'intervalle $]x_{i_0-1},x_{i_0}[$, donc aussi sur chacun des deux nouveaux intervalles $]x_{i_0-1},z[$ et $]z,x_{i_0}[$. On aura donc

$$\mathrm{E}(\varphi,\sigma)=\sum_{j=1}^{i_0-1}(x_j-x_{j-1})\,c_j+(z-x_{i_0-1})\,c_{i_0}+(x_{i_0}-z)\,c_{i_0}+\sum_{j=i_0+1}^{n}(x_j-x_{j-1})\,c_j=$$

$$=\sum_{j=1}^{i_0-1}(x_j-x_{j-1})\,c_j+(x_{i_0}-x_{i_0-1})\,c_{i_0}+\sum_{j=i_0+1}^{n}(x_j-x_{j-1})\,c_j=\mathrm{E}(\varphi,\pi).$$

Si maintenant ρ est une autre subdivision adaptée à φ, on peut considérer une subdivision σ plus fine que π et ρ, et d'après ce qui précède,

$$\mathrm{E}(\varphi,\pi)=\mathrm{E}(\varphi,\sigma)=\mathrm{E}(\varphi,\rho).$$

Définition 4.1.2. Soit φ une fonction réelle en escalier définie sur l'intervalle $[a,b]$ et soit $\pi = (x_0 = a < x_1 < \ldots < x_n = b)$ une subdivision adaptée à φ, c'est à dire que φ est constante égale à c_i sur chacun des intervalles ouverts $]x_{i-1}, x_i[$ de la subdivision, pour $i = 1, \ldots, n$. On définit *l'intégrale de φ sur $[a,b]$* par l'expression

$$\int_a^b \varphi(t)\,dt = \sum_{i=1}^n (x_i - x_{i-1})\,c_i.$$

Parfois, on notera simplement $\int_a^b \varphi$ pour abréger.

Proposition 4.1.1. Linéarité et majoration de l'intégrale des fonctions en escalier. *Si φ et ψ sont deux fonctions réelles en escalier sur $[a,b]$, on a*

$$\forall \lambda, \mu \in \mathbb{R}, \quad \int_a^b (\lambda\varphi + \mu\psi)(t)\,dt = \lambda \int_a^b \varphi(t)\,dt + \mu \int_a^b \psi(t)\,dt\,;$$

$$\Bigl(\varphi \le \psi \textit{ sur } [a,b]\Bigr) \Rightarrow \Bigl(\int_a^b \varphi(t)\,dt \le \int_a^b \psi(t)\,dt\Bigr)\,; \quad \Bigl|\int_a^b \varphi(t)\,dt\Bigr| \le \int_a^b \bigl|\varphi(t)\bigr|\,dt.$$

Démonstration. Il suffit de choisir une subdivision σ qui soit adaptée à la fois à la fonction φ et à ψ. Les deux premières propriétés sont alors très faciles à vérifier sur la formule $\mathrm{E}(\varphi, \sigma)$ qui définit l'intégrale. La troisième s'obtient en encadrant f entre $-|f|$ et $|f|$, et en appliquant les deux premières propriétés, ou bien directement par l'inégalité triangulaire appliquée à l'expression $\mathrm{E}(\varphi, \pi)$.

Remarque. En dehors de la deuxième propriété qui concerne l'ordre, tout marcherait aussi bien pour des fonctions en escalier à valeurs vectorielles (voir la section suivante), que ce soit pour la définition de l'intégrale ou pour sa linéarité. On pourrait aussi adapter la troisième propriété de la proposition précédente pour montrer que la norme de l'intégrale vectorielle est majorée par l'intégrale de la norme.

Intégrabilité par encadrement

Soit f une fonction réelle bornée sur $[a,b]$, disons bornée par le nombre C, c'est à dire que $|f| \le \mathrm{C}$ sur $[a,b]$; on peut alors trouver des fonctions en escalier $\varphi_1 \le f$, par exemple la fonction constante $\varphi_1 = -\mathrm{C}$, et également des fonctions en escalier $\varphi_2 \ge f$, par exemple $\varphi_2 = \mathrm{C}$. Fixons une fonction en escalier $\varphi_2 \ge f$. On a alors pour toute fonction en escalier $\varphi_1 \le f$ l'inégalité $\varphi_1 \le \varphi_2$, donc $\int_a^b \varphi_1(t)\,dt \le \int_a^b \varphi_2(t)\,dt$. L'ensemble des intégrales $\int_a^b \varphi_1(t)\,dt$ est donc (non vide et) majoré par $\int_a^b \varphi_2(t)\,dt$; sa borne supérieure M_1 vérifie donc l'inégalité $\mathrm{M}_1 \le \int_a^b \varphi_2(t)\,dt$. Si maintenant on fait varier φ_2, on voit que la borne inférieure m_2 de l'ensemble des valeurs des intégrales $\int_a^b \varphi_2(t)\,dt$ (où on suppose toujours $\varphi_2 \ge f$) vérifie $m_2 \ge \mathrm{M}_1$. La fonction f est sympathique lorsque ces deux nombres sont égaux : c'est dans ce cas qu'on dit que f est intégrable *au sens de Riemann*.

Définition 4.1.3. Soit f une fonction bornée à valeurs réelles définie sur $[a,b]$; on dit que la fonction f est *Riemann-intégrable* sur $[a,b]$ si pour tout nombre réel $\varepsilon > 0$, il existe deux fonctions en escalier φ_1 et φ_2 sur $[a,b]$ telles que $\varphi_1 \le f \le \varphi_2$ sur $[a,b]$ et

$$\int_a^b \varphi_2(t)\,dt - \int_a^b \varphi_1(t)\,dt = \int_a^b (\varphi_2 - \varphi_1)(t)\,dt < \varepsilon.$$

Cette définition revient exactement à dire que les deux nombres M_1 et m_2 introduits ci-dessus sont égaux. On définit alors l'intégrale de f comme étant la valeur commune,

$$\int_a^b f(t)\,dt = \sup\{\int_a^b \varphi_1(t)\,dt : \varphi_1 \le f\} = \inf\{\int_a^b \varphi_2(t)\,dt : \varphi_2 \ge f\}$$

où les fonctions φ_1 et φ_2 sont bien entendu choisies en escalier. Parfois, on notera simplement $\int_a^b f$ pour abréger.

Remarque. Si f est une fonction Riemann-intégrable sur l'intervalle $[a, b]$, elle est bornée sur $[a, b]$ **par définition**.

Remarque idéologique. L'intégrale est une limite de sommes de "petits bouts". Dans un problème physique, l'intégrale a la même "dimension" (au sens des physiciens) que les petits bouts $f(t)\,dt$. Un certain nombre d'étudiants se sont laissé enfoncer dans le crâne en terminale l'idée qu'une intégrale représente une surface. Si cette interprétation est utile dans certains cas, elle est bien nuisible à la compréhension de la plupart des intégrales physiques (telles que travail, flux, etc. . .).

Exercice. Soit $A = [0, 1] \cap \mathbb{Q}$ l'ensemble des nombres rationnels de $[0, 1]$; montrer que la fonction indicatrice de A n'est pas R-intégrable (exercice très classique).

Le fait que cette fonction χ_A ne soit pas R-intégrable est un défaut de l'intégrale de Riemann. On a de bonnes raisons de penser que cette fonction devrait avoir une intégrale nulle, dans une bonne théorie de l'intégration. C'est ce qui se passe avec la théorie plus moderne de *l'intégrale de Lebesgue*.

Intégrabilité par approximation

Pour démontrer certaines propriétés, et pour traiter ensuite le cas vectoriel (section suivante), il est commode de transformer un peu la caractérisation de l'intégrabilité par encadrement pour obtenir une caractérisation *par approximation en escalier*. On obtient une définition équivalente de l'intégrabilité d'une fonction réelle f définie sur $[a, b]$ en disant : pour tout $\varepsilon > 0$, il existe deux fonctions en escalier φ et ψ sur $[a, b]$ telles que

$$|f - \varphi| \le \psi \quad \text{et} \quad \int_a^b \psi < \varepsilon.$$

Pour passer de la formulation par encadrement, avec deux fonctions φ_1, φ_2 en escalier telles que $\varphi_1 \le f \le \varphi_2$ à celle de la ligne précédente, il suffit de poser par exemple $\varphi = \varphi_1$ et $\psi = \varphi_2 - \varphi_1$; dans ce cas $|f - \varphi| = f - \varphi_1 \le \varphi_2 - \varphi_1 = \psi$. Inversement, étant données φ et ψ qui donnent l'approximation, il suffit de poser $\varphi_1 = \varphi - \psi$, $\varphi_2 = \varphi + \psi$.

Si f est R-intégrable et si on dispose d'une approximation de la forme $|f - \varphi| \le \psi$, avec φ, ψ en escalier, l'encadrement $\varphi - \psi \le f \le \varphi + \psi$, la définition de l'intégrale de f et les propriétés de l'intégrale des fonctions en escalier montrent que $\int_a^b \varphi - \int_a^b \psi \le \int_a^b f \le \int_a^b \varphi + \int_a^b \psi$ donc

$$(*) \qquad \Big|\int_a^b f - \int_a^b \varphi\Big| \le \int_a^b \psi.$$

Exercice. Supposons que la fonction f définie sur $[a, b]$ vérifie la propriété suivante : pour tout $\varepsilon > 0$, il existe deux fonctions R-intégrables g et h sur $[a, b]$ telles que $|f - g| \le h$ et $\int_a^b h < \varepsilon$. Montrer que f est R-intégrable sur $[a, b]$.

Proposition 4.1.2. Linéarité et majoration. *Si f_1 et f_2 sont* R*-intégrables sur $[a, b]$, leurs combinaisons linéaires sont* R*-intégrables sur $[a, b]$ et*

$$\forall \lambda_1, \lambda_2 \in \mathbb{R}, \quad \int_a^b (\lambda_1 f_1 + \lambda_2 f_2)(t)\, dt = \lambda_1 \int_a^b f_1(t)\, dt + \lambda_2 \int_a^b f_2(t)\, dt.$$

Si f et g sont R*-intégrables sur $[a, b]$ et si $f \le g$ sur $[a, b]$, il en résulte que*

$$\int_a^b f(t)\, dt \le \int_a^b g(t)\, dt.$$

Si f est R*-intégrable sur $[a, b]$, la fonction $t \to |f(t)|$ est* R*-intégrable sur $[a, b]$ et on a l'inégalité*

$$\Bigl| \int_a^b f(t)\, dt \Bigr| \le \int_a^b |f(t)|\, dt.$$

Démonstration. Montrons d'abord que $\lambda_1 f_1 + \lambda_2 f_2$ est R-intégrable. Supposons donné $\varepsilon > 0$. Choisissons quatre fonctions en escalier $\varphi_1, \psi_1, \varphi_2, \psi_2$ telles que $|f_j - \varphi_j| < \psi_j$ et $\int_a^b \psi_j < \varepsilon$ pour $j = 1, 2$. On sait d'après la relation $(*)$ que cela implique l'inégalité $|\int_a^b f_j - \int_a^b \varphi_j| \le \int_a^b \psi_j < \varepsilon$. Posons $f = \lambda_1 f_1 + \lambda_2 f_2$, $\varphi = \lambda_1 \varphi_1 + \lambda_2 \varphi_2$ et d'autre part $\psi = |\lambda_1| \psi_1 + |\lambda_2| \psi_2$. Supposons pour fixer les idées que $|\lambda_1| + |\lambda_2| \le 1$. On aura

$$|\int_a^b f - \int_a^b \varphi| = \bigl|(\lambda_1 f_1 + \lambda_2 f_2) - (\lambda_1 \varphi_1 + \lambda_2 \varphi_2)\bigr| \le |\lambda_1|\, \psi_1 + |\lambda_2|\, \psi_2 = \psi$$

ainsi que

$$\int_a^b \psi = \int_a^b (|\lambda_1|\, \psi_1 + |\lambda_2|\, \psi_2) < (|\lambda_1| + |\lambda_2|)\, \varepsilon \le \varepsilon,$$

et cette approximation est possible pour tout $\varepsilon > 0$, ce qui montre que $f = \lambda_1 f_1 + \lambda_2 f_2$ est R-intégrable. On sait alors que son intégrale vérifie

$$\Bigl|\int_a^b f - \int_a^b \varphi\Bigr| = \Bigl|\int_a^b f - \bigl(\lambda_1 \int_a^b \varphi_1 + \lambda_2 \int_a^b \varphi_2\bigr)\Bigr| \le \int_a^b \psi < \varepsilon$$

et par ailleurs

$$\Bigl|\bigl(\lambda_1 \int_a^b f_1 + \lambda_2 \int_a^b f_2\bigr) - \bigl(\lambda_1 \int_a^b \varphi_1 + \lambda_2 \int_a^b \varphi_2\bigr)\Bigr| < (|\lambda_1| + |\lambda_2|)\, \varepsilon \le \varepsilon,$$

ce qui montre que

$$\Bigl|\int_a^b (\lambda_1 f_1 + \lambda_2 f_2)(t)\, dt - \bigl(\lambda_1 \int_a^b f_1(t)\, dt + \lambda_2 \int_a^b f_2(t)\, dt\bigr)\Bigr| < 2\varepsilon$$

pour tout $\varepsilon > 0$, ce qui donne l'égalité voulue.

Terminons avec les deux dernières propriétés. Supposons que $f \le g$. L'ensemble des fonctions en escalier $\le f$ est contenu dans celui des fonctions en escalier $\le g$, donc le sup

$\int_a^b f$ des intégrales de la première famille est inférieur ou égal au sup $\int_a^b g$ de la deuxième famille. La vérification du dernier point est très simple par approximation : si $|f - \varphi| \le \psi$ et $\int_a^b \psi < \varepsilon$, on aura aussi

$$\big||f| - |\varphi|\big| \le \psi,$$

donc $|\varphi|$ est une fonction en escalier qui donne l'approximation souhaitée pour la fonction valeur absolue $|f|$. L'inégalité sur les intégrales résulte de la deuxième propriété, appliquée à l'encadrement $-|f| \le f \le |f|$.

Exercice. Si $\varphi(x_1, x_2)$ est une fonction lipschitzienne sur $\mathbb{R}^2$, et si f, g sont deux fonctions réelles R-intégrables sur $[a, b]$, montrer que $t \to \varphi(f(t), g(t))$ est R-intégrable sur $[a, b]$. Cas particulier : $\max(f, g)$, $\min(f, g)$.

Proposition 4.1.3. *Soient f_1 et f_2 deux fonctions Riemann-intégrables sur $[a, b]$; la fonction produit $f_1 f_2$ est Riemann-intégrable sur $[a, b]$.*

Démonstration. Soit M une borne pour les fonctions $|f_1|$ et $|f_2|$, et soit φ_1 une approximation en escalier de f_1, telle que $|f_1 - \varphi_1| \le \psi_1$, avec ψ_1 en escalier et $\int_a^b \psi_1 < \varepsilon/(2\mathrm{M})$. Puisque $|f_1| \le \mathrm{M}$, on peut supposer que $|\varphi_1| \le \mathrm{M}$ (micro-exercice). Soit de même φ_2 une approximation en escalier de f_2, telle que $|f_2 - \varphi_2| \le \psi_2$, avec ψ_2 en escalier et $\int_a^b \psi_2 < \varepsilon/(2\mathrm{M})$; on a alors

$$f_1 f_2 - \varphi_1 \varphi_2 = (f_1 - \varphi_1) f_2 + \varphi_1 (f_2 - \varphi_2),$$

donc

$$|f_1 f_2 - \varphi_1 \varphi_2| \le \mathrm{M}\psi_1 + \mathrm{M}\psi_2 = \psi,$$

et on obtient $\int_a^b \psi < \varepsilon$.

ATTENTION. Le résultat précédent ne marche que parce que les fonctions R-intégrables sont bornées par définition. Dans une théorie plus générale de l'intégration (intégrales généralisées absolument convergentes, ou bien théorie de Lebesgue), il faudra supposer que **l'une des deux** fonctions intégrables f_1 ou f_2 est de plus **bornée** pour pouvoir conclure que $f_1 f_2$ est intégrable.

Si on considère par exemple la théorie de l'intégrale généralisée absolument convergente sur $[0, 1]$, on voit que $f(x) = x^{-1/2}$ est “intégrable” sur $[0, 1]$, mais son carré $f^2(x) = f(x).f(x) = 1/x$ ne l'est plus.

Sommes de Riemann, théorème de Riemann

Etant donnée une subdivision π de $[a, b]$, on dira que

$$\delta(\pi) = \max\{x_i - x_{i-1} : 1 \le i \le n\}$$

est le *pas*, ou le *module* de la subdivision. On définira une *subdivision pointée* (π, ξ) de l'intervalle $[a, b]$ en choisissant pour chaque indice $i = 1, \ldots, n$ un point ξ_i quelconque dans l'intervalle *fermé* $[x_{i-1}, x_i]$ de la subdivision π. Etant donnée une subdivision pointée (π, ξ), on associe à chaque fonction réelle f définie sur $[a, b]$ la *somme de Riemann*

$$\Sigma_{\pi,\xi}(f) = \sum_{i=1}^{n} (x_i - x_{i-1}) f(\xi_i).$$

Il serait peut être préférable que la notation de la somme de Riemann rappelle l'intervalle $[a,b]$, mais on peut prétendre avec un peu de mauvaise foi que $[a,b]$ est implicitement donné par la subdivision π.

Les sommes de Riemann sont les intégrales de certaines fonctions en escalier : si ξ est un choix de points pour la subdivision π, la somme de Riemann $\Sigma_{\pi,\xi}(f)$ est l'intégrale de la fonction en escalier

$$\varphi = \sum_{i=1}^{n} f(\xi_i)\chi_{]x_{i-1},x_i[}$$

qui est égale à $f(\xi_i)$ sur chaque intervalle ouvert $]x_{i-1},x_i[$. On voit facilement que :

si f et g sont deux fonctions réelles quelconques définies sur $[a,b]$ et si (π,ξ) est une subdivision pointée de $[a,b]$, on a

$$\Sigma_{\pi,\xi}(\lambda f+\mu g) = \lambda\Sigma_{\pi,\xi}(f)+\mu\Sigma_{\pi,\xi}(g)\,; \qquad \Bigl(f \le g \text{ sur } [a,b]\Bigr) \Rightarrow \Bigl(\Sigma_{\pi,\xi}(f) \le \Sigma_{\pi,\xi}(g)\Bigr).$$

Remarque. Si A est un sous-ensemble de $\mathbb{R}$, on notera χ_{A} la fonction définie sur $\mathbb{R}$ par $\chi_{\mathrm{A}}(x)=1$ si $x\in \mathrm{A}$ et $\chi_{\mathrm{A}}(x)=0$ si $x\notin \mathrm{A}$. On dit que χ_{A} est la *fonction caractéristique*, ou bien *fonction indicatrice* de A.

Soit c un nombre tel que $a\le c\le b$ et désignons par $|c,b]$ l'un quelconque des deux intervalles $]c,b]$ ou $[c,b]$; soit (π,ξ) une subdivision pointée de l'intervalle $[a,b]$, avec $\pi=(x_0=a<\ldots<x_n=b)$; on a alors en posant $\mathrm{J}=|c,b]$

$$(**) \qquad \Bigl|\Sigma_{\pi,\xi}(\chi_{\mathrm{J}}) - \int_a^b \chi_{\mathrm{J}}(t)\,dt\Bigr| = \bigl|\Sigma_{\pi,\xi}(\chi_{\mathrm{J}})-(b-c)\bigr| \le \delta(\pi).$$

On voit donc que les sommes de Riemann de la fonction χ_{J} convergent vers l'intégrale $\int_a^b \chi_{\mathrm{J}} = b-c$ de la fonction χ_{J} lorsque le pas $\delta(\pi)$ de la subdivision π tend vers 0. On verra un peu plus loin que cette convergence des sommes de Riemann vers l'intégrale est vraie pour toutes les fonctions Riemann-intégrables (c'est ce que nous appelons *théorème de Riemann*).

L'inégalité $(**)$ ci-dessus est presque évidente, ce sont uniquement les problèmes de "bord" qui peuvent nous embêter un peu. Si c n'est pas point de subdivision de π, il existe exactement un intervalle $[x_{k-1},x_k]$ de la subdivision π qui contient le point c, et $x_{k-1}<c<x_k$. On est donc sûr que $\chi_{\mathrm{J}}(\xi_i)=0$ pour $i<k$ et $\chi_{\mathrm{J}}(\xi_i)=1$ si $i>k$, et seule la valeur de $\chi_{\mathrm{J}}(\xi_k)$ est incertaine, 0 ou 1 selon que ξ_k est ou n'est pas dans J. On a donc

$$b-x_k = \sum_{i=k+1}^{n}(x_i-x_{i-1}).1 \le \Sigma_{\pi,\xi}(\chi_{\mathrm{J}}) \le \sum_{i=k}^{n}(x_i-x_{i-1}).1 = b-x_{k-1},$$

et ces deux valeurs extrêmes qui encadrent $b-c$ diffèrent entre elles de $x_k-x_{k-1}\le\delta(\pi)$, donc $|\Sigma_{\pi,\xi}(\chi_{\mathrm{J}})-(b-c)|\le\delta(\pi)$. Le cas où c est point de subdivision est un peu plus embêtant, et laissé au lecteur.

Proposition 4.1.4. *Pour toute fonction φ en escalier sur $[a,b]$, les sommes de Riemann $\Sigma_{\pi,\xi}(\varphi)$ convergent vers $\int_a^b\varphi$ lorsque $\delta(\pi)$ tend vers 0.*

Démonstration. La fonction φ est combinaison linéaire de fonctions indicatrices d'intervalles, que l'on peut choisir de la forme précédente χ_{J}. Il suffit donc d'appliquer la linéarité des sommes de Riemann et la remarque précédente.

Théorème 4.1.1. *Soit f une fonction à valeurs réelles définie sur $[a,b]$; si la fonction f est Riemann-intégrable sur $[a,b]$, alors pour tout nombre réel $\varepsilon > 0$, il existe $\eta > 0$ tel que pour toute subdivision pointée (π, ξ) de $[a,b]$ telle que $\delta(\pi) < \eta$ on ait*

$$\Big|\Sigma_{\pi,\xi}(f) - \int_a^b f(t)\,dt\Big| < \varepsilon.$$

Autrement dit, les sommes de Riemann convergent vers l'intégrale de f lorsque le pas de la subdivision tend vers 0.

Inversement, si les sommes de Riemann de f convergent vers un nombre réel I *lorsque le pas de la subdivision π tend vers* 0, *la fonction f est* R*-intégrable sur $[a,b]$ et son intégrale est égale* à I.

Démonstration. Supposons que f soit R-intégrable, et déduisons que les sommes de Riemann convergent vers $\int_a^b f$. Soit $\varepsilon > 0$ donné et soit $\varphi_1 \le f \le \varphi_2$ un encadrement de f par deux fonctions en escalier, tel que $(\int_a^b f) - \varepsilon/2 < \int_a^b \varphi_1 \le \int_a^b \varphi_2 < (\int_a^b f) + \varepsilon/2$. Puisque φ_1 et φ_2 sont en escalier, on sait que leurs sommes de Riemann respectives tendent vers les intégrales, donc il existe $\eta > 0$ tel que $|\Sigma_{\pi,\xi}(\varphi_j) - \int_a^b \varphi_j| < \varepsilon/2$ pour toute subdivision pointée (π, ξ) de module $< \eta$, et pour $j = 1, 2$. Il en résulte que

$$\Big(\int_a^b f\Big) - \varepsilon < \Big(\int_a^b \varphi_1\Big) - \frac{\varepsilon}{2} < \Sigma_{\pi,\xi}(\varphi_1) \le \Sigma_{\pi,\xi}(f) \le$$

$$\Sigma_{\pi,\xi}(\varphi_2) \le \Big(\int_a^b \varphi_2\Big) + \frac{\varepsilon}{2} < \Big(\int_a^b f\Big) + \varepsilon,$$

ce qui montre que $|\Sigma_{\pi,\xi}(f) - \int_a^b f| < \varepsilon$ pour toute subdivision pointée de module $< \eta$, et termine la démonstration du premier point.

Supposons inversement que les sommes de Riemann de f convergent vers un nombre réel I lorsque le pas de la subdivision π tend vers 0. Choisissons $\varepsilon > 0$ et soit $\eta > 0$ tel que $|\Sigma_{\pi,\xi}(f) - \mathrm{I}| < \varepsilon/4$ pour toute subdivision pointée (π, ξ) de pas $< \eta$. Soit π une subdivision de diamètre $< \eta$; pour chaque intervalle $\mathrm{I}_j =]x_{j-1}, x_j[$ de π, choisissons un point $\xi_j \in \mathrm{I}_j$ tel que $f(\xi_j) \sim \inf f(\mathrm{I}_j)$, disons $f(\xi_j) < \inf f(\mathrm{I}_j) + \varepsilon/(4(b-a))$. La fonction en escalier φ_1 égale à $\inf f(\mathrm{I}_j)$ sur chaque intervalle $]x_{j-1}, x_j[$ et à $f(x_j)$ au point x_j vérifie $\varphi_1 \le f$ et

$$0 \le \Sigma_{\pi,\xi}(f) - \int_a^b \varphi_1 < \varepsilon/4$$

donc $|\mathrm{I} - \int_a^b \varphi_1| < \varepsilon/2$. De même on pourra trouver φ_2 en escalier telle que $f \le \varphi_2$ et $|\mathrm{I} - \int_a^b \varphi_2| < \varepsilon/2$, ce qui donne bien un encadrement de f par deux fonctions en escalier avec $\int_a^b (\varphi_2 - \varphi_1) < \varepsilon$. L'intégrale de f étant comprise entre celles de φ_1 et φ_2, il en résulte aussi que $|\mathrm{I} - \int_a^b f| < \varepsilon$, et ceci pour tout $\varepsilon > 0$, donc $\int_a^b f = \mathrm{I}$.

Exemple. Si f est R-intégrable sur $[0,1]$, on a

$$\int_0^1 f(t)\,dt = \lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{n} f\Big(\frac{i}{n}\Big).$$

Cette formule s'applique plus ou moins directement dans un grand nombre d'exercices, par exemple dans l'exercice classique suivant : montrer que

$$\lim_{n\to+\infty} \sum_{k=1}^{n} \frac{1}{n+k} = \ln 2.$$

Fonctions intégrables usuelles

Théorème 4.1.2. Intégrabilité des fonctions continues. *Si f est continue sur $[a,b]$, elle est* R*-intégrable sur $[a,b]$.*

Démonstration. On sait que toute fonction réelle continue f sur un intervalle fermé borné $[a,b]$ est uniformément continue, ce qui nous dit qu'étant donné $\varepsilon > 0$, il existe $\eta > 0$ tel que $|f(x) - f(y)| < \varepsilon/(b-a)$ chaque fois que $|x-y| < \eta$. Il en résulte que pour toute subdivision π de module $< \eta$, la fonction en escalier φ_π définie par $\varphi_\pi(x) = f(x_i)$ si $x \in]x_{i-1}, x_i]$, pour $i = 1, \ldots, n$ et $\varphi_\pi(a) = f(a)$, c'est à dire

$$\varphi_\pi = f(a)\chi_{[a,a]} + \sum_{i=1}^{n} f(x_i)\chi_{]x_{i-1},x_i]}$$

vérifie $|f - \varphi_\pi| < \varepsilon/(b-a)$ sur l'intervalle $[a,b]$. On obtient une approximation en escalier de f en posant $\varphi = \varphi_\pi$ et en prenant pour ψ la fonction constante $\varepsilon/(b-a)$. On a $|f - \varphi| \le \psi$ et $\int_a^b \psi = \varepsilon$.

Théorème 4.1.3. *Si f est de classe* C^1 *sur $[a,b]$, on a*

$$f(b) - f(a) = \int_a^b f'(t)\,dt.$$

Démonstration. Considérons une subdivision $\pi = (x_0, \ldots, x_n)$. Pour tout $i = 1, \ldots, n$ on peut trouver d'après le théorème des accroissements finis un point $\xi_i \in [x_{i-1}, x_i]$ tel que

$$f(x_i) - f(x_{i-1}) = (x_i - x_{i-1})f'(\xi_i)$$

donc $\Sigma_{\pi,\xi}(f') = f(b) - f(a)$. Quand on fait tendre le diamètre de π vers 0, cette somme de Riemann particulière doit tendre vers l'intégrale de f', puisque f' est R-intégrable.

Théorème 4.1.4. *Si f est une fonction monotone sur l'intervalle borné $[a,b]$, alors f est Riemann-intégrable sur $[a,b]$.*

Démonstration. Supposons par exemple f croissante. Découpons l'intervalle $[a,b]$ en n parties égales, chacune de longueur $h = (b-a)/n$, et limitées par les points $x_0 = a < x_1 = a + h < x_2 = a + 2h < \ldots < x_n = b$. Définissons deux fonctions étagées φ_1 et φ_2 sur $[a,b]$ en posant $\varphi_1(x) = f(x_{i-1})$ pour tout $x \in [x_{i-1}, x_i[$, et $\varphi_2(x) = f(x_i)$ pour tout $x \in]x_{i-1}, x_i]$, pour tout $i = 1, \ldots, n$. Pour être complet, on pose aussi $\varphi_1(b) = f(b)$ et $\varphi_2(a) = f(a)$. On voit que $\varphi_1 \le f \le \varphi_2$,

$$\int_a^b \varphi_1 = h(f(a) + f(a+h) + \cdots + f(a + (n-1)h))$$

et

$$\int_a^b \varphi_2 = h(f(a+h) + f(a+2h) + \cdots + f(b)).$$

On a donc $\int_a^b (\varphi_2 - \varphi_1) = h(f(b) - f(a)) = (b-a)\big(f(b) - f(a)\big)/n$, quantité que l'on peut rendre aussi petite que l'on veut, en choisissant l'entier n très grand.

Relation de Chasles

Soient $a < b < c$ et f une fonction à valeurs réelles définie sur $[a, c]$; alors f est Riemann-intégrable sur $[a, c]$ si et seulement si elle est intégrable sur $[a, b]$ et sur $[b, c]$, et dans ce cas

$$\int_a^c f = \int_a^b f + \int_b^c f.$$

On note que $\int_a^a f = 0$. Dans le cas où $b < a$, on convient que $\int_a^b f(t)\,dt = -\int_b^a f(t)\,dt$. Si on se donne a, b, c dans un ordre quelconque, on a alors la relation

$$\int_a^c f(t)\,dt = \int_a^b f(t)\,dt + \int_b^c f(t)\,dt.$$

Rappelons les propriétés de majoration : si $a \le b$, on a $\left|\int_a^b f(t)\,dt\right| \le \int_a^b |f(t)|\,dt$, mais dans le cas où on ne sait pas que $a \le b$, il faut écrire la majoration sous la forme

$$\left|\int_a^b f(t)\,dt\right| \le \left|\int_a^b |f(t)|\,dt\right|.$$

En particulier, si on a $|f(t)| \le \mathrm{M}$ pour tout point t entre a et b, on en déduit la majoration $\left|\int_a^b f(t)\,dt\right| \le \mathrm{M}\,|b-a|$, quelle que soit la position respective de a et b.

Intégrale et primitives

Soit f une fonction réelle continue définie sur un intervalle ouvert I ; soit a un point fixé de I et posons

$$\mathrm{F}(x) = \int_a^x f(t)\,dt$$

pour tout $x \in \mathrm{I}$ (avec la convention ci-dessus lorsque $x < a$).

Théorème 4.1.5. *La fonction* F *est une primitive de f sur* I. *On a donc montré que toute fonction continue sur un intervalle ouvert* I *admet des primitives.*

Démonstration. Montrons que pour tout $x \in \mathrm{I}$, le nombre $f(x)$ est la dérivée de F au point x. On a

$$\left|\frac{\mathrm{F}(x+h) - \mathrm{F}(x)}{h} - f(x)\right| = \left|\frac{1}{h}\int_x^{x+h} (f(t) - f(x))\,dt\right| \le \left|\frac{1}{h}\int_x^{x+h} |f(t) - f(x)|\,dt\right| \le \varepsilon$$

si $|h|$ est choisi assez petit pour que $x + h \in \mathrm{I}$ et $|f(t) - f(x)| < \varepsilon$ pour tout t entre les points x et $x + h$.

4.2. Le cas vectoriel.

Dans ce paragraphe on considère des fonctions définies sur un intervalle $[a, b]$ et à valeurs dans un espace vectoriel réel F de dimension finie. On supposera que F est muni d'une norme. On appellera *fonction en escalier de* $[a, b]$ *dans* F une fonction φ de $[a, b]$ dans F pour laquelle il existe un entier $n \ge 1$ et une subdivision $x_0 = a < x_1 < \ldots < x_n = b$ de $[a, b]$ telle que φ soit constante sur chaque intervalle $]x_{i-1}, x_i[$, pour $i = 1, \ldots, n$.

Si c_i désigne la valeur constante (vectorielle) de φ sur l'intervalle $]x_{i-1}, x_i[$, on définit l'intégrale de φ sur l'intervalle $[a, b]$ par la formule

$$\int_a^b \varphi(t)\,dt = \sum_{i=1}^n (x_i - x_{i-1})\,c_i \in \mathrm{F}.$$

(on doit vérifier que le résultat ne dépend que de φ : la vérification est la même que dans le cas réel). On vérifie facilement les propriétés de linéarité de l'intégrale des fonctions en escalier vectorielles, ainsi que la majoration $\| \int_a^b \varphi\| \le \int_a^b \|\varphi\|$ (appliquer l'inégalité triangulaire à la somme de vecteurs qui définit l'intégrale).

Définition 4.2.1. Soit f une fonction définie sur $[a,b]$ et à valeurs dans F ; on dit que f est *intégrable au sens de Riemann* (nous écrirons en abrégé : R-intégrable) si pour tout $\varepsilon > 0$ donné, on peut trouver une fonction *vectorielle* φ en escalier et ψ en escalier réelle telles que

$$\forall x \in [a,b],\ \|f(x) - \varphi(x)\| \le \psi(x) \text{ et } \int_a^b \psi(t)\,dt < \varepsilon.$$

La définition du vecteur *intégrale* demande un petit raisonnement. Considérons une suite quelconque d'approximations $\|f - \varphi_m\| \le \psi_m$, avec (φ_m) suite de fonctions en escalier vectorielles et (ψ_m) suite de fonctions en escalier réelles telle que $\lim_m \int_a^b \psi_m = 0$. Etant donnés m et p, on pourra écrire $\|\varphi_m - \varphi_p\| \le \|f - \varphi_m\| + \|f - \varphi_p\| \le \psi_m + \psi_p$, ce qui entraîne que $\int_a^b \|\varphi_m - \varphi_p\|$ tend vers 0 lorsque $m, p \to +\infty$. Il en résulte que

$$\Big\|\int_a^b \varphi_m - \int_a^b \varphi_p\Big\|$$

tend aussi vers 0, donc la suite des intégrales $(\int_a^b \varphi_m)$ est une suite de Cauchy dans l'espace F, donc une suite convergente puisque F est complet (comme tout espace vectoriel réel de dimension finie). Il est naturel de définir l'intégrale de f comme étant le vecteur limite I de la suite des intégrales des φ_m, mais il reste un petit point à compléter : la limite ne dépend pas de la suite (φ_m, ψ_m) choisie. Si (φ, ψ) est une approximation quelconque de f telle que $\|f - \varphi\| \le \psi$ et $\int_a^b \psi < \varepsilon$, on aura

$$\|\varphi_m - \varphi\| \le \|f - \varphi_m\| + \|f - \varphi\| \le \psi_m + \psi.$$

Si on fait tendre m vers l'infini, l'intégrale de φ_m tend vers I et celle de ψ_m vers 0, ce qui entraîne que

$$\Big\|\mathrm{I} - \int_a^b \varphi\Big\| \le \int_a^b \psi < \varepsilon.$$

On voit alors que le vecteur $\mathrm{I} \in \mathrm{F}$ est uniquement déterminé par cette propriété : c'est l'unique vecteur I de F tel que pour tout $\varepsilon > 0$ donné et pour toute approximation $\|f-\varphi\| \le \psi$, $\int_a^b \psi < \varepsilon$, on ait $\Big\|\mathrm{I} - \int_a^b \varphi\Big\| < \varepsilon$. On notera encore l'intégrale par le symbole

$$\int_a^b f(t)\,dt = \mathrm{I} \in \mathrm{F}$$

ou bien $\int_a^b f$ pour abréger.

Remarque hors programme. Cette définition reste la définition correcte si on cherche à étendre la définition de l'intégrabilité à des fonctions à valeurs dans un espace vectoriel normé de dimension infinie (qu'on aura besoin de supposer complet, pour faire converger les suites de Cauchy ci-dessus).

Supposons donnée une base de F et choisissons pour norme la norme euclidienne des coordonnées dans cette base. Si $t \to f(t)$ est une fonction de $[a,b]$ dans F, désignons par f_i les fonctions coordonnées de f (ce sont des fonctions à valeurs réelles). On voit facilement que f est Riemann-intégrable au sens de la définition précédente si et seulement si toutes les fonctions coordonnées f_i sont Riemann-intégrables : on peut en effet approcher f par une fonction en escalier vectorielle φ si et seulement si on peut approcher toutes les fonctions coordonnées f_i par des fonctions en escalier réelles φ_i. Dans ce cas, l'intégrale de Riemann de f est le vecteur de F dont les composantes dans la base choisie sont les intégrales des fonctions f_i.

Il en résulte plusieurs conséquences faciles, obtenues simplement en passant aux fonctions coordonnées : si f est R-intégrable sur $[a,b]$, elle est bornée. Si f est continue de $[a,b]$ dans l'espace vectoriel F, elle est R-intégrable. Si f est de classe C^1 de $[a,b]$ dans F, on a $f(b)-f(a)=\int_a^b f'(t)\,dt$.

Etant donnée une subdivision pointée (π,ξ) de l'intervalle $[a,b]$, on associe à chaque fonction f définie sur $[a,b]$ et à valeurs dans F la *somme de Riemann* (vectorielle)

$$\Sigma_{\pi,\xi}(f)=\sum_{i=1}^{n}(x_i-x_{i-1})f(\xi_i)\in \mathrm{F}.$$

En appliquant le théorème de Riemann aux fonctions coordonnées, on obtient le même énoncé dans le cas vectoriel :

Soit f une fonction vectorielle définie sur $[a,b]$; si la fonction f est Riemann-intégrable sur $[a,b]$, alors pour tout nombre réel $\varepsilon>0$, il existe $\eta>0$ tel que pour toute subdivision pointée (π,ξ) de $[a,b]$ telle que $\delta(\pi)<\eta$ on ait

$$\left\|\Sigma_{\pi,\xi}(f)-\int_a^b f(t)\,dt\right\|<\varepsilon.$$

Autrement dit, les sommes de Riemann convergent vers l'intégrale de f lorsque le pas de la subdivision tend vers 0.

Proposition 4.2.1. Linéarité et majoration. *Si f et g sont deux fonctions* R-*intégrables sur $[a,b]$, à valeurs dans* F, *la fonction $\lambda f+\mu g$ est intégrable sur $[a,b]$ et on a*

$$\int_a^b(\lambda f+\mu g)=\lambda\int_a^b f+\mu\int_a^b g.$$

De plus, la fonction réelle $t\to\|f(t)\|$ est R-*intégrable, et on a*

$$\left\|\int_a^b f(t)\,dt\right\|\le\int_a^b\|f(t)\|\,dt.$$

Démonstration. On montre que la fonction $t\to\|f(t)\|$ est R-intégrable quand f l'est comme on a montré que $|f|$ est intégrable dans le cas réel. Pour vérifier les autres propriétés, il suffit de passer à la limite dans les propriétés correspondantes des sommes de Riemann, qui sont évidentes.

Le résultat précédent s'applique en particulier à la majoration de l'intégrale des fonctions à valeurs complexes, en considérant $\mathbb{C}$ comme un espace normé sur $\mathbb{R}$ et le module d'un nombre complexe comme la norme de cet espace.

Corollaire 4.2.1. *Si f est une fonction Riemann-intégrable sur $[a,b]$, à valeurs complexes, la fonction réelle $t\to|f(t)|$ est* R-*intégrable, et on a*

$$\left|\int_a^b f(t)\,dt\right|\le\int_a^b|f(t)|\,dt.$$

Proposition 4.2.2. *Soient f,f_1,f_2 trois fonctions à valeurs dans* F *et* R-*intégrables sur $[a,b]$, et h réelle et* R-*intégrable sur $[a,b]$; les fonctions suivantes sont Riemann-intégrables sur $[a,b]$:*

$$hf,\ \ f_1\,.\,f_2\ \text{(produit scalaire)},\ \ t\to\|f(t)\|.$$

Démonstration par approximation.

Un exemple d'application de la majoration de la norme

On va retrouver une version simple du théorème des Accroissements Finis vectoriel. Supposons f de classe C^1 d'un ouvert U de E (espace vectoriel réel de dimension finie) à valeurs dans F. On suppose que U contient un segment vectoriel $[\mathrm{A}, \mathrm{B}]$. On a alors en posant $g(t) = f(\mathrm{A} + t(\mathrm{B} - \mathrm{A}))$ pour $t \in [0, 1]$

$$f(\mathrm{B}) - f(\mathrm{A}) = g(1) - g(0) = \int_0^1 g'(t)\, dt \in \mathrm{F}$$

ce qui donne la majoration

$$\|f(\mathrm{B}) - f(\mathrm{A})\| \le \int_0^1 \|g'(t)\|\, dt \le$$

$$\le \sup_{t\in[0,1]} \|g'(t)\| = \|\mathrm{B} - \mathrm{A}\| \sup\{\|df_x\| : x \in [\mathrm{A}, \mathrm{B}]\}.$$

Le résultat reste vrai en supposant simplement que f soit différentiable dans U (c'est à dire en enlevant l'hypothèse de continuité de la différentielle), mais il demande alors une démonstration différente (bien que l'idée de fond soit la même ; voir le chapitre de Calcul Différentiel).

4.3. Intégrale double

On va s'intéresser à des fonctions définies sur un rectangle contenu dans $\mathbb{R}^2$. En fait, il ne s'agira pas d'un rectangle quelconque, mais d'un rectangle de côtés parallèles aux axes de coordonnées, de la forme $\mathrm{P} = [a, b] \times [c, d]$. Pour ne pas répéter sans cesse cette longue phrase, on conviendra d'appeler *pavé borné* un tel ensemble $\mathrm{P} = [a, b] \times [c, d]$, où a, b, c, d sont réels avec $a \le b$ et $c \le d$.

Considérons donc un pavé borné fixé $\mathrm{P} = [a, b] \times [c, d]$ dans le plan $\mathbb{R}^2$. On appelle *subdivision* π du pavé P la donnée d'une subdivision $\pi_x = (x_0 = a < \ldots < x_m = b)$ de $[a, b]$ et d'une subdivision $\pi_y = (y_0 = c < \ldots < y_n = d)$ de $[c, d]$. On dira que les mn rectangles $\mathrm{R}_{i,j} =]x_{i-1}, x_i[\times]y_{j-1}, y_j[$, pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$, sont les rectangles (ouverts) de la subdivision. On aura aussi besoin de parler des *segments de bordure* qui limitent ces rectangles : les segments horizontaux sont égaux à $]x_{i-1}, x_i[\times \{y_j\}$, pour $i = 1, \ldots, m$ et $j = 0, \ldots, n$; les segments verticaux sont les $\{x_i\} \times]y_{j-1}, y_j[$, pour $i = 0, \ldots, m$ et $j = 1, \ldots, n$. On notera $|\mathrm{Q}|$ la surface d'un pavé Q, donc on posera $|\mathrm{R}_{i,j}| = (x_i - x_{i-1})(y_j - y_{j-1})$.

On dit que la subdivision σ du pavé P est *plus fine* que la subdivision π si chaque rectangle $\mathrm{R}_{i,j}$ de la subdivision π est découpé en rectangles plus petits dans la subdivision σ. Cela revient à dire que σ_x est plus fine que π_x **et** σ_y plus fine que π_y. Etant données deux subdivisions π et ρ quelconques du pavé P, on peut toujours trouver une subdivision σ plus fine que π et que ρ en même temps : il suffit de prendre sur chaque coordonnée un ensemble de points de subdivision qui contienne tous les points de subdivision de π **et** de ρ.

Fonctions en escalier et valeur de leur intégrale

Définition 4.3.1. Une fonction réelle φ définie sur le pavé borné $\mathrm{P} = [a, b] \times [c, d]$ est dite *fonction en escalier sur le pavé* P s'il existe une subdivision π du pavé P (avec $\pi_x = (x_0 = a < \ldots < x_m = b)$ et $\pi_y = (y_0 = c < \ldots < y_n = d)$) telle que φ soit constante sur chaque rectangle ouvert $]x_{i-1}, x_i[\times]y_{j-1}, y_j[$, pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$, et constante aussi sur tous les segments de bordure (segments ouverts) de ces rectangles, horizontaux et verticaux (cette dernière condition n'est pas importante pour le calcul des intégrales, elle est plutôt de nature technique).

Si cette condition est réalisée, on dira que π est une subdivision de P adaptée à la fonction en escalier φ. On vérifie que toute subdivision σ plus fine que π sera encore adaptée à φ. Si ψ est une deuxième fonction en escalier sur P et ρ une subdivision adaptée à ψ, on peut trouver une subdivision σ plus fine que π et ρ, qui sera donc adaptée à la fois à φ et à ψ. Les deux fonctions φ et ψ seront donc constantes sur les rectangles ouverts et sur les segments de bordure de la subdivision σ, ce qui implique que toute fonction combinaison linéaire $\lambda\varphi + \mu\psi$ de ces deux fonctions en escalier sera elle aussi constante sur les intervalles de σ.

Si I est un sous-intervalle de $[a, b]$ et J un sous-intervalle de $[c, d]$ (sous-intervalles ouverts, ou fermés, ou semi-ouverts, qui peuvent être réduits à un point), on vérifie facilement que la fonction indicatrice χ_{Q} du "rectangle" $\mathrm{Q} = \mathrm{I} \times \mathrm{J}$ produit des deux intervalles, est une fonction en escalier.

Inversement, étant donnée une subdivision π, les rectangles ouverts $\mathrm{R}_{i,j}$ sont de la forme $\mathrm{I}\times\mathrm{J}$ avec $\mathrm{I} =]x_{i-1}, x_i[$ et $\mathrm{J} =]y_{j-1}, y_j[$, et les segments de bordure sont aussi de cette forme, avec par exemple $\mathrm{I} =]x_{i-1}, x_i[$ et $\mathrm{J} = \{y_j\}$ pour les segments horizontaux ; les ensembles réduits aux sommets (x_i, y_j) sont les produits $\{x_i\} \times \{y_j\}$, et toute fonction en escalier φ telle que π soit adaptée à φ est combinaison linéaire des fonctions indicatrices de ces divers ensembles.

Une fonction φ est donc une fonction en escalier si et seulement si φ est une combinaison linéaire de fonctions indicatrices χ_{Q_k} de sous-rectangles $\mathrm{Q}_k \subset \mathrm{P}$, de la forme $\mathrm{Q}_k = \mathrm{I}_k \times \mathrm{J}_k$, où I_k et J_k sont des sous-intervalles de $[a, b]$ et $[c, d]$ respectivement. On note que pour toute fonction en escalier $\varphi(x_1, x_2)$ sur $[a, b] \times [c, d]$, la fonction $x_2 \to \varphi(x_1, x_2)$ est une fonction en escalier sur l'intervalle $[c, d]$, pour tout x_1 fixé dans $[a, b]$. Pour expliquer cette remarque, on considère d'abord la fonction indicatrice d'un rectangle $\mathrm{Q} = \mathrm{I} \times \mathrm{J}$, où I et J sont deux intervalles, contenus respectivement dans $[a, b]$ et dans $[c, d]$. On remarque que

$$\chi_{\mathrm{Q}}(x_1, x_2) = \chi_{\mathrm{I}}(x_1)\, \chi_{\mathrm{J}}(x_2)$$

par conséquent pour x_1 fixé, la fonction $x_2 \to \chi_{\mathrm{Q}}(x_1, x_2)$ est la fonction en escalier $\chi_{\mathrm{I}}(x_1)\, \chi_{\mathrm{J}}$, multiple scalaire de la fonction χ_{J}. Si $\varphi = \sum_k \lambda_k \chi_{\mathrm{Q}_k}$, où $\mathrm{Q}_k = \mathrm{I}_k \times \mathrm{J}_k$, on en déduit que $x_2 \to \varphi(x_1, x_2)$ est la combinaison linéaire $\sum_k \lambda_k \chi_{\mathrm{I}_k}(x_1)\, \chi_{\mathrm{J}_k}$, qui est une fonction en escalier sur $[c, d]$.

On définit l'intégrale des fonctions en escalier de façon naturelle : désignons par $c_{i,j}$ la valeur constante de φ dans le rectangle ouvert $\mathrm{R}_{i,j} =]x_{i-1}, x_i[\times]y_{j-1}, y_j[$. On pose

$$\iint_{\mathrm{P}} \varphi = \sum_{i=1}^{m} \sum_{j=1}^{n} (x_i - x_{i-1})(y_j - y_{j-1}) c_{i,j} = \sum_{i=1}^{m} \sum_{j=1}^{n} |\mathrm{R}_{i,j}|\, c_{i,j}.$$

En toute rigueur, il faudrait vérifier, comme dans le cas de l'intégrale simple, que cette quantité ne dépend que de φ et pas de la subdivision adaptée particulière.

Proposition 4.3.1. Linéarité et majoration de l'intégrale des fonctions en escalier. *Si φ et ψ sont deux fonctions réelles en escalier sur le pavé borné* P*, on a*

$$\iint_{\mathrm{P}} (\lambda\varphi + \mu\psi) = \lambda \iint_{\mathrm{P}} \varphi + \mu \iint_{\mathrm{P}} \psi\,;$$

$$\Big(\varphi \le \psi \textit{ sur } \mathrm{P}\Big) \Rightarrow \Big(\iint_{\mathrm{P}} \varphi \le \iint_{\mathrm{P}} \psi\Big)\,; \quad \Big|\iint_{\mathrm{P}} \varphi\Big| \le \iint_{\mathrm{P}} |\varphi|.$$

Démonstration. Il suffit de choisir une subdivision σ qui soit adaptée à la fois à la fonction φ et à ψ.

Intégrabilité par encadrement

Définition 4.3.2. Soit f une fonction réelle définie sur le pavé borné P ; on dit que f est *intégrable au sens de Riemann* (nous écrirons en abrégé : R-intégrable) sur P si pour tout nombre réel $\varepsilon > 0$, il existe deux fonctions en escalier φ_1 et φ_2 sur P telles que $\varphi_1 \le f \le \varphi_2$ sur P et

$$\iint_{\mathrm{P}} \varphi_2 - \iint_{\mathrm{P}} \varphi_1 = \iint_{\mathrm{P}} (\varphi_2 - \varphi_1) < \varepsilon.$$

Comme dans le cas de l'intégrale simple, on pose ensuite

$$\iint_{\mathrm{P}} f(x,y)\,dx\,dy = \sup\{\iint_{\mathrm{P}} \varphi_1 : \varphi_1 \le f\} = \inf\{\iint_{\mathrm{P}} \varphi_2 : \varphi_2 \ge f\}$$

où les fonctions φ_1 et φ_2 sont bien entendu choisies en escalier. Parfois, on notera simplement $\iint_{\mathrm{P}} f$ pour abréger.

Remarque. Si f est Riemann-intégrable sur le pavé borné P, elle est bornée sur ce pavé.

Intégrabilité par approximation

Comme dans le cas de l'intégrale simple, il est commode de transformer un peu la caractérisation de l'intégrabilité par encadrement pour obtenir une caractérisation *par approximation en escalier*. On obtient une définition équivalente de l'intégrabilité d'une fonction réelle f définie sur P en disant : pour tout $\varepsilon > 0$, il existe deux fonctions en escalier φ et ψ sur P telles que

$$|f - \varphi| \le \psi \quad \text{et} \quad \iint_{\mathrm{P}} \psi < \varepsilon.$$

On passe d'une formulation à l'autre comme dans le cas de l'intégrale simple : si on a $\varphi_1 \le f \le \varphi_2$ on pose par exemple $\varphi = \varphi_1$ et $\psi = \varphi_2 - \varphi_1$; si $|f - \varphi| \le \psi$, il suffit de poser $\varphi_1 = \varphi - \psi$, $\varphi_2 = \varphi + \psi$.

Proposition 4.3.2. Linéarité et majoration. *Si f et g sont deux fonctions* R-*intégrables sur le pavé* P, *la fonction $\lambda f + \mu g$ est* R-*intégrable sur* P *pour tous $\lambda, \mu \in \mathbb{R}$ et on a*

$$\iint_{\mathrm{P}} (\lambda f + \mu g) = \lambda \iint_{\mathrm{P}} f + \mu \iint_{\mathrm{P}} g\,;$$

$$\Big(f \le g \textit{ sur } \mathrm{P}\Big) \Rightarrow \Big(\iint_{\mathrm{P}} f \le \iint_{\mathrm{P}} g\Big).$$

De plus, la fonction $(x,y) \to |f(x,y)|$ est R-*intégrable sur* P *et*

$$\Big|\iint_{\mathrm{P}} f\Big| \le \iint_{\mathrm{P}} |f|.$$

Démonstration : la même que pour l'intégrale simple.

Exemples et exercices.

1. Si f est bornée, et si elle est nulle en dehors d'une droite Δ horizontale ou verticale, alors elle est R-intégrable sur tout pavé borné et son intégrale est nulle.
2. Soit $\mathrm{A} = [0,1] \cap \mathbb{Q}$; montrer que la fonction indicatrice de $\mathrm{A} \times \mathrm{A}$ n'est pas R-intégrable sur le carré $[0,1] \times [0,1]$.
3. Soit g une fonction R-intégrable sur $[a,b]$ et posons $f(x,y) = g(x)$ pour tout $(x,y) \in [a,b] \times [c,d] = \mathrm{P}$. Montrer que f est R-intégrable sur le pavé P.

Proposition 4.3.3. *Soient f_1, f_2 deux fonctions réelles Riemann-intégrables sur le pavé borné* P *; les fonctions $f_1 f_2$, $\max(f_1, f_2)$, $|f_1|$ sont Riemann-intégrables sur* P.

Démonstration par encadrement ou approximation en escalier.

Théorème 4.3.1. *Si f est continue sur $[a,b] \times [c,d]$, elle est* R-*intégrable sur ce rectangle.*

Démonstration par continuité uniforme et approximation en escalier : si f est continue sur le pavé fermé borné P, elle est uniformément continue sur P ; pour tout $\varepsilon > 0$ donné, il existe donc $\eta > 0$ tel que $|f(x,y) - f(x',y')| < \varepsilon$ dès que la distance des points (x,y) et (x',y') est $< \eta$. On choisit alors une subdivision $\pi = (\mathrm{R}_{i,j})$ de P telle que tous les rectangles de P aient un diamètre $< \eta$ (c'est à dire que tout couple de points de $\mathrm{R}_{i,j}$ a une distance $< \eta$). On prend ensuite un point $\xi_{i,j}$ quelconque dans $\mathrm{R}_{i,j}$, un point quelconque ξ_α dans chaque segment de bordure S_α, et on définit une fonction en escalier φ sur P, égale à $f(\xi_{i,j})$ sur $\mathrm{R}_{i,j}$, pour chaque rectangle $\mathrm{R}_{i,j}$, à $f(\xi_\alpha)$ sur chaque segment de bordure S_α et à $f(x_i, y_j)$ pour chaque sommet (x_i, y_j) des rectangles de π. On a alors $|f - \varphi| < \varepsilon$ sur P, ce qui permet de conclure.

On définira une *subdivision pointée* (π, ξ) en choisissant un point $\xi_{i,j}$ quelconque dans chaque rectangle $\mathrm{R}_{i,j}$ de la subdivision π, pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$. Etant donnée une subdivision pointée (π, ξ), on associe à chaque fonction f définie sur le pavé P la *somme de Riemann* double

$$\Sigma^{(2)}_{\pi,\xi}(f) = \sum_{i=1}^{m} \sum_{j=1}^{n} (x_i - x_{i-1})(y_j - y_{j-1}) f(\xi_{i,j}) = \sum_{i=1}^{m} \sum_{j=1}^{n} |\mathrm{R}_{i,j}|\, f(\xi_{i,j}).$$

On a encore un théorème de Riemann dans ce cadre. On dira que

$$\delta(\pi) = \max\{\delta(\pi_x), \delta(\pi_y)\}$$

est le *pas*, ou le *module* de la subdivision π.

Théorème 4.3.2. *Soit f une fonction à valeurs réelles définie sur le pavé borné* P *; si la fonction f est Riemann-intégrable sur* P, *alors pour tout nombre réel $\varepsilon > 0$, il existe $\eta > 0$ tel que pour toute subdivision pointée (π, ξ) de* P *telle que $\delta(\pi) < \eta$ on ait*

$$\Bigl| \Sigma^{(2)}_{\pi,\xi}(f) - \iint_{\mathrm{P}} f(x,y)\, dx\, dy \Bigr| < \varepsilon.$$

Autrement dit, les sommes de Riemann convergent vers l'intégrale de f lorsque le pas de la subdivision tend vers 0.

Calcul des intégrales doubles

Lemme. *Soit φ une fonction en escalier sur le pavé borné $\mathrm{P} = [a,b] \times [c,d]$, à valeurs réelles ; pour tout x_1 fixé dans $[a,b]$, la fonction $x_2 \to \varphi(x_1,x_2)$ est en escalier sur $[c,d]$, et la fonction $x_1 \to \int_c^d \varphi(x_1,x_2)\,dx_2$ est en escalier sur $[a,b]$. De plus,*

$$\iint_{\mathrm{P}} \varphi = \int_a^b \left(\int_c^d \varphi(x_1,x_2)\,dx_2 \right) dx_1.$$

Démonstration. On a déjà noté que pour toute fonction en escalier $\varphi(x_1,x_2)$ sur le pavé $[a,b]\times[c,d]$, la fonction $x_2 \to \varphi(x_1,x_2)$ est une fonction en escalier sur l'intervalle $[c,d]$, pour tout x_1 fixé dans $[a,b]$. Par linéarité de l'intégrale, il suffit de montrer le résultat lorsque $\varphi = \chi_{\mathrm{Q}}$, où Q est un rectangle $\mathrm{I}\times\mathrm{J}$ contenu dans P, avec $\mathrm{I} = |\alpha,\beta|$ et $\mathrm{J} = |\gamma,\delta|$. On a alors $|\alpha,\beta| \subset [a,b]$, $|\gamma,\delta| \subset [c,d]$ et

$$\int_c^d \chi_{\mathrm{Q}}(x_1,x_2)\,dx_2 = \chi_{\mathrm{I}}(x_1) \int_c^d \chi_{\mathrm{J}}(x_2)\,dx_2 = (\delta-\gamma)\chi_{\mathrm{I}}(x_1),$$

et en tant que fonction de x_1, l'intégrale en x_2 est bien une fonction en escalier. Ensuite

$$\int_a^b (\delta-\gamma)\chi_{\mathrm{I}}(x_1)\,dx_1 = (\delta-\gamma)(\beta-\alpha) = |\mathrm{Q}| = \iint_{\mathrm{P}} \chi_{\mathrm{Q}}.$$

Le résultat précédent, généralisé convenablement, est l'outil fondamental pour le calcul des intégrales doubles.

Théorème 4.3.3. (Calcul des intégrales doubles). *Soit f une fonction à valeurs réelles Riemann-intégrable sur le pavé borné $\mathrm{P} = [a,b]\times[c,d]$ et telle que pour tout $x_1 \in [a,b]$, la fonction partielle $x_2 \to f(x_1,x_2)$ soit Riemann-intégrable sur $[c,d]$; alors la fonction $x_1 \to \int_c^d f(x_1,x_2)\,dx_2$ est Riemann-intégrable sur $[a,b]$ et on a*

$$\iint_{\mathrm{P}} f = \int_a^b \left(\int_c^d f(x_1,x_2)\,dx_2 \right) dx_1.$$

Démonstration. Soit $\varepsilon > 0$ donné ; il existe deux fonctions en escalier φ_1 et φ_2 sur le pavé P telles que $\varphi_1 \le f \le \varphi_2$ sur P et $\iint_{\mathrm{P}}(\varphi_2-\varphi_1) < \varepsilon$. Il résulte de l'encadrement que pour tout $x_1 \in [a,b]$, on a

$$\Phi_1(x_1) = \int_c^d \varphi_1(x_1,x_2)\,dx_2 \le \mathrm{F}(x_1) = \int_c^d f(x_1,x_2)\,dx_2 \le \Phi_2(x_1) = \int_c^d \varphi_2(x_1,x_2)\,dx_2.$$

On sait d'après le lemme précédent que $\int_a^b(\Phi_2-\Phi_1) = \iint_{\mathrm{P}}(\varphi_2-\varphi_1) < \varepsilon$ et que Φ_1, Φ_2 sont en escalier, et on vient de dire que $\Phi_1 \le \mathrm{F} \le \Phi_2$. Comme ε est arbitraire, tout ceci montre que F est intégrable sur $[a,b]$, et

$$\iint_{\mathrm{P}} \varphi_1 = \int_a^b \Phi_1 \le \int_a^b \mathrm{F} \le \int_a^b \Phi_2 = \iint_{\mathrm{P}} \varphi_2,$$

donc puisqu'on a aussi $\iint_{\mathrm{P}} \varphi_1 \le \iint_{\mathrm{P}} f \le \iint_{\mathrm{P}} \varphi_2$,

$$\Big| \iint_{\mathrm{P}} f - \int_a^b \mathrm{F} \Big| < \varepsilon.$$

Puisque ε est arbitraire, il en résulte que $\iint_{\mathrm{P}} f = \int_a^b \mathrm{F}$.

Exemple. Supposons que g, h soient deux fonctions réelles d'une seule variable, intégrables respectivement sur $[a, b]$ et $[c, d]$. Définissons une fonction f sur $[a, b] \times [c, d]$ en posant $f(x, y) = g(x)h(y)$. Alors f est intégrable sur P et

$$\iint_{\mathrm{P}} g(x)h(y)\,dx\,dy = \Big(\int_a^b g(x)\,dx\Big)\Big(\int_c^d h(y)\,dy\Big).$$

Remarque. Bien entendu, il existe un théorème analogue obtenu en échangeant les rôles de x_1 et x_2 dans l'énoncé précédent. Si la fonction f vérifie les hypothèses dans les deux cas, on en déduit le *théorème d'interversion* des ordres d'intégration,

$$\int_a^b \Big(\int_c^d f(x_1, x_2)\,dx_2\Big)\,dx_1 = \int_c^d \Big(\int_a^b f(x_1, x_2)\,dx_1\Big)\,dx_2.$$

Dans certains exemples, l'une des directions est calculable facilement (par calcul de primitive par exemple) alors que l'autre ne l'est pas.

Un cas particulier d'application du théorème précédent 4.3.3 est celui où la fonction f est continue sur $\mathrm{P} = [a, b] \times [c, d]$. Dans ce cas, on sait que la fonction f est R-intégrable sur P et toutes les fonctions $x_1 \to f(x_1, x_2)$ ou $x_2 \to f(x_1, x_2)$ sont continues, donc R-intégrables, donc la formule du théorème s'applique, dans les deux sens, à toute fonction continue sur P. On a donc toujours dans ce cas la propriété d'interversion des ordres d'intégration :

Corollaire 4.3.1. *Soit f une fonction à valeurs réelles, continue sur le pavé borné $\mathrm{P} = [a, b] \times [c, d]$; la fonction f est Riemann-intégrable sur* P *et on a*

$$\int_a^b \Big(\int_c^d f(x_1, x_2)\,dx_2\Big)\,dx_1 = \iint_{\mathrm{P}} f = \int_c^d \Big(\int_a^b f(x_1, x_2)\,dx_1\Big)\,dx_2.$$

Exercices.

1. Calculer le volume de la sphère de rayon r dans $\mathbb{R}^3$, en intégrant $\sqrt{r^2 - x^2 - y^2}$ sur un disque de rayon r dans $\mathbb{R}^2$.

2. Soit f une fonction de classe C^2 sur le rectangle $\mathrm{P} = [a, b] \times [c, d]$; calculer

$$\iint_{\mathrm{P}} \frac{\partial^2 f}{\partial x \partial y}(x, y)\,dx\,dy.$$

Inversement, étant donnée g continue, on définit une fonction G par

$$\mathrm{G}(x, y) = \iint_{[a,x]\times[b,y]} g(s, t)\,ds\,dt\,; \quad \text{calculer} \quad \frac{\partial^2 \mathrm{G}}{\partial x \partial y}.$$

3. Calculer

$$\iint_{\mathrm{P}} \mathrm{e}^{-xt} \sin x\,dx\,dt$$

sur le rectangle $\mathrm{P} = [0, \mathrm{A}] \times [0, \mathrm{B}]$ et en déduire la valeur de $\int_0^{+\infty} \frac{\sin x}{x}\,dx$ en faisant tendre A et B vers $+\infty$.

Une application. Dérivée d'intégrales dépendant d'un paramètre

Supposons donnée une fonction réelle f, définie et continue sur $[a,b]\times \mathrm{I}$, où I est un intervalle de $\mathbb{R}$. Posons pour $\lambda \in \mathrm{I}$

$$\mathrm{F}(\lambda) = \int_a^b f(x,\lambda)\,dx.$$

Il est facile de montrer que la fonction F est continue sur l'intervalle I. Il suffit de montrer que f est continue sur tout intervalle fermé $[c,d]$ contenu dans I. La fonction f est alors uniformément continue sur $[a,b]\times[c,d]$, ce qui entraîne que F est uniformément continue sur $[c,d]$; en effet, pour tout $\varepsilon > 0$ donné, il existe $\eta > 0$ tel que $|f(x,\lambda) - f(x',\lambda')| < \varepsilon$ dès que $|x-x'| + |\lambda - \lambda'| < \eta$. On a alors si $\lambda, \lambda' \in [c,d]$ et $|\lambda - \lambda'| < \eta$

$$|\mathrm{F}(\lambda) - \mathrm{F}(\lambda')| \leq \int_a^b |f(x,\lambda) - f(x,\lambda')|\,dx \leq \varepsilon|b-a|.$$

On va s'occuper d'une question plus délicate, la dérivabilité de la fonction F. On supposera maintenant que I est un intervalle ouvert.

Proposition 4.3.4. *Supposons que f admette une dérivée partielle $\dfrac{\partial f}{\partial \lambda}$ continue sur $[a,b]\times \mathrm{I}$. La fonction* F *est alors dérivable sur l'intervalle* I, *et*

$$\frac{d\mathrm{F}}{d\lambda}(\lambda_0) = \int_a^b \frac{\partial f}{\partial \lambda}(x,\lambda_0)\,dx$$

pour tout $\lambda_0 \in \mathrm{I}$.

Démonstration. Remarquons d'abord que

$$\mathrm{G}(\lambda) = \int_a^b \frac{\partial f}{\partial \lambda}(x,\lambda)\,dx$$

est une fonction continue sur I d'après le paragraphe précédent. Considérons comme avant un intervalle fermé $[c,d]$ contenu dans I. Pour tout $\lambda \in [c,d]$, on obtient en calculant l'intégrale double de $\dfrac{\partial f}{\partial \lambda}$ sur le rectangle $[a,b]\times[c,\lambda]$

$$\int_c^\lambda \Big(\int_a^b \frac{\partial f}{\partial \lambda}(x,t)\,dx\Big)\,dt = \int_a^b \Big(\int_c^\lambda \frac{\partial f}{\partial \lambda}(x,t)\,dt\Big)\,dx =$$

$$\int_a^b \big(f(x,\lambda) - f(x,c)\big)\,dx = \mathrm{F}(\lambda) - \mathrm{F}(c)$$

d'où le résultat, puisqu'on a montré que

$$\mathrm{F}(\lambda) = \mathrm{F}(c) + \int_c^\lambda \mathrm{G}(t)\,dt$$

pour tout $\lambda \in \mathrm{I}$ tel que $\lambda > c$.

Remarque. Si I n'est pas ouvert, par exemple si $\mathrm{I} = [c,d]$, on peut montrer que F admet $\mathrm{G}(c)$ pour dérivée à droite au point c.

Il existe de nombreuses variations possibles pour des énoncés concernant la dérivabilité des intégrales généralisées dépendant d'un paramètre. On se contentera d'un cas simple qui n'utilise que les résultats sur les séries de fonctions normalement convergentes. On considère des intégrales généralisées, soit sur un intervalle borné semi-ouvert $[a, b[$, soit sur $[a, +\infty[$. Pour unifier la notation, on désignera par c soit un nombre fini $> a$, soit $c = +\infty$.

Proposition 4.3.5. *On suppose donnée une fonction $g \geq 0$ sur $[a, c[$, telle que son intégrale généralisée soit (absolument) convergente,*

$$\int_a^c g(x)\, dx < +\infty.$$

On suppose donnée une fonction f définie et continue sur $[a, c[\times \mathrm{I}$, et telle que pour tout $(x, \lambda) \in [a, c[\times \mathrm{I}$ on ait

$$\Big| \frac{\partial f}{\partial \lambda}(x, \lambda) \Big| \leq g(x).$$

On suppose aussi que l'intégrale $\int_a^c f(x, \lambda)\, dx$ converge pour tout $\lambda \in \mathrm{I}$. Sous ces conditions la fonction F définie sur I par

$$\mathrm{F}(\lambda) = \int_a^c f(x, \lambda)\, dx$$

est une fonction dérivable sur l'intervalle I, et

$$\mathrm{F}'(\lambda) = \int_a^c \frac{\partial f}{\partial \lambda}(x, \lambda)\, dx.$$

Démonstration. On introduit une suite (a_n) strictement croissante, qui tend vers c, et telle que $a_0 = a$. Pour toute fonction h dont l'intégrale généralisée sur $[a, c[$ est convergente, on peut écrire

$$\int_a^c h(x)\, dx = \lim_n \int_{a_0}^{a_n} h(x)\, dx = \sum_{n=0}^{+\infty} \int_{a_n}^{b_n} h(x)\, dx,$$

où on a posé $b_n = a_{n+1}$. Posons

$$u_n(\lambda) = \int_{a_n}^{b_n} f(x, \lambda)\, dx.$$

D'après les résultats précédents, on sait que chacune des fonctions (u_n) est dérivable sur l'intervalle I, avec

$$u_n'(\lambda) = \int_{a_n}^{b_n} \frac{\partial f}{\partial \lambda}(x, \lambda)\, dx.$$

D'autre part, pour tout $\lambda \in \mathrm{I}$,

$$|u_n'(\lambda)| \leq \int_{a_n}^{b_n} \Big| \frac{\partial f}{\partial \lambda} \Big| (x, \lambda)\, dx \leq \int_{a_n}^{b_n} g(x)\, dx = v_n,$$

et la série numérique positive $\sum v_n$ converge puisque l'intégrale généralisée de g est finie. La série de fonctions $\sum u_n'()$ est donc normalement convergente sur I. Il en résulte que la fonction somme de la série est dérivable, sa dérivée étant égale à la fonction somme de la série dérivée. Or

$$\sum_{n=0}^{+\infty} u_n(\lambda) = \int_a^c f(x,\lambda)\,dx = \mathrm{F}(\lambda)$$

et

$$\mathrm{F}'(\lambda) = \sum_{n=0}^{+\infty} u_n'(\lambda) = \int_a^c \frac{\partial f}{\partial \lambda}(x,\lambda)\,dx.$$

Un exemple : la fonction Γ. On pose pour tout $s > 0$

$$\Gamma(s) = \int_0^{+\infty} x^{s-1}\,\mathrm{e}^{-x}\;dx.$$

Dans le cas $0 < s < 1$, il s'agit d'une intégrale généralisée avec deux problèmes de convergence, en 0 et en $+\infty$. On voit que $\Gamma(1) = 1$. Une intégration par parties donne la relation fondamentale $\Gamma(t+1) = t\,\Gamma(t)$, d'où il résulte que $\Gamma(2) = 1.\Gamma(1) = 1$, $\Gamma(3) = 2\Gamma(2) = 2$, et de proche en proche on obtient que $\Gamma(n) = (n-1)!$ On voit aussi que $\Gamma(t)$ tend vers $+\infty$ lorsque $t \to 0$; en effet,

$$\Gamma(t) \geq \int_0^1 x^{t-1}\,\mathrm{e}^{-x}\;dx \geq \mathrm{e}^{-1}\int_0^1 x^{t-1}\,dt = \frac{\mathrm{e}^{-1}}{t}\cdot$$

On va montrer que Γ est dérivable en tout point $s > 0$. Il suffit de trouver un intervalle $\mathrm{I} =]c,d[$ contenant s tel que la proposition précédente soit applicable à $]0,+\infty[\times[c,d]$. Etant donné $s > 0$, on choisit c, d tels que $0 < c < s < d$. On voit que $x^{s-1} \leq x^{c-1}$ pour $0 < x < 1$ et $x^{s-1} \leq x^{d-1}$ pour $x \geq 1$. Pour tout $s \in]c,d[$ on peut donc majorer la fonction

$$\frac{\partial f}{\partial s}(x,s) = x^{s-1}\,\ln x\;\mathrm{e}^{-x}$$

par la fonction d'une seule variable g définie par

$$g(x) = |\ln x|\,x^{c-1}\,\mathrm{e}^{-x}\;\text{ si } 0 < x \leq 1,\quad g(x) = |\ln x|\,x^{d-1}\,\mathrm{e}^{-x}\;\text{ si } x > 1.$$

On vérifie facilement que cette fonction g est "intégrable" sur $]0,+\infty[$, ce qui achève la vérification. On montrera avec des arguments similaires que Γ est indéfiniment dérivable. Par exemple,

$$\Gamma''(t) = \int_0^{+\infty} (\ln x)^2 x^{t-1}\,\mathrm{e}^{-x}\;dx,$$

ce qui montre que la dérivée seconde est > 0 sur $]0,+\infty[$. La fonction Γ est donc (strictement) convexe. Puisqu'elle tend vers $+\infty$ en 0 et en $+\infty$, elle admet un minimum unique, qui est situé entre 1 et 2 (puisque $\Gamma(1) = \Gamma(2) = 1$). Les tables numériques donnent ce minimum t_0 un peu au dessous de 3/2, à savoir $t_0 \simeq 1,46$..

On verra plus loin que

$$\int_0^{+\infty} \mathrm{e}^{-x^2/2}\;dx = \sqrt{\pi/2}.$$

Le changement de variable $u = x^2/2$ donne $\Gamma(1/2) = \sqrt{\pi}$, et la relation de récurrence de la fonction Γ donne $\Gamma(3/2) = \sqrt{\pi}/2$, valeur proche du minimum.

Ensembles quarrables

Avant de donner la définition, on va introduire une convention de notation commode. Supposons que f soit une fonction définie sur $\mathbb{R}^2$, mais nulle en dehors d'un pavé borné P, et R-intégrable sur ce pavé P. Pour tout pavé borné Q contenant P, on aura alors que f est R-intégrable sur Q et que $\iint_{\mathrm{Q}} f = \iint_{\mathrm{P}} f$. On posera donc simplement $\iint f$ pour représenter l'intégrale sur n'importe quel pavé assez grand pour que f soit identiquement nulle en dehors. On dira que f est R-*intégrable à support borné*.

Définition 4.3.3. On dit qu'un ensemble $\mathrm{A} \subset \mathbb{R}^2$ est *quarrable* si A est un ensemble borné et si la fonction indicatrice χ_{A} est R-intégrable (à support borné) sur $\mathbb{R}^2$.

Exemples simples d'ensembles quarrables.

Soit g une fonction positive sur $[a, b]$ et R-intégrable ; l'ensemble

$$\mathrm{A} = \{(x, y) : a \leq x \leq b,\ 0 \leq y \leq g(x)\}$$

est un ensemble quarrable.

Soit en effet $\varphi_1 \leq g \leq \varphi_2$ un encadrement de g par deux fonctions en escalier sur $[a, b]$ telles que $\int_a^b (\varphi_2 - \varphi_1) < \varepsilon$. On peut supposer que $0 \leq \varphi_1$. On a alors un encadrement d'ensembles $\mathrm{B}_1 \subset \mathrm{A} \subset \mathrm{B}_2$ où on a posé $\mathrm{B}_j = \{(x, y) : a \leq x \leq b,\ 0 \leq y \leq \varphi_j(x)\}$ pour $j = 1, 2$. Les ensembles B_1 et B_2 sont des unions finies de rectangles, et on vérifie facilement que $\iint \chi_{\mathrm{B}_j} = \int_a^b \varphi_j$. On a obtenu un encadrement de χ_{A} entre χ_{B_1} et χ_{B_2}, et ces deux fonctions en escalier ont des intégrales aussi voisines que souhaité, donc χ_{A} est intégrable.

En appliquant ce principe plusieurs fois à des découpages convenables, on voit que la plupart des ensembles A limités par des courbes d'équations raisonnables sont quarrables (disques, polygones, etc. . .).

Intégrale sur un sous-ensemble

Soit A un ensemble quarrable et soit f une fonction définie sur A ; soit P un pavé borné contenant A ; désignons (de façon un peu incorrecte) par $\chi_{\mathrm{A}} f$ la fonction égale à f sur A et à 0 en dehors de A. On dira que f est intégrable sur A si $\chi_{\mathrm{A}} f$ est intégrable sur $\mathbb{R}^2$, et on posera

$$\iint_{\mathrm{A}} f(x, y)\, dx\, dy = \iint \chi_{\mathrm{A}} f.$$

4.4. Changement de variable. Coordonnées polaires.

Soient U et V deux ouverts quarrables de $\mathbb{R}^2$ et T une application bijective de U sur V, qui soit de classe C^1 ainsi que la bijection inverse (il en résulte en particulier que la différentielle $d\mathrm{T}_x$ est inversible en tout point $x \in \mathrm{U}$). Si f est une fonction réelle R-intégrable sur l'ensemble V, on a la formule de changement de variables

$$\iint_{\mathrm{V}} f(v)\, dv_1 dv_2 = \iint_{\mathrm{U}} f(\mathrm{T}(u))\, |\det(d\mathrm{T}_u)|\, du_1 du_2.$$

On va essayer d'expliquer (très vaguement) le cas où T est une transformation linéaire bijective. Le point crucial est de se rappeler que si Q est un parallélogramme, la surface (orientée) de l'image T(Q) est égale à celle de Q multipliée par le déterminant de T. De plus, si T est linéaire, sa différentielle $d\mathrm{T}_u$ en tout point $u \in \mathrm{U}$ est égale à T, donc le déterminant jacobien de la formule ci-dessus est constamment égal à $\det \mathrm{T}$ dans le cas linéaire.

On va découper l'intégrale double en petits bouts et appliquer la remarque précédente à chaque petit bout. Puisque l'ouvert V est quarrable, on peut "l'approcher" par une famille de pavés R_k deux à disjoints ; on choisit ensuite un point ξ_k dans R_k, et on a

$$\iint_{\mathrm{V}} f \sim \sum_k |\mathrm{R}_k| f(\xi_k)$$

si le découpage est assez fin. On a alors une approximation de U par la famille de parallélogrammes $\mathrm{P}_k = \mathrm{T}^{-1}(\mathrm{R}_k)$, et une approximation de la fonction g définie sur U par $g(u) = f(\mathrm{T}(u))$ par $\sum_k \chi_{\mathrm{P}_k} f(\xi_k) = \sum_k \chi_{\mathrm{P}_k} g(\xi'_k)$, où $\xi'_k = \mathrm{T}^{-1}(\xi_k)$. On en déduit

$$\iint_{\mathrm{U}} f(\mathrm{T}(u)) \sim \sum_k |\mathrm{P}_k| f(\xi_k),$$

et on conclut par le fait que $|\mathrm{R}_k| = |\det \mathrm{T}|\,|\mathrm{P}_k|$ pour tout k.

Un autre cas raisonnablement facile est celui où la transformation est de la forme $\mathrm{T}(u_1, u_2) = (\mathrm{S}(u_1, u_2), u_2)$. Dans ce cas la formule de calcul des intégrales doubles permet de ramener le théorème à un changement de variable dans une intégrale simple. Certains changements complexes peuvent se ramener à une composition de plusieurs changements de ce type plus simple. C'est le cas par exemple pour les coordonnées polaires qui sont examinées au paragraphe suivant : pour obtenir le demi-plan $\mathrm{V} = \{(x, y) : x > 0\}$, on peut commencer par le changement bijectif $(r, \theta) \to (r, y)$, considéré sur le domaine $\mathrm{U} = \{(r, \theta) : r > 0,\ -\pi/2 < \theta < \pi/2\}$ suivi du changement $(r, y) \to (x, y)$.

Coordonnées polaires

On pose

$$x = r\cos\theta, \quad y = r\sin\theta,$$

où (x, y) varie dans l'ouvert $\mathrm{V} = \mathbb{R}^2 \setminus \{(x, 0) : x \le 0\}$ et (r, θ) dans l'ensemble ouvert $\mathrm{U} =]0, +\infty[\, \times\,]-\pi, \pi[$. On a ici $\mathrm{T}(r, \theta) = (r\cos\theta, r\sin\theta)$. Le déterminant jacobien de T au point $(r, \theta) \in \mathrm{U}$ est égal à r, donc $dx\,dy$ est remplacé par $r\,dr\,d\theta$.

Développons un exemple hyper-classique, le calcul de l'intégrale $\int_{-\infty}^{+\infty} \mathrm{e}^{-x^2/2}\,dx$ par l'intermédiaire du calcul de l'intégrale double de la fonction f définie sur $\mathbb{R}^2$ par la formule $f(x, y) = \mathrm{e}^{-(x^2+y^2)/2}$. On calcule d'abord l'intégrale de la fonction f sur le disque de rayon R

$$\mathrm{D}_\mathrm{R} = \{(x, y) : x^2 + y^2 \le \mathrm{R}^2\}.$$

On obtient par changement de variable polaire

$$\iint_{\mathrm{D}_\mathrm{R}} f = \int_{-\pi}^{\pi} \Bigl(\int_0^{\mathrm{R}} \mathrm{e}^{-r^2/2}\, r\,dr \Bigr) d\theta = 2\pi(1 - e^{-\mathrm{R}^2/2}).$$

On compare cette intégrale à l'intégrale de la fonction f sur les carrés

$$\mathrm{C}_\mathrm{R} = \{(x, y) : |x| \le \mathrm{R}\,;\ |y| \le \mathrm{R}\},$$

qui est égale à $\left(\int_{\mathrm{R}}^{\mathrm{R}} \mathrm{e}^{-x^2/2}\ dx\right)^2$, et en encadrant $\mathrm{D_R} \subset \mathrm{C_R} \subset \mathrm{D}_{\mathrm{R}\sqrt{2}}$ puis en passant à la limite quand $\mathrm{R} \to +\infty$, on obtient

$$\int_{-\infty}^{+\infty} \mathrm{e}^{-x^2/2}\ \frac{dx}{\sqrt{2\pi}} = 1.$$

Cette fonction est appellée *fonction gaussienne* et l'intégrale correspondante *intégrale gaussienne*. Elles jouent un très grand rôle en probabilités et en statistiques.

4.5. Compléments, applications

Intégrale triple

La théorie se développe de façon tout à fait analogue à celle de l'intégrale double. On a en particulier un théorème de calcul des intégrales triples par trois intégrations simples successives. On peut aussi généraliser toute la théorie à $\mathbb{R}^n$.

Longueur d'un arc

Soit $\varphi : [0, \mathrm{T}] \to \mathbb{R}^2$ un arc de courbe paramétré ; si la quantité

$$\mathrm{L_N} = \sum_{k=1}^{\mathrm{N}} \left\| \varphi\left(k\frac{\mathrm{T}}{\mathrm{N}}\right) - \varphi\left((k-1)\frac{\mathrm{T}}{\mathrm{N}}\right) \right\|$$

tend vers une limite L quand $\mathrm{N} \to \infty$, on dit que L est la longueur de l'arc de courbe (la norme utilisée est la norme usuelle sur $\mathbb{R}^2$). Si φ est de classe C^1, on montre que la limite existe et que

$$\mathrm{L} = \int_0^{\mathrm{T}} \|\varphi'(t)\|\, dt.$$

Exercice. Calculer la longueur d'une arche de cycloïde, paramétrée par $x = \mathrm{R}(\theta - \sin(\theta))$, $y = \mathrm{R}(1 - \cos(\theta))$, lorsque θ varie de 0 à 2π.

Volumes et aires de révolution

On considère le volume limité par la surface d'équation

$$r = f(z)$$

en coordonnées cylindriques, où $f \geq 0$ est définie pour z entre deux “altitudes” z_0 et z_1, avec $z_0 < z_1$ et par les deux plans horizontaux $z = z_0$ et $z = z_1$. Le volume limité par cette surface et par les deux plans d'altitude $z = z_1$ et $z = z_2$ est donné par la formule

$$\mathrm{V} = \pi \int_{z_0}^{z_1} (f(z))^2\, dz.$$

Pour comprendre cette formule, il faut imaginer qu'on découpe le volume en petites tranches circulaires découpées par les plans horizontaux d'altitudes z et $z + dz$.
Exemple simple : calculer le volume de la sphère.

Surface de révolution

Dans la même situation que ci-dessus, on peut s'intéresser à la surface de l'objet précédent (sans tenir compte des surfaces des deux “bases”, c'est à dire des deux sections aux altitudes z_1 et z_2). On obtient la formule

$$\mathrm{A} = 2\pi \int_{z_0}^{z_1} f(z)\sqrt{1 + (f'(z))^2}\, dz.$$

Exemple : calculer la surface d'une tranche de sphère, entre deux altitudes z_0 et z_1. Qu'observez-vous ?

4.6. Calculs approchés des intégrales simples

Méthode des trapèzes

Soient f une fonction de classe C^2 sur l'intervalle $[a,b]$, N un entier ≥ 1 et $h = (b-a)/N$. On approxime l'intégrale de f entre a et b par la formule

$$T_N = h\left(\frac{1}{2}f(a) + f(a+h) + f(a+2h) + \cdots + f(a+(N-1)h) + \frac{1}{2}f(b)\right)$$

et on obtient la majoration suivante pour l'erreur

$$\left|\int_a^b f(t)\,dt - T_N\right| \leq \frac{(b-a)M_2}{12}\,h^2$$

où

$$M_2 = \max\{|f''(x)| : a \leq x \leq b\}.$$

La méthode des trapèzes revient à approximer la fonction f sur chaque intervalle de longueur h de la subdivision par une fonction affine qui coïncide avec f en chaque extrémité de l'intervalle.

L'évaluation de l'erreur est fondée sur le calcul suivant : on considère un petit intervalle $[u,v]$, on désigne par m son milieu et on pose $h = v - u$. On désigne par $a(t)$ la fonction affine telle que $a(u) = f(u)$ et $a(v) = f(v)$. On obtient par deux intégrations par parties convenablement choisies un calcul de l'erreur commise dans l'intégrale de u à v en remplaçant f par la fonction affine a :

$$\int_u^v (f(t)-a(t))\,dt = [(f(t)-a(t))(t-m)]_u^v - \int_u^v (f'(t)-a'(t))(t-m)\,dt$$

$$= \left[(f'(t)-a'(t))(h^2/8-(t-m)^2/2)\right]_u^v - \int_u^v f''(t)(h^2/8-(t-m)^2/2)\,dt =$$

$$= -\int_u^v f''(t)(h^2/8-(t-m)^2/2)\,dt.$$

On observe que $h^2/8 - (t-m)^2/2$ reste positif sur l'intervalle $[u,v]$, ce qui permet de majorer la valeur absolue de l'intégrale précédente par

$$M_2\int_u^v (h^2/8-(t-m)^2/2)\,dt = M_2\,\frac{h^3}{12}$$

ce qui donne le résultat annoncé en ajoutant tous les petits bouts...

Méthode de Simpson

Soient f une fonction de classe C^4 sur l'intervalle $[a,b]$, N un entier ≥ 1 et posons $h = (b-a)/(2N)$; on découpe l'intervalle $[a,b]$ en un nombre pair 2N d'intervalles de longueur h. La méthode de Simpson, qui est beaucoup plus performante que la méthode des trapèzes dans la plupart des cas, consiste à approcher f par un trinôme sur chaque groupe de deux intervalles successifs. Puisqu'on est passé d'approximation affine à approximation trinôme, on pourrait s'attendre à ce que l'ordre de l'erreur passe de h^2 à h^3, mais il se produit un petit miracle, qui fait que l'ordre de l'erreur sera en fait h^4.

On approche l'intégrale par

$$S_N = \frac{h}{3}\Big(f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) + \cdots$$
$$+ \cdots + 2f(a+(2N-2)h) + 4f(a+(2N-1)h) + f(b)\Big)$$

et on obtient la majoration suivante pour l'erreur :

$$\left|\int_a^b f(t)\,dt - S_N\right| \leq \frac{(b-a)M_4}{180}\,h^4$$

où

$$M_4 = \max\{|f^{(4)}(x)| : a \leq x \leq b\}.$$

Lemme. *Pour tout polynôme* P *de degré* ≤ 3, *on a l'égalité*

$$\int_{u-h}^{u+h} \mathrm{P}(t)\,dt = \frac{h}{3}\big(\mathrm{P}(u-h) + 4\mathrm{P}(u) + \mathrm{P}(u+h)\big).$$

Démonstration. Soit Q une primitive de P ; alors Q est un polynôme de degré ≤ 4. D'après la formule de Taylor pour les polynômes,

$$\mathrm{Q}(u+h) = \mathrm{Q}(u) + h\mathrm{Q}'(u) + \frac{h^2}{2}\mathrm{Q}''(u) + \frac{h^3}{6}\mathrm{Q}^{(3)}(u) + \frac{h^4}{24}\mathrm{Q}^{(4)}(u),$$

et de même pour $\mathrm{Q}(u-h)$, donc

$$\int_{u-h}^{u+h} \mathrm{P}(t)\,dt = 2h\mathrm{Q}'(u) + \frac{h^3}{3}\mathrm{Q}^{(3)}(u) = 2h\mathrm{P}(u) + \frac{h^3}{3}\mathrm{P}''(u).$$

En appliquant maintenant la formule de Taylor au calcul de $\mathrm{P}(u-h) + 4\mathrm{P}(u) + \mathrm{P}(u+h)$, on vérifie le lemme. Une autre façon de procéder consiste à remarquer d'abord qu'on peut supposer $u = 0$ par translation (la fonction $x \to \mathrm{P}(x-u)$ est une autre fonction polynomiale de même degré), puis qu'il suffit de montrer le lemme pour les quatre fonctions 1, x, x^2 et x^3, puisqu'on obtiendra tous les polynômes de degré ≤ 3 par combinaisons linéaires de ces quatre. Il reste à observer que $\int_{-h}^{h} dt = 2h = \frac{h}{3}(1+4+1)$, $\int_{-h}^{h} t\,dt = 0 = \frac{h}{3}(-h+0+h)$, $\int_{-h}^{h} t^2\,dt = 2h^3/3 = \frac{h}{3}((-h)^2+0+h^2)$, $\int_{-h}^{h} t^3\,dt = 0 = \frac{h}{3}((-h)^3+0+h^3)$.

La formule précise du calcul de l'erreur donnée ci-dessus pour la méthode de Simpson est assez délicate à démontrer ; la preuve suit un chemin analogue à celui que nous avons indiqué pour la méthode des trapèzes. Indiquons une approche plus simple, mais qui fournit de moins bonnes constantes. Soit $[u-h, u+h]$ “un bout” dans l'application de la méthode de Simpson ; considérons la formule de Taylor à l'ordre 4, appliquée à la fonction f au point u. Soit P le polynôme de degré 3 figurant avant le reste (intégral ou de Lagrange) ; nous savons que

$$|f(x) - \mathrm{P}(x)| \le \frac{(x-u)^4}{24}\,\mathrm{M}_4.$$

Il en résulte

$$\Big|\int_{u-h}^{u+h} f(t)\,dt - \int_{u-h}^{u+h} \mathrm{P}(t)\,dt\Big| \le \mathrm{M}_4\,\frac{h^5}{60}.$$

Par ailleurs, en posant $\mathrm{S}_{u,h}(g) = (h/3)(g(u-h) + 4g(u) + g(u+h))$, on obtient

$$|\mathrm{S}_{u,h}(f) - \mathrm{S}_{u,h}(\mathrm{P})| \le \frac{h}{3}\mathrm{M}_4\frac{h^4}{12} = \mathrm{M}_4\,\frac{h^5}{36}.$$

L'erreur sur un “bout” est donc $\le 2\mathrm{M}_4h^5/45$. En multipliant par le nombre N de bouts, on obtient une majoration de l'erreur totale par $(b-a)\mathrm{M}_4h^4/45$.

5. Equations différentielles

On va étudier principalement dans ce chapitre les équations différentielles du premier ordre, qui s'écrivent (symboliquement) sous la forme

$$y' = \mathrm{F}(x, y) \tag{1}$$

où x est une variable réelle, $y \in \mathbb{R}^d$, F est une fonction à valeurs dans $\mathbb{R}^d$, définie sur une partie $\Gamma \subset \mathbb{R} \times \mathbb{R}^d$. L'idée est que y représente une fonction dérivable $x \to y(x)$, dont la dérivée y' doit vérifier la relation (1). Il faut cependant prendre quelques précautions pour définir la notion de solution de l'équation (1).

Exemple. Cherchons les solutions de l'équation $y' = y^2$ telles que $y(0) = a$, avec $a > 0$. On voit que "la" solution "explose" à un point $x_1 > 0$ dépendant de la valeur de a. En effet, on trouve

$$y(x) = \frac{a}{1 - ax}$$

qui tend vers $+\infty$ lorsque $x \to 1/a$ (par valeurs inférieures). On voit donc qu'on ne peut pas imposer en même temps et *a priori* l'intervalle sur lequel on cherche une solution et la valeur de la solution en un point de cet intervalle ; c'est ce qui conduira à faire figurer un intervalle I dans la définition d'une solution. La variable, que nous avons appelée x jusqu'ici, peut aussi bien s'appeler t (si on pense qu'elle représente le *temps* par exemple), mais on réservera la lettre y pour représenter la fonction.

Définition 5.0.1. Une *solution de l'équation* (1) est un couple (I, φ), où I est un intervalle ouvert de $\mathbb{R}$, φ une fonction dérivable de I dans $\mathbb{R}^d$, telle que $(x, \varphi(x)) \in \Gamma$ pour tout $x \in \mathrm{I}$ et telle que

$$\forall x \in \mathrm{I}, \quad \varphi'(x) = \mathrm{F}(x, \varphi(x)).$$

L'une des façons de poser un problème d'équation différentielle est la suivante : on se donne une valeur initiale $x_0 \in \mathbb{R}$ et une donnée initiale $y^{(0)} \in \mathbb{R}^d$, telles que $(x_0, y^{(0)}) \in \Gamma$, et on cherche la ou les solutions (I, φ) telles que $x_0 \in \mathrm{I}$ et $\varphi(x_0) = y^{(0)}$ (problème avec *donnée initiale*). On cherche ensuite à trouver le plus grand intervalle possible sur lequel on puisse définir chaque solution (notion de solution maximale qui sera étudiée plus loin).

Dans certains cas, l'équation est de la forme $y' = \mathrm{F}(y)$, c'est à dire que F ne dépend pas de la variable x. On dit dans ce cas qu'il s'agit d'une équation *autonome*. On verra qu'on peut toujours ramener théoriquement le problème général avec F définie sur $\mathbb{R} \times \mathbb{R}^d$ à un problème autonome, mais où l'espace des valeurs sera $\mathbb{R}^{d+1}$ au lieu de $\mathbb{R}^d$.

Pour une équation autonome, l'évolution d'une solution φ entre x_0 initial et x ne dépend que de la différence $x - x_0$. Pour être plus précis, disons que lorsque (I, φ) est une solution de $y' = \mathrm{F}(y)$, avec $0 \in \mathrm{I}$ et telle que $\varphi(0) = y^{(0)}$, on obtiendra une solution ψ telle que $\psi(x_0) = y^{(0)}$, définie sur un intervalle contenant x_0 en faisant tout simplement une translation,

$$\psi(x) = \varphi(x - x_0).$$

On peut aussi étudier des équations différentielles d'ordre plus élevé, par exemple les équations du second ordre, de la forme $y'' = \mathrm{F}(x, y, y')$. On verra que ces équations se ramènent aussi, au moins du point de vue théorique, à des équations du premier ordre, mais au prix du remplacement de l'espace des valeurs $\mathbb{R}^d$ par un espace de dimension plus grande.

5.1. Equations différentielles linéaires à coefficients constants

On va étudier pour commencer l'une des équations différentielles les plus simples, et dans le cas $d = 1$

$$y' = ay \tag{L1}$$

avec $a \in \mathbb{R}$. Si $a = 0$ on obtient l'équation simple mais importante $y' = 0$: ses seules solutions sont de la forme (I, c), où c est une fonction constante sur l'intervalle I.

Les solutions de (L1) sont bien connues mais on va les introduire par le procédé des approximations successives qui interviendra à nouveau plus loin pour résoudre des équations différentielles très générales. On cherche à résoudre (L1) avec la condition $y(0) = y_0$, où $y_0 \in \mathbb{R}$ est donné. L'idée de base est que l'opération d'intégration est "plus stable" que l'opération de dérivation. Si on a une approximation raisonnable ψ_1 d'une solution φ, on peut espérer améliorer cette approximation en résolvant $\psi_2' = a\psi_1$, puis en remplaçant ψ_1 par ψ_2 ; voici une façon d'expliciter cette idée d'amélioration : supposons que ψ_1 soit une approximation de la solution idéale φ, approximation suffisamment bonne pour avoir, par exemple, le même DL d'ordre 2 au voisinage de 0 que la solution idéale. On voit qu'en définissant ψ_2 par $\psi_2(0) = y_0$ et $\psi_2' = a\psi_1$, c'est maintenant ψ_2' qui a le même DL d'ordre 2 que $\varphi' = a\varphi$, donc par intégration de DL on conclut que le DL d'ordre 3 de ψ_2 au voisinage de 0 coïncide avec celui de φ, et on a bien obtenu une certaine amélioration. On peut ensuite envisager de recommencer cette opération une infinité de fois et essayer d'obtenir une vraie solution à la limite.

On définit donc une suite de fonctions $(\varphi_n)_{n\ge 0}$ en posant $\varphi_0(x) = y_0$ (fonction constante ; c'est le choix le plus simple mais on pourrait prendre à peu près n'importe quoi d'autre) et en définissant φ_{n+1} par $\varphi_{n+1}'(x) = a\varphi_n(x)$ et $\varphi_{n+1}(0) = y_0$. On obtient

$$\varphi_n(x) = y_0\big(1 + ax + \frac{a^2x^2}{2!} + \cdots + \frac{a^nx^n}{n!}\big).$$

On voit que $\varphi_n(x)$ converge vers $y_0\,\mathrm{e}^{ax}$ pour tout $x \in \mathbb{R}$. On pourrait en déduire que le couple $(\mathbb{R}, x \to \mathrm{e}^{ax})$ est solution, mais ce serait idiot tellement cette affirmation est évidente par le calcul (mais cet argument de passage à la limite devra être utilisé plus loin pour le cas général).

Unicité des solutions de L1

Soit (I, φ) une solution de l'équation L1 ; posons $\psi(x) = \mathrm{e}^{-ax}\,\varphi(x)$ pour tout $x \in \mathrm{I}$. On voit que $\psi'(x) = \mathrm{e}^{-ax}(\varphi'(x) - a\varphi(x)) = 0$, donc ψ est constante sur l'intervalle I, ce qui montre qu'il existe une constante C telle que pour tout $x \in \mathrm{I}$, on ait $\varphi(x) = \mathrm{C}\,\mathrm{e}^{ax}$. On a donc montré que les seules solutions de l'équation (L1) sont de la forme $(\mathrm{I}, x \to \mathrm{C}\,\mathrm{e}^{ax})$.

Equation avec second membre

On suppose maintenant donnée une fonction continue g définie sur un intervalle I, et on cherche à résoudre l'équation, dite *équation avec second membre*

$$y' - ay = g(x). \tag{L1S}$$

Supposons que (I, φ) soit solution de (L1S) et posons $\psi(x) = \mathrm{e}^{-ax}\,\varphi(x)$ comme avant. On aura ici

$$\psi'(x) = \mathrm{e}^{-ax}(\varphi'(x) - a\varphi(x)) = \mathrm{e}^{-ax}\,g(x).$$

Supposons que x_0 soit un point de l'intervalle I. On obtiendra en calculant $\psi(x) = \psi(x_0) + \int_{x_0}^{x} \psi'(u)\,du$ l'expression de φ sous la forme

$$\varphi(x) = \mathrm{e}^{a(x-x_0)}\ \varphi(x_0) + \mathrm{e}^{ax} \int_{x_0}^{x} \mathrm{e}^{-au}\, g(u)\, du.$$

Linéarité

Il est clair que si φ_1 et φ_2 sont deux solutions de $y' = ay$ définies sur le même intervalle I, les combinaisons linéaires $\lambda_1\varphi_1 + \lambda_2\varphi_2$ sont également solutions. Cette remarque très simple s'étend immédiatement aux équations linéaires à coefficients variables, $y'(x) = a(x)y(x)$, où a est une fonction continue définie sur un intervalle I. Il n'est pas difficile de résoudre cette équation par calcul de primitive. Si $x \to b(x)$ est une primitive de a sur I, on vérifie que $\varphi(x) = \mathrm{e}^{b(x)}$ est solution. La transformation $\psi(x) = \mathrm{e}^{-b(x)}\,\varphi(x)$ permet encore de montrer l'unicité et de résoudre l'équation avec second membre $y'(x) = a(x)y(x) + g(x)$.

Remarquons que si φ_1 et φ_2 sont deux solutions de l'équation avec second membre $y'(x) = a(x)y(x) + g(x)$, la différence $\varphi_1 - \varphi_2$ est solution de l'équation sans second membre $y'(x) = a(x)y(x)$. D'où le principe suivant : pour résoudre l'équation avec second membre, il suffit de trouver une solution particulière de cette équation avec second membre, et de lui ajouter la “solution générale” de l'équation sans second membre.

Cas complexe

Si on suppose maintenant que $a \in \mathbb{C}$, la fonction $y(x)$ doit être supposée à valeurs complexes. A part ça tout est pareil, y compris le cas avec second membre (et avec une fonction g elle aussi à valeurs complexes).

5.2. Systèmes différentiels linéaires

On considère maintenant l'équation différentielle linéaire du premier ordre générale, qui s'écrit sous la forme

$$\text{(L)} \qquad y' = \mathrm{A}y$$

où A est une matrice réelle $d \times d$, et où y, y' sont aussi des fonctions à valeurs dans $\mathbb{R}^d$. Si on considère les fonctions coordonnées $y_1, \ldots, y_d$ de la fonction vectorielle y, on voit que l'équation (L) est équivalente au *système différentiel* suivant

$$\begin{aligned} y_1'(t) &= a_{1,1}\, y_1(t) + \cdots + a_{1,d}\, y_d(t) \\ y_2'(t) &= a_{2,1}\, y_1(t) + \cdots + a_{2,d}\, y_d(t) \\ &\cdots \\ y_d'(t) &= a_{d,1}\, y_1(t) + \cdots + a_{d,d}\, y_d(t), \end{aligned}$$

où les $(a_{i,j})$ sont les coefficients de la matrice A.

Avant de développer la théorie générale, étudions quelques exemples très simples, dans le cas $d = 2$. Pour aider l'intuition, on imagine que $y(t)$ est la position au temps t d'un point mobile dans $\mathbb{R}^2$ dont la vitesse $y'(t)$ vérifie à tout instant t la relation $y'(t) = \mathrm{A}y(t)$. On va s'intéresser à la trajectoire parcourue par le mobile, et trouver si possible une équation cartésienne pour celle-ci.

Quelques exemples simples avec description des trajectoires

Considérons par exemple le système différentiel

$$y_1' = y_1, \quad y_2' = -y_2.$$

On trouve $y_1(t) = \mathrm{C}\,\mathrm{e}^t$, $y_2(t) = \mathrm{D}\,\mathrm{e}^{-t}$. En supposant $\mathrm{C} \neq 0$, on trouve l'équation cartésienne $y_2 = \mathrm{CD}/y_1$ pour la trajectoire du mobile. C'est une trajectoire hyperbolique. Si on remplace la première équation par $y_1' = ay_1$, avec $a > 0$, on trouve des courbes de la forme $y_2 = \mathrm{C}^{1/a}\mathrm{D}\, {y_1}^{-1/a}$, qui ne sont plus des hyperboles mais ont un peu la même allure générale (les axes de coordonnées sont asymptotes).

Remarquons que dans tous les cas, si on change tous les signes dans la matrice du système (c'est à dire que l'on remplace la matrice A du système par la matrice $-\mathrm{A}$), les trajectoires seront les mêmes mais parcourues dans le sens inverse. Dans tous ces exemples le point 0 est fixe, stable ou instable. Si on étudie le système

$$y_1' = -y_1, \quad y_2' = -2y_2$$

on trouve des trajectoires paraboliques qui tendent vers 0 lorsque $t \to +\infty$ (en fait le mouvement se fait sur une demi-parabole. Si on change les signes, le mouvement "part de 0" lorsque $t = -\infty$ et s'éloigne à l'infini sur la même demi-parabole quand $t \to +\infty$).

Si la matrice A est diagonalisable sur $\mathbb{R}$, on obtiendra un exemple d'un des types précédents en se plaçant dans une base de $\mathbb{R}^2$ formée de vecteurs propres de A. Si la matrice A a la forme de Jordan,

$$y_1' = ay_1 + y_2, \quad y_2' = ay_2,$$

on trouve d'abord $y_2(t) = \mathrm{C}\,\mathrm{e}^{at}$, puis on reporte dans la première équation, en obtenant ainsi une équation avec second membre. Il vient

$$y_1(t) = \mathrm{D}\,\mathrm{e}^{at} + \mathrm{C}t\,\mathrm{e}^{at}.$$

La trajectoire a une équation moins sympathique, de la forme $y_1 = cy_2 + dy_2 \ln(y_2)$. Considérons pour finir le système

$$y_1' = -y_2, \quad y_2' = y_1.$$

Ici la matrice n'a pas de valeur propre réelle, mais on peut étendre le problème en système différentiel complexe, et revenir au problème initial par changement de base. On obtient des trajectoires circulaires. Si on remplace la première équation par $y_1' = -by_2$, avec $b > 0$ et $b \neq 1$, on trouve des trajectoires en spirale.

Cas général. Exponentielle de matrices

Commençons par quelques remarques sur l'équation différentielle linéaire à coefficients constants générale $y' = \mathrm{A}y$. On va s'intéresser plus particulièrement à la résolution de l'équation avec donnée initiale. On se donne un vecteur $y^{(0)} \in \mathbb{R}^d$, et on cherche une solution φ définie sur $\mathbb{R}$ et telle que $\varphi(0) = y^{(0)}$ (par un changement de variable simple, on pourra traiter de même toute condition initiale de la forme $\psi(t_0) = y^{(0)}$, en posant simplement $\psi(t) = \varphi(t - t_0)$).

Remarquons avant toute chose que la fonction nulle est toujours solution de l'équation $y' = \mathrm{A}y$. C'est l'un des aspects de la *linéarité* de l'équation. Les remarques sur la linéarité du cas $d = 1$ sont encore valables ici. Il est clair que si φ_1 et φ_2 sont deux solutions de $y' = \mathrm{A}y$ définies sur le même intervalle I, les combinaisons linéaires $c_1\varphi_1 + c_2\varphi_2$ sont également solutions.

1. Si $y^{(0)}$ est un vecteur propre de A, tel que $\mathrm{A}y^{(0)} = \lambda y^{(0)}$, on voit facilement que

$$\varphi(t) = \mathrm{e}^{\lambda t}\, y^{(0)}$$

est solution de l'équation. Si on a une base $(v^{(1)}, \ldots, v^{(d)})$ de $\mathbb{R}^d$ formée de vecteurs propres de A, avec $\mathrm{A}v^{(j)} = \lambda_j v^{(j)}$ pour $j = 1, \ldots, d$, il est facile de résoudre l'équation avec n'importe quelle donnée initiale $y^{(0)}$. On peut décomposer $y^{(0)}$ dans la base de vecteurs propres,

$$y^{(0)} = \sum_{j=1}^{d} c_j v^{(j)}$$

et on constate que

$$\varphi(t) = \sum_{j=1}^{d} c_j \,\mathrm{e}^{\lambda_j t}\, v^{(j)}$$

est solution de l'équation, avec $\varphi(0) = y^{(0)}$.

2. Si $\lambda = \alpha + i\beta$ est une valeur propre complexe de la matrice A, avec $\alpha, \beta \in \mathbb{R}$, $\beta \neq 0$, il est facile d'en déduire des solutions réelles de l'équation. Soit en effet $z = x + iy \in \mathbb{C}^d$ un vecteur propre complexe de A, avec $x, y \in \mathbb{R}^d$; on voit de la même façon que $\varphi(t) = \mathrm{e}^{\lambda t}\, z$ définit une solution complexe de l'équation. En séparant partie réelle et partie imaginaire, on obtient deux solutions réelles,

$$t \to \mathrm{e}^{\alpha t}(\cos(\beta t)x - \sin(\beta t)y), \quad t \to \mathrm{e}^{\alpha t}(\sin(\beta t)x + \cos(\beta t)y).$$

La combinaison des remarques 1 et 2 permet de traiter le cas où la matrice A est diagonalisable sur $\mathbb{C}$.

3. On peut aussi utiliser la triangulation (sur $\mathbb{R}$ ou sur $\mathbb{C}$, selon le cas). Illustrons ceci sur un exemple simple. Supposons que A soit écrite sous la forme $\mathrm{A} = \mathrm{V}^{-1}\mathrm{TV}$, avec V inversible et T triangulaire supérieure, par exemple

$$\mathrm{A} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} a & 3 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

Supposons donnée la condition initiale $y^{(0)} = (4, 5)$. En posant $u(t) = \mathrm{V}y(t)$, on voit que l'équation $y' = \mathrm{A}y$ se ramène à l'équation triangulaire $u' = \mathrm{T}u$, avec condition initiale $u^{(0)} = \mathrm{V}y^{(0)} = (-1, 9)$, c'est à dire

$$u_1' = au_1 + 3u_2, \quad u_2' = u_2,$$

$u_1(0) = -1$, $u_2(0) = 9$. On obtient immédiatement avec la deuxième équation $u_2(t) = 9\,\mathrm{e}^t$. En reportant dans la première, on a

$$u_1' - au_1 = 27\,\mathrm{e}^t,$$

équation avec second membre qui se résout comme on a dit plus haut. On observera une distinction selon que $a = 1$ ou $a \neq 1$. Pour finir, on écrira $y(t) = \mathrm{V}^{-1}u(t)$. Cette méthode s'applique aussi si on utilise la triangulation complexe. Les calculs intermédiaires font appel à des fonctions complexes, mais la dernière étape (retour à y) doit les faire disparaître.

Approche théorique générale

Le même principe d'approximations successives que dans le cas $d = 1$ fournit ici une suite d'approximations $(\varphi_n)_{n\geq 0}$ à valeurs dans $\mathbb{R}^d$ définies par

$$\varphi_n(t) = \big(\mathrm{I}_d + t\mathrm{A} + \frac{t^2}{2!}\mathrm{A}^2 + \cdots + \frac{t^n}{n!}\mathrm{A}^n\big)\, y^{(0)}$$

pour tout $n \geq 0$ et tout $t \in \mathbb{R}$, et conduit à introduire l'exponentielle de matrice $\mathrm{e}^{t\mathrm{A}}$ pour résoudre le problème. Rappelons que

$$\mathrm{e}^{t\mathrm{A}} = \sum_{n=0}^{+\infty} \frac{t^n}{n!}\,\mathrm{A}^n = \lim_{n\to+\infty} \sum_{k=0}^{n} \frac{t^k}{k!}\,\mathrm{A}^k.$$

Notons que $\mathrm{e}^{t\mathrm{I}_d} = \mathrm{e}^t\,\mathrm{I}_d$ pour tout t réel (ou complexe) ; en particulier, si 0_d désigne la matrice nulle de taille $d \times d$, on a $\mathrm{e}^{0_d} = \mathrm{I}_d$. Rappelons les propriétés qui ont été démontrées dans le poly de MT241.

Lemme. *Si les matrices* A *et* B *commutent, on a*

$$\mathrm{e}^{\mathrm{A}+\mathrm{B}} = \mathrm{e}^{\mathrm{A}}\,\mathrm{e}^{\mathrm{B}}\,.$$

En particulier, la matrice e^{A} *est inversible et son inverse est* $\mathrm{e}^{-\mathrm{A}}$.

On en déduit la propriété importante $\mathrm{e}^{t\mathrm{A}}\,\mathrm{e}^{-t\mathrm{A}} = \mathrm{I}_d$. On rappelle aussi que

$$(**) \qquad \frac{d}{dt}\,\mathrm{e}^{t\mathrm{A}} = \mathrm{A}\,\mathrm{e}^{t\mathrm{A}} = \mathrm{e}^{t\mathrm{A}}\,\mathrm{A}.$$

Il est alors facile de voir à partir des propriétés précédentes la première partie de la proposition qui suit. La deuxième partie concerne l'unicité des solutions.

Proposition 5.2.1. *Pour tout* $y^{(0)} \in \mathbb{R}^d$ *donné, la fonction* φ *à valeurs dans* $\mathbb{R}^d$*, définie sur* $\mathbb{R}$ *par*

$$\forall t \in \mathbb{R}, \quad \varphi(t) = \mathrm{e}^{t\mathrm{A}}\,y^{(0)}$$

est une solution de l'équation (L). *Si* (I, φ) *est une solution de l'équation* $y' = \mathrm{A}y$*, il existe* $y^{(0)} \in \mathbb{R}^d$ *tel que*

$$\forall t \in \mathrm{I}, \quad \varphi(t) = \mathrm{e}^{t\mathrm{A}}\,y^{(0)}.$$

Toute solution (I, φ) *est donc restriction à* I *d'une solution définie sur* $\mathbb{R}$ *par* $t \to \mathrm{e}^{t\mathrm{A}}\,y^{(0)}$.

Démonstration. La première partie résulte immédiatement du calcul de la dérivée de $\mathrm{e}^{t\mathrm{A}}$. Pour la deuxième, on pose comme toujours $\psi(t) = \mathrm{e}^{-t\mathrm{A}}\,\varphi(t)$, qui donne par dérivation (d'un produit matriciel)

$$\psi'(t) = \mathrm{e}^{-t\mathrm{A}}(-\mathrm{A}\varphi(t) + \varphi'(t)) = 0,$$

donc ψ est constante sur I, égale disons à $y^{(0)}$. En multipliant $\psi(t)$ par $\mathrm{e}^{t\mathrm{A}}$, on obtient le résultat voulu.

Résumé. Pour tout intervalle non vide $\mathrm{I} \subset \mathbb{R}$*, pour tout point* $t_0 \in \mathrm{I}$ *et pour tout vecteur* $y^{(0)}$ *donné dans* $\mathbb{R}^d$ *il existe une unique solution* (I, φ) *de l'équation* $y' = \mathrm{A}y$ *vérifiant la condition initiale* $\varphi(t_0) = y^{(0)}$*. Cette solution peut s'exprimer par* $\varphi(t) = \mathrm{e}^{(t-t_0)\mathrm{A}}\,y^{(0)}$.

Linéarité. Equation avec second membre

On a dit que si φ_1 et φ_2 sont deux solutions de l'équation différentielle $y' = \mathrm{A}y$ définies sur le même intervalle I, les combinaisons linéaires $c_1\varphi_1 + c_2\varphi_2$ sont également solutions, donc l'ensemble des solutions de l'équation $y' = \mathrm{A}y$ est un sous-espace vectoriel $\mathcal{E}$ de l'espace des fonctions de I dans $\mathbb{R}^d$; cet espace $\mathcal{E}$ des solutions est de dimension d. En effet pour $t_0 \in \mathrm{I}$ fixé, l'application Φ définie par $\varphi \in \mathcal{E} \to \varphi(t_0) \in \mathbb{R}^d$ est bijective de $\mathcal{E}$ sur $\mathbb{R}^d$ d'après les résultats de la proposition 5.2.1 : étant donné $y^{(0)} \in \mathbb{R}^d$ quelconque, la fonction φ définie par $\varphi(t) = \mathrm{e}^{(t-t_0)\mathrm{A}}\, y^{(0)}$ est une solution telle que $\varphi(t_0) = y^{(0)}$, ce qui montre la surjectivité de Φ, et cette solution est unique, ce qui donne l'injectivité.

Si $v^{(1)}, \ldots, v^{(d)}$ sont des données initiales indépendantes dans $\mathbb{R}^d$, les solutions correspondantes restent indépendantes pour toute valeur de t : c'est clair puisque la matrice $\mathrm{e}^{t\mathrm{A}}$ est inversible.

Exercice. Wronskien. Soient $y^{(1)}, \ldots, y^{(d)}$ des solutions de l'équation $y' = \mathrm{A}y$. Si on désigne par $\mathrm{M}(t)$ la matrice $d \times d$ dont les colonnes sont les vecteurs $y^{(1)}(t), \ldots, y^{(d)}(t)$, montrer que la fonction $f(t) = \det \mathrm{M}(t)$ vérifie l'équation différentielle $f' = (\operatorname{tr} \mathrm{A})\, f$. On appelle ce déterminant le *déterminant Wronskien*.

Une bonne partie des remarques très simples sur la linéarité s'étend encore aux équations linéaires à coefficients variables, $y'(x) = \mathrm{A}(x)y(x)$, où A est une fonction continue définie sur un intervalle I et à valeurs matrices $d \times d$. En revanche, il n'est plus possible en général de résoudre cette équation par calcul de primitive.

Exercice. Soit $x \to \mathrm{B}(x)$ une primitive de $x \to \mathrm{A}(x)$ sur I, c'est à dire une fonction à valeurs matrices dont les coefficients sont des primitives des coefficients correspondants de $\mathrm{A}(x)$. Si pour tout $x \in \mathrm{I}$ la matrice $\mathrm{B}(x)$ commute avec la matrice $\mathrm{A}(x)$, montrer que l'équation différentielle linéaire $y'(x) = \mathrm{A}(x)y(x)$ admet $x \to \mathrm{e}^{\mathrm{B}(x)}\, y^{(0)}$ pour solution.

Remarquons que si φ_1 et φ_2 sont deux solutions de l'équation avec second membre $y'(x) = \mathrm{A}(x)y(x) + g(x)$ (où g est une fonction vectorielle continue sur I), la différence $\varphi_1 - \varphi_2$ est solution de l'équation “sans second membre” $y'(x) = \mathrm{A}(x)y(x)$. On peut donc encore dire que pour résoudre l'équation avec second membre, il suffit de trouver une solution particulière, et de lui ajouter la “solution générale” de l'équation sans second membre.

Dans le cas de l'équation avec second membre à coefficients constants $y' - \mathrm{A}y = g(t)$, où g est une fonction continue définie sur un intervalle I, à valeurs dans $\mathbb{R}^d$, la même méthode que lorsque $d = 1$ permet de résoudre le problème. Si on se donne $t_0 \in \mathrm{I}$, on voit que les solutions s'expriment par la formule

$$\varphi(t) = \mathrm{e}^{(t-t_0)\mathrm{A}}\, \varphi(t_0) + \mathrm{e}^{t\mathrm{A}} \int_{t_0}^{t} \mathrm{e}^{-u\mathrm{A}}\, g(u)\, du.$$

Considérons le problème du ressort, $y'' + y = 0$ ou bien $y'' + y = f(t)$. On va le transformer en système différentiel 2×2 en considérant la donnée vectorielle $\mathrm{Y} = (y, y') = (\mathrm{Y}_1, \mathrm{Y}_2)$. On a les équations

$$\mathrm{Y}_1' = y' = \mathrm{Y}_2\,; \quad \mathrm{Y}_2' = y'' = -y = -\mathrm{Y}_1$$

ce qui donne notre système différentiel. On va voir sur cet exemple comment fonctionnent les méthodes théoriques introduites ci-dessus. Il s'agit dans ce paragraphe d'une *illustration*, ça n'est pas la méthode pratique que vous appliquerez par la suite pour résoudre une équation différentielle du second ordre à coefficients constants.

Tout d'abord, la matrice du système qui traduit l'équation $y'' + y = 0$ est

$$\mathrm{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \text{et} \quad \mathrm{e}^{t\mathrm{A}} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}$$

(l'exponentielle est une rotation). On trouve donc la solution générale du système homogène,

$$\mathrm{e}^{t\mathrm{A}} \begin{pmatrix} a \\ b \end{pmatrix} = a \begin{pmatrix} \cos t \\ -\sin t \end{pmatrix} + b \begin{pmatrix} \sin t \\ \cos t \end{pmatrix}.$$

Passons à l'équation du ressort dont l'oscillation est forcée par un terme $f(t)$, l'équation étant alors $y'' + y = f(t)$. En passant au système on a

$$\mathrm{Y}' = \mathrm{AY} + \begin{pmatrix} 0 \\ f(t) \end{pmatrix}.$$

On peut théoriquement résoudre par la formule

$$\mathrm{e}^{t\mathrm{A}} \int_0^t \mathrm{e}^{-u\mathrm{A}}\, g(u)\, du \quad \text{où} \quad g(u) = \begin{pmatrix} 0 \\ f(u) \end{pmatrix} ;$$

dans le cas particulier où $f(t) = \sin(\omega t)$, les calculs sont pénibles mais faisables. On aperçoit le problème de résonance quand $\omega = 1$.

Calcul pratique d'exponentielles de matrices

On se sert de la théorie de la réduction des matrices complexes, qui utilise la décomposition de $\mathbb{C}^d$ en sous-espaces caractéristiques de la matrice A et la triangulation des blocs caractéristiques. On sait ainsi que la matrice A peut s'écrire $\mathrm{A} = \mathrm{VMV}^{-1}$, où V est inversible, et où M est une matrice diagonale par blocs, où chaque bloc diagonal est une matrice triangulaire T_α, avec une valeur constante $\lambda_\alpha \in \mathbb{C}$ sur la diagonale (λ_α est l'une des valeurs propres (complexes) de la matrice A). On obtient alors $\mathrm{e}^{t\mathrm{A}} = \mathrm{V}\mathrm{e}^{t\mathrm{M}}\,\mathrm{V}^{-1}$, et $\mathrm{e}^{t\mathrm{M}}$ est diagonale par blocs, chaque bloc étant égal à $\mathrm{e}^{t\mathrm{T}_\alpha}$. Il reste à expliquer l'allure de $\mathrm{e}^{t\mathrm{T}}$, où T est l'un quelconque de ces blocs diagonaux de M, de taille $p \times p$. On peut écrire $\mathrm{T} = \lambda \mathrm{I}_p + \mathrm{N}$, où N est nilpotente, ce qui donnera puisque $\lambda \mathrm{I}_p$ commute avec N et que $\mathrm{N}^p = 0$

$$\mathrm{e}^{t\mathrm{T}} = \mathrm{e}^{t\lambda \mathrm{I}_p}\, \mathrm{e}^{t\mathrm{N}} = \mathrm{e}^{t\lambda}\big(\mathrm{I}_p + t\mathrm{N} + \frac{t^2}{2!}\mathrm{N}^2 + \cdots + \frac{t^{p-1}}{(p-1)!}\mathrm{N}^{p-1}\big).$$

Si T a la forme de Jordan on a une information un peu plus précise.

Conséquence. *Si toutes les valeurs propres (réelles ou complexes) de la matrice* A *sont de partie réelle* < 0, *la matrice* $\mathrm{e}^{t\mathrm{A}}$ *tend vers* 0 *lorsque* $t \to +\infty$ *(on verra plus bas que c'est si et seulement si). Il en résulte que dans ce cas toutes les solutions de* $y' = \mathrm{A}y$ *tendent vers* 0 *lorsque* $t \to +\infty$, *quelle que soit la donnée initiale* $y^{(0)}$.

Pour montrer que $\mathrm{e}^{t\mathrm{A}}$ tend vers 0 quand $t \to +\infty$, il suffit de montrer que $\mathrm{e}^{t\mathrm{T}}$ tend vers 0 pour chacun des blocs triangulaires $\mathrm{T} = \mathrm{T}_\alpha$ de la matrice M introduite ci-dessus ; on a supposé que chacune des valeurs propres λ de A a une partie réelle < 0, donc $\lambda = a + ib$ avec $a < 0$. On a donc $|\,\mathrm{e}^{t\lambda}\,| = \mathrm{e}^{at}$ qui tend vers 0 rapidement, ce qui entraîne que $\mathrm{e}^{t\lambda}\, t^j$ tend vers 0 quand $t \to +\infty$ pour tout $j = 0, \ldots, p-1$, donc chacun des morceaux de la décomposition de $\mathrm{e}^{t\mathrm{T}}$ donnée ci-dessus tend vers 0. Inversement, si $\mathrm{e}^{t\mathrm{A}}$ tend vers 0 quand $t \to +\infty$, toutes les valeurs propres de la matrice A ont une partie réelle < 0 : si $\lambda = a + ib$ est valeur propre, il existe un vecteur Z complexe non nul tel que $\mathrm{AZ} = \lambda \mathrm{Z}$, et alors $e^{t\mathrm{A}}\mathrm{Z} = \mathrm{e}^{t\lambda}\,\mathrm{Z}$ tend vers 0, donc $\mathrm{e}^{t\lambda}$ tend vers 0 quand $t \to +\infty$, ce qui entraîne $a < 0$.

Equations différentielles linéaires d'ordre plus élevé, à coefficients constants

Considérons d'abord l'équation du second ordre

$$y'' + py' + qy = 0$$

avec $p, q \in \mathbb{R}$. On peut tranformer cette équation en système, en posant $y_1 = y$ et $y_2 = y'$,

$$y_1' = y_2, \quad y_2' = -qy_1 - py_2.$$

On obtient que la matrice du système est la matrice transposée de la matrice compagnon du polynôme $\mathrm{P} = q + p\mathrm{X} + \mathrm{X}^2$. Les valeurs propres de la matrice sont donc les racines du trinôme P. Si les valeurs propres λ, μ sont distinctes, les solutions complexes sont de la forme $c\,\mathrm{e}^{\lambda t} + d\,\mathrm{e}^{\mu t}$. Si ces valeurs propres ne sont pas réelles (quand $p^2 - 4q < 0$), on a $\lambda = \alpha + i\beta$ et $\mu = \alpha - i\beta$, avec $\alpha, \beta \in \mathbb{R}$ et les solutions réelles sont de la forme

$$\mathrm{e}^{\alpha t}(c\cos(\beta t) + d\sin(\beta t)).$$

Si α est racine double du trinôme, les solutions sont de la forme $\mathrm{e}^{\alpha t}(c + dt)$.

Exemple : oscillation d'un ressort avec amortissement. Résoudre l'équation $y'' + fy' + y = 0$, où $f \geq 0$, selon la valeur de $f \geq 0$ (qui représente un terme de frottement ; oscillations amorties, cas apériodique selon la valeur de f).

Considérons maintenant l'équation différentielle linéaire à coefficients constants d'ordre quelconque. L'équation différentielle d'ordre n

$$(*) \qquad y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y' + a_0 y = 0$$

se ramène à un système $n \times n$, dont la matrice A est la transposée de la matrice compagnon M(P) du polynôme

$$\mathrm{P} = \mathrm{X}^n + a_{n-1}\mathrm{X}^{n-1} + \cdots + a_1\mathrm{X} + a_0.$$

Les valeurs propres sont donc les racines du polynôme P.

On compare la résolution de l'équation $(*)$ à celle du système

$$(**) \qquad \mathrm{Y}' = \mathrm{AY}.$$

On établit facilement les principes suivants :

1– Si (I, φ) est solution de l'équation $(*)$, et si on définit la fonction vectorielle Φ sur I par $\Phi(t) = (\varphi(t), \varphi'(t), \ldots, \varphi^{(n-1)}(t)) \in \mathbb{R}^n$ (ou $\mathbb{C}^n$), le couple (I, Φ) est solution du système $(**)$.

2– Si (I, Φ) est solution du système $(**)$, et si $\Phi(t) = (\varphi_1(t), \ldots, \varphi_n(t))$, alors (I, φ_1) est solution de l'équation $(*)$ (et en fait chacune des fonctions coordonnées φ_j donne une solution).

La correspondance $\Phi \to \varphi_1$ est donc une bijection linéaire de l'espace vectoriel des solutions de $(**)$ sur l'espace des solutions de l'équation $(*)$. Cet espace est donc de dimension n (réel ou complexe, selon le cadre fixé). De plus, si t_0 est un point fixé de I, l'application $\varphi \to (\varphi(t_0), \varphi'(t_0), \ldots, \varphi^{(n-1)}(t_0))$ est une bijection linéaire de l'espace des solutions de $(*)$ sur $\mathbb{R}^n$ (ou $\mathbb{C}^n$).

On rappelle que les valeurs propres de A sont les racines de P. Si $\lambda \in \mathbb{C}$ est une telle racine, on voit facilement que $\mathrm{X} = (1, \lambda, \lambda^2, \ldots, \lambda^{n-1})$ vérifie $\mathrm{AX} = \lambda\mathrm{X}$, ce qui donne une solution $\Phi(t) = \mathrm{e}^{\lambda t}\,\mathrm{X}$ pour le système, donc $\varphi(t) = \mathrm{e}^{\lambda t}$ pour l'équation $(*)$.

Si les racines $(\lambda_1, \ldots, \lambda_n)$ de P sont deux à deux distinctes, on obtient ainsi une base de l'espace des solutions, formée des n fonctions $t \to \mathrm{e}^{\lambda_j t}$, pour $j = 1, \ldots, n$.

Exercice. Retrouver l'indépendance des solutions dans ce cas en utilisant l'isomorphisme ci-dessus pour $t_0 = 0$ (on rencontrera le déterminant classique de Vandermonde).

Cas des racines multiples pour P

Si
$$\mathrm{P} = (\mathrm{X} - \lambda_1)^{r_1} \dots (\mathrm{X} - \lambda_p)^{r_p}$$
est la décomposition de P avec $(\lambda_1, \dots, \lambda_p)$ racines distinctes, on obtient une base de l'espace des solutions de $(*)$ en faisant la liste des fonctions
$$t \to \mathrm{e}^{\lambda_j t}, t\,\mathrm{e}^{\lambda_j t}, \dots, t^{r_j - 1}\,\mathrm{e}^{\lambda_j t},$$
pour $j = 1, \dots, p$.

5.3. Le cas général du premier ordre. Théorème de Cauchy-Lipschitz

L'objectif principal de ce paragraphe est d'obtenir un théorème général pour l'existence de solution pour l'équation générale $y' = \mathrm{F}(x, y)$. On parlera surtout du cas de dimension un (la fonction y est réelle, et F est alors une fonction de deux variables réelles), mais il faut aussi comprendre l'importance théorique du cas où y est une fonction vectorielle. Les résultats vectoriels donnent en particulier l'existence de solutions pour des équations différentielles d'ordre plus élevé.

Supposons que l'on veuille étudier une équation générale du second ordre, qui se présente sous la forme
$$y'' = \mathrm{G}(x, y, y')$$
où G est une fonction continue définie sur un ouvert de $\mathbb{R}^3$, et où y représente une fonction réelle d'une variable. On peut ramener cette équation à une équation vectorielle du premier ordre, par une transformation tout à fait analogue à celle qui a été utilisée pour ramener une équation linéaire du second ordre à un système linéaire du premier ordre. On définit une fonction vectorielle F sur une partie de $\mathbb{R} \times \mathbb{R}^2$, à valeurs dans $\mathbb{R}^2$, par
$$\mathrm{F}(x, y_1, y_2) = \begin{pmatrix} y_2 \\ \mathrm{G}(x, y_1, y_2) \end{pmatrix}.$$
Si $x \to \Phi(x) = (\varphi_1(x), \varphi_2(x))$ vérifie l'équation du premier ordre vectorielle $\mathrm{Y}' = \mathrm{F}(x, \mathrm{Y})$, on voit alors que φ_1 est solution de l'équation $y'' = \mathrm{G}(x, y, y')$. Une transformation du même type permet aussi de ramener une équation non homogène à une équation homogène.

On va considérer dans ce paragraphe l'équation $y'(x) = \mathrm{F}(x, y(x))$, où F est définie et continue sur un ouvert $\Gamma \subset \mathbb{R} \times \mathbb{R}^d$. Remarquons que si (I, φ) est solution, la fonction φ est de classe C^1 sur I, puisque $x \to \varphi'(x) = \mathrm{F}(x, \varphi(x))$ est obtenue comme composition des fonctions continues $x \to (x, \varphi(x))$ et F.

Exercice. Si F est de classe C^k, montrer que toute solution φ de $y' = \mathrm{F}(x, y)$ est de classe C^{k+1}. Si F est C^∞, montrer que toute solution est C^∞ aussi.

Problème d'unicité

Une question qui se pose immédiatement est celle de l'unicité de la solution correspondant à une donnée initiale. Il est intéressant de noter que les deux questions d'existence et d'unicité trouveront une réponse raisonnablement simple dans le cas où F est lipschitzienne (dans la variable y). On peut obtenir des théorèmes plus généraux pour l'existence de solutions, mais ils sont très largement hors programme.

Considérons l'équation $y' = 2|y|^{1/2}$, avec la condition initiale $y(0) = 0$. La fonction φ définie sur $\mathbb{R}$ par $\varphi(t) = 0$ est solution, mais ψ définie par $\psi(t) = -t^2$ pour $t \leq 0$ et

$\psi(t) = t^2$ pour $t > 0$ est une autre solution. En fait, on trouve 4 solutions en changeant de branche au point 0.

Plus généralement, cherchons à résoudre cette équation avec la condition initiale $y(x_0) = y_0 > 0$. Si φ est solution dans un voisinage de x_0 et vérifie $\varphi(x_0) = y_0 > 0$, on aura $\varphi > 0$ dans un voisinage de x_0, ce qui donnera $\varphi'/(2\sqrt{\varphi}) = 1$ dans le même voisinage, donc

$$\varphi(x) = (x - x_0 + \sqrt{y_0})^2$$

dans ce voisinage. On peut considérer cette solution tant que φ' reste > 0, c'est à dire lorsque $x > x_0 - \sqrt{y_0}$. De la même façon, si on cherche une solution telle que $\varphi(x_1) = y_1 < 0$, on trouvera une fonction de la forme $\varphi(x) = -(x-b)^2$, qui sera solution pour $x < b$. On peut chercher à raccorder deux telles solutions, on posant $\varphi(x) = (x-a)^2$ pour $x > a$, puis $\varphi(x) = -(x-b)^2$ pour $x \le b < a$ et $\varphi(x) = 0$ si $b \le x \le a$. On vérifie à la main que cette fonction est solution sur $\mathbb{R}$ tout entier.

On remarque que 0 est aussi solution, et qu'au lieu de continuer par $(x-a)^2$ pour $x > a$, on aurait pu définir une autre solution en posant $\varphi_1(x) = 0$ pour $x > a$; de même, on pourrait avoir $\varphi(x) = 0$ pour $x \le a$ et $\varphi(x) = (x-a)^2$ pour $x > a$. On obtient ainsi une famille de solutions $\varphi_{a,b}$, où a est réel ou $+\infty$ et $b \le a$ est réel ou $-\infty$, définies par $\varphi_{a,b}(x) = -(x-b)^2$ pour $x \le b$, $\varphi_{a,b}(x) = 0$ si $b < x < a$ et $\varphi_{a,b}(x) = (x-a)^2$ si $a \le x$. On pourra montrer plus loin qu'on a bien obtenu ainsi toutes les solutions sur $\mathbb{R}$ de cette équation.

Pour chaque donnée initiale $\varphi(x_0) = 0$ passe quatre solutions qui sont deux à deux distinctes dans tout intervalle ouvert I contenant x_0 : il n'y a pas d'unicité locale dans cet exemple. On va voir que ces difficultés de non unicité sont évitées lorsque la fonction F vérifie une condition de Lipschitz convenable.

Définition 5.3.1. Soit W un sous-ensemble non vide de $\mathbb{R} \times \mathbb{R}^d$; on dit qu'une fonction $\mathrm{F} : (x,y) \to \mathrm{F}(x,y) \in \mathbb{R}^d$ définie sur W est M-*lipschitzienne en* y si

$$\|\mathrm{F}(x,y_1) - \mathrm{F}(x,y_2)\| \le \mathrm{M}\,\|y_1 - y_2\|$$

pour tous $(x,y_1) \in \mathrm{W}$, $(x,y_2) \in \mathrm{W}$. On dit que F est lipschitzienne en y sur l'ensemble W s'il existe une constante M telle que F soit M-lipschitzienne en y sur W.

Exercice. Si W est ouvert, on dit que F est *localement lipschitzienne en* y *sur l'ensemble* W si pour tout point $(x,y) \in \mathrm{W}$ il existe un voisinage $\mathrm{W}_1 \subset \mathrm{W}$ de (x,y) dans lequel F est lipschitzienne en y. En utilisant le théorème de Bolzano-Weierstrass, montrer qu'une fonction F localement lipschitzienne en y sur un ouvert $\Gamma \subset \mathbb{R} \times \mathbb{R}^d$ vérifie une estimation globale sur tout fermé borné $\mathrm{K} \subset \Gamma$, c'est à dire qu'il existe M (dépendant de K) tel que pour tous (x,y_1), (x,y_2) dans K on ait $|\mathrm{F}(x,y_1) - \mathrm{F}(x,y_2)| \le \mathrm{M}\,|y_1 - y_2|$.

Théorème 5.3.1. (Lemme de Gronwall). *Soit f une fonction réelle continue sur* $[0,\mathrm{T}]$ *telle que*

$$\forall t \in [0,\mathrm{T}], \quad f(t) \le \mathrm{C} + \mathrm{D}\int_0^t f(u)\,du,$$

où C *et* D *sont deux constantes, avec* $\mathrm{D} \ge 0$. *On a alors*

$$\forall t \in [0,\mathrm{T}], \quad f(t) \le \mathrm{C}\,\mathrm{e}^{\mathrm{D}t}\,.$$

Démonstration. On montre d'abord que si une fonction continue h définie sur $[0,\mathrm{T}]$ vérifie

$$h(t) \le \mathrm{D}\int_0^t h(u)\,du$$

pour tout $t \in [0, \mathrm{T}]$, on a $h(t) \leq 0$ pour tout $t \in [0, \mathrm{T}]$; posons $\mathrm{H}(t) = \int_0^t h(u)\,du$ et $\psi(t) = \mathrm{e}^{-\mathrm{D}t}\,\mathrm{H}(t)$. Alors H est dérivable et $\mathrm{H}' = h$ (parce que h est continue), $\psi(0) = 0$ et $\psi'(t) = \mathrm{e}^{-\mathrm{D}t}\big(h(t) - \mathrm{DH}(t)\big) \leq 0$ pour tout $t \in]0, \mathrm{T}[$, donc $\psi \leq 0$ sur $[0, \mathrm{T}]$, donc $\mathrm{H} \leq 0$ sur $[0, \mathrm{T}]$. Puisque $\mathrm{D} \geq 0$ et $h \leq \mathrm{DH}$, il en résulte bien que $h \leq 0$ sur $[0, \mathrm{T}]$

On suppose ensuite que f vérifie l'hypothèse du lemme de Gronwall, et on pose $h(t) = f(t) - \mathrm{C}\,\mathrm{e}^{\mathrm{D}t}$. On voit que $h(t) \leq \mathrm{D}\int_0^t h(u)\,du$ pour tout $t \in [0, \mathrm{T}]$, d'où le résultat par ce qui précède.

ATTENTION. Le résultat n'est pas vrai si $\mathrm{D} < 0$.

Corollaire 5.3.1. *Soit g une fonction de classe C^1 sur un intervalle $]a, b[$, à valeurs réelles ou vectorielles, et telle que $\|g'(t)\| \leq \mathrm{D}\,\|g(t)\|$ pour tout $t \in]a, b[$. On a alors pour tout $t_0 \in]a, b[$ fixé*

$$\forall t \in]a, b[, \quad \|g(t)\| \leq \|g(t_0)\|\,\mathrm{e}^{\mathrm{D}|t-t_0|}.$$

Démonstration. Traitons le cas où t est entre t_0 et b ; pour u entre 0 et $\mathrm{T} = b - t_0 > 0$ on écrit $f(u) = \|g(t_0 + u)\|$ et

$$f(u) = \Big\|g(t_0) + \int_0^u g'(t_0+s)\,ds\Big\| \leq \|g(t_0)\| + \int_0^u \|g'(t_0+s)\|\,ds \leq \|g(t_0)\| + \mathrm{D}\int_0^u f(s)\,ds,$$

puis on applique le lemme de Gronwall à f. On procède de même de l'autre côté, lorsque t est entre a et x_0, en posant cette fois $f(u) = \|g(t_0 - u)\|$.

Corollaire 5.3.2. *Soit g une fonction de classe C^1 sur un intervalle ouvert* I, *à valeurs réelles ou vectorielles, telle que $g(t_0) = 0$ pour un certain point $t_0 \in \mathrm{I}$, et telle que $\|g'(t)\| \leq \mathrm{D}\,\|g(t)\|$ pour tout point $t \in \mathrm{I}$. On a alors $g = 0$ sur* I.

Unicité, comparaison de solutions partant de points voisins

Supposons que $\mathrm{F}(x, y)$ soit M-lipschitzienne en y pour (x, y) dans un voisinage V de (t_0, y_0) et soient φ_1, φ_2 deux solutions définies dans un même intervalle I contenant t_0, et telles que $(t, \varphi_1(t)) \in \mathrm{V}$, $(t, \varphi_2(t)) \in \mathrm{V}$ pour tout $t \in \mathrm{I}$; on aura en posant $g(s) = \varphi_1(t_0 + s) - \varphi_2(t_0 + s)$

$$\|g'(s)\| = \|\varphi_1'(t_0 + s) - \varphi_2'(t_0 + s)\| = \|\mathrm{F}(t_0 + s, \varphi_1(t_0 + s)) - \mathrm{F}(t_0 + s, \varphi_2(t_0 + s))\| \leq$$

$$\leq \mathrm{M}\,\|\varphi_1(t_0 + s) - \varphi_2(t_0 + s)\| = \mathrm{M}\,\|g(s)\|.$$

D'après le premier des deux corollaires précédents, ceci permet de conclure pour tout point t de I

$$\|\varphi_1(t) - \varphi_2(t)\| \leq \|\varphi_1(t_0) - \varphi_2(t_0)\|\,\mathrm{e}^{\mathrm{M}|t-t_0|}.$$

Ce résultat donne en particulier l'unicité : si φ_1 et φ_2 sont deux solutions telles que $\varphi_1(t_0) = \varphi_2(t_0)$, l'inégalité précédente signifie que $\varphi_1(t) = \varphi_2(t)$ pour tout $t \in \mathrm{I}$.

Proposition 5.3.1. Unicité locale. *Supposons* F *continue et* M*-lipschitzienne en y dans un voisinage* V *de $(t_0, y^{(0)})$. Si φ et ψ sont deux solutions définies dans un intervalle ouvert* I *contenant t_0 et si $\varphi(t_0) = \psi(t_0) = y^{(0)}$, il existe un intervalle ouvert* $\mathrm{J} \subset \mathrm{I}$ *contenant t_0 tel que*

$$\forall t \in \mathrm{J}, \quad \varphi(t) = \psi(t).$$

Démonstration. Les fonctions φ et ψ étant solutions sont dérivables par définition, donc continues au point t_0, donc on peut trouver un petit intervalle J contenant t_0 tel que $(t, \varphi(t)) \in \mathrm{V}$ et $(t, \psi(t)) \in \mathrm{V}$ pour tout $t \in \mathrm{J}$. On peut alors appliquer le résultat ci-dessus.

Théorème 5.3.2. *Soit* (I, φ) *une solution de l'équation* $y' = \mathrm{F}(x, y)$. *On suppose que* F *est lipschitzienne en* y *au voisinage de chaque point* $(x, \varphi(x))$, $x \in \mathrm{I}$. *Si* x_0 *est un point de* I *et* (I, ψ) *une solution telle que* $\psi(x_0) = \varphi(x_0)$, *on aura* $\psi = \varphi$ *sur* I.

Démonstration. Soit $b \in \mathrm{I}$ tel que $x_0 < b$ et posons

$$\mathrm{A} = \{c \in [x_0, b] : \varphi = \psi \text{ sur } [x_0, c[\}.$$

D'après l'hypothèse et la proposition précédente, on sait que A contient au moins un point $c > x_0$. L'ensemble A est majoré par b par définition. On pose $m = \sup \mathrm{A}$, et on va montrer que $m = b$. On vérifie d'abord que $\psi = \varphi$ sur $[x_0, m[$: pour tout x tel que $x_0 \le x < m$, il existe, par définition de m comme borne supérieure de A, au moins un point $c \in \mathrm{A}$ tel que $x < c$, donc $\psi(x) = \varphi(x)$ puisque ψ et φ coïncident sur $[x_0, c[$. Il en résulte que $\psi(m) = \varphi(m)$ par continuité. Si on avait $m < b$, on pourrait d'après la proposition précédente prolonger l'égalité sur un voisinage de m, donc au delà de m, ce qui contredirait la définition de m.

On a donc $m = b$, donc $\psi = \varphi$ sur $[x_0, b[$, et ceci pour tout b de I plus grand que x_0. On procède de même à gauche de I, et on a ainsi obtenu le résultat annoncé.

Application. Suite et fin de l'exercice $y' = 2\sqrt{|y|}$. Bien entendu, la fonction $\mathrm{F}(x, y) = 2|y|^{1/2}$ du contre-exemple à l'unicité mentionné plus haut n'est pas lipschitzienne en y au voisinage des points de la forme $(x_0, 0)$. On va voir comment le théorème précédent permet néanmoins de prouver que l'on avait bien identifié toutes les solutions de cette équation (on va considérer les solutions définies sur $\mathbb{R}$; le lecteur essaiera d'adapter à un intervalle quelconque).

Si $(\mathbb{R}, \psi)$ est une solution de cette équation, on va montrer qu'elle est nécessairement de l'un des types $\varphi_{a,b}$ évoqués précédemment. L'une des possibilités est $\psi = 0$. Sinon, supposons que $\psi(x_0) \neq 0$. Si $\psi(x_0) > 0$, on trouve une solution $(]a, +\infty[, \varphi)$ de la forme $\varphi(x) = (x - a)^2$ qui coïncide avec ψ au point x_0 (prendre $a = x_0 - \sqrt{\psi(x_0)}$) : en effet, la fonction $\mathrm{F}(x, y) = 2\sqrt{|y|}$ est lipschitzienne en y au voisinage de chacun des points $(x, (x - a)^2)$, $x > a$, donc d'après le théorème précédent on déduit que $\psi(x) = (x - a)^2$ pour tout $x > a$, et $\psi(a) = 0$ par continuité.

Si $\psi(x_0) < 0$, on montre de la même façon que $\psi(x) = -(x - b)^2$ pour tout $x < b$. Ces deux arguments permettent d'identifier toutes les solutions.

Si on a deux équations différentielles $y' = \mathrm{F}_1(y)$ et $y' = \mathrm{F}_2(y)$ telles que $\mathrm{F}_1(y) - \mathrm{F}_2(y) = o(y)$ pour y voisin de 0, on peut faire des choses... Si par exemple $\mathrm{F}_1(y) = -\sin(y)$, et si $\mathrm{F}_2(y) = -y$, on peut chercher à comparer le comportement des solutions de l'équation plus compliquée $y' = -\sin(y)$ à celui des solutions de l'équation simple $y' = -y$. Pour la deuxième on trouve e^{-t} comme solution. Soit donc y une solution de la première et posons $z(t) = \mathrm{e}^t\, y(t)$; on aura comme toujours $z'(t) = \mathrm{e}^t(y' + y) = \mathrm{e}^t(y - \sin(y))$, et on peut majorer $|z'(t)| \le \mathrm{e}^t\, \varepsilon |y(t)| = \varepsilon |z(t)|$ quand $y(t)$ se trouve dans un petit voisinage de 0. Par Gronwall on obtient $|z(t)| \le |z(0)|\, \mathrm{e}^{\varepsilon t}$, ce qui donnera $|y(t)| \le \mathrm{e}^{-(1-\varepsilon)t}$, montrant au moins que $y(t)$ tend vers 0 lorsque t tend vers $+\infty$.

Cette discussion se rattache au problème de la *stabilité*. On suppose que $\mathrm{F}(0) = 0$. L'équation autonome $y' = \mathrm{F}(y)$ admet alors la solution évidente $y = 0$. On dit que le point 0 est stable si pour $y^{(0)}$ assez voisin de $y = 0$, les solutions de $y' = \mathrm{F}(y)$ vérifiant la condition initiale $y(0) = y^{(0)}$ tendent vers 0 lorsque t tend vers $+\infty$. Dans les bons cas, cette étude peut se faire en comparant à une équation linéaire. C'est ce que nous venons d'indiquer pour l'équation $y' = -\sin(y)$.

Si on considère l'équation autonome $y' = \mathrm{F}(y)$ au voisinage de $y = 0$ et si la différentielle $(d\mathrm{F})_0$ a une matrice dont toutes les valeurs propres (réelles ou complexes) ont une partie réelle < 0, il existe un voisinage V de 0 tel que toute solution partant de $y^{(0)} \in \mathrm{V}$ au temps $t = 0$ tende vers le point 0 quand $t \to +\infty$.

La démonstration est un peu délicate et nous allons seulement l'évoquer. On a d'abord besoin du lemme suivant.

Lemme. *Si A est la matrice d'un endomorphisme u d'un espace vectoriel réel E, et si on a $\mathrm{Re}\,\lambda \le \alpha$ pour toute valeur propre λ (réelle ou complexe) de A, il existe pour tout $\varepsilon > 0$ une base de E telle qu'en désignant par B la nouvelle matrice de u dans cette base, on ait*

$$\mathrm{X}\,.\,\mathrm{BX} \le (\alpha + \varepsilon)\,\mathrm{X}\,.\,\mathrm{X}$$

pour tout $\mathrm{X} \in \mathbb{R}^n$.

Idée de démonstration, dans le cas où la matrice A est diagonalisable sur $\mathbb{C}$ (cette propriété ne dépend pas de la base choisie : elle peut s'exprimer en disant que le polynôme minimal a toutes ses racines –y compris les racines complexes– simples). On sait qu'on peut trouver une base de E dans laquelle la nouvelle matrice $\mathrm{B} = \mathrm{V}^{-1}\mathrm{AV}$ est diagonale par blocs de taille 1×1 (pour les valeurs propres réelles) ou de taille 2×2 (pour les valeurs propres complexes non réelles), et de façon que les blocs 2×2 soient de la forme

$$\mathrm{C} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

avec $b \neq 0$; un tel bloc correspond à un couple de valeurs propres complexes $\lambda = a \pm ib$. Par hypothèse on a $a \le \alpha$ pour chacun de ces blocs 2×2 et $\lambda \le \alpha$ pour les valeurs propres réelles. Le lecteur vérifiera que dans cette base on a bien le résultat annoncé (la raison est que pour $\mathrm{Y} \in \mathbb{R}^2$, on a $\mathrm{Y}\,.\,\mathrm{CY} = a\,\mathrm{Y}\,.\,\mathrm{Y} \le \alpha\mathrm{Y}\,.\,\mathrm{Y}$. On devra ensuite ajouter les contributions des différents blocs, y compris les blocs de taille 1 pour lesquels la majoration est directe).

Dans le cas général, on peut y arriver en montrant que toute matrice réelle A peut être arbitrairement approchée par des matrices réelles A' qui ont des valeurs propres complexes deux à deux distinctes (et sont donc diagonalisables sur $\mathbb{C}$; c'est un exercice un peu difficile, que l'on peut faire en raisonnant par récurrence sur la taille de la matrice). C'est ce remplacement de A par A' qui introduira la perturbation $\varepsilon > 0$ dans l'énoncé, alors qu'elle n'apparaissait pas dans le cas précédent.

Supposons maintenant que l'endomorphisme $(d\mathrm{F})_0$ de $\mathbb{R}^d$ ait une matrice A telle que $\mathrm{Re}\,\lambda \le \alpha < 0$ pour toute valeur propre λ de A. On raisonnera alors en choisissant de tout exprimer dans la base $\mathbf{f}$ de $\mathbb{R}^d$ donnée par le lemme précédent. Choisissons $\varepsilon > 0$ tel que $\beta = \alpha + 2\varepsilon < 0$. On écrit un développement limité d'ordre 1

$$y' = \mathrm{F}(y) = (d\mathrm{F})_0(y) + r(y),$$

où le reste $r(y)$ est $o(y)$. On aura en coordonnées dans cette base $\mathbf{f}$, pour y assez voisin de 0, en désignant par Y le vecteur colonne des coordonnées de y dans la base, par $\Phi(\mathrm{Y})$ le vecteur colonne des coordonnées de $\mathrm{F}(y)$ et par B la matrice de la différentielle $(d\mathrm{F})_0$ dans la base $\mathbf{f}$

$$\mathrm{Y}' = \Phi(\mathrm{Y}) = \mathrm{BY} + \mathrm{R}(\mathrm{Y}),$$

où le “reste” $\mathrm{R}(\mathrm{Y})$ est $o(\mathrm{Y})$; on choisit alors $\eta > 0$ de façon que $\|\mathrm{R}(\mathrm{Y})\| \le \varepsilon\|\mathrm{Y}\|$ chaque fois que $\|\mathrm{Y}\| < \eta$, et on désigne par V le voisinage de 0 formé de tous les y tels que $\|\mathrm{Y}\| < \eta$. On aura alors l'inégalité $\mathrm{Y}\,.\,\mathrm{R}(\mathrm{Y}) \le \|\mathrm{Y}\|\,\|\mathrm{R}(\mathrm{Y})\|$ qui implique pour tout $y \in \mathrm{V}$

$$\mathrm{Y}\,.\,\mathrm{Y}' = \mathrm{Y}\,.\,\mathrm{BY} + \mathrm{Y}\,.\,\mathrm{R}(\mathrm{Y}) \le (\alpha + 2\varepsilon)\,\mathrm{Y}\,.\,\mathrm{Y} \le 0.$$

Mais cette équation montre que $\mathrm{Y}\,.\,\mathrm{Y} = \|\mathrm{Y}\|^2$ est alors décroissant ; si la valeur initiale $y(0) = y_0$ est dans V, c'est à dire que $\|\mathrm{Y}_0\| < \eta$, alors $y(t)$ restera dans V pour toutes les valeurs $t > 0$, et de plus on aura $\|\mathrm{Y}(t)\|^2 \le \mathrm{e}^{2\beta t}$ qui tend vers 0 lorsque $t \to +\infty$.

Proposition 5.3.2. *On suppose* $(x, y) \to \mathrm{F}(x, y)$ *continue et* M-*lipschitzienne en* y *sur* $\mathrm{I} \times \mathbb{R}^d$, *où* I *est un intervalle ouvert de* $\mathbb{R}$ *(borné ou non, qui peut être égal à* $\mathbb{R}$ *tout entier dans certains cas). Pour tout* $(t_0, y^{(0)}) \in \mathrm{I} \times \mathbb{R}^d$, *il existe une solution* φ *définie sur* I *telle que* $\varphi(t_0) = y^{(0)}$.

On remarquera que dans cette proposition, on a pu garantir l'existence d'une solution définie sur un intervalle I donné à l'avance. C'est un événement assez rare pour qu'il vaille la peine d'être mentionné, et qui provient du caractère *globalement lipschitzien* de la fonction F sur l'ensemble $\Gamma = \mathrm{I} \times \mathbb{R}^d$.

Démonstration. On va utiliser la méthode des approximations successives qui consiste à introduire une suite (φ_n) qui converge vers la solution cherchée. On supposera pour simplifier que $d = 1$ et on prendra $t_0 = 0 \in \mathrm{I}$. Il existe par hypothèse une constante M telle que

$$|\mathrm{F}(x, y_1) - \mathrm{F}(x, y_2)| \le \mathrm{M}\,|y_1 - y_2|$$

pour tous $y_1, y_2 \in \mathbb{R}$ et tout $x \in \mathrm{I}$. On pose $\varphi_0(x) = y_0$ et pour $n \ge 0$,

$$\varphi_{n+1}(x) = y_0 + \int_0^x \mathrm{F}(t, \varphi_n(t))\,dt.$$

Soient S, T tels que $\mathrm{S} < 0$, $\mathrm{T} > 0$ et $[\mathrm{S}, \mathrm{T}] \subset \mathrm{I}$. On voit que

$$\varphi_1(x) - \varphi_0(x) = \int_0^x \mathrm{F}(t, y_0)\,dt$$

qui peut être majoré par une constante C quand $x \in [\mathrm{S}, \mathrm{T}]$ (parce que $t \to \mathrm{F}(t, y_0)$ est continue, donc bornée sur $[\mathrm{S}, \mathrm{T}]$. La constante $\mathrm{C} = \mathrm{C}(\mathrm{S}, \mathrm{T})$ dépend de S et T). On va montrer par récurrence que pour tout $x \in [\mathrm{S}, \mathrm{T}]$, on a

$$|\varphi_n(x) - \varphi_{n-1}(x)| \le \mathrm{C}\,\frac{\mathrm{M}^{n-1}|x|^{n-1}}{(n-1)!}.$$

C'est vrai pour $n = 1$. Supposons cette propriété vraie pour l'entier $n \ge 1$, et passons à $n + 1$. On a pour tout $x \ge 0$

$$|\varphi_{n+1}(x) - \varphi_n(x)| \le \int_0^x |(\mathrm{F}(t, \varphi_n(t)) - \mathrm{F}(t, \varphi_{n-1}(t))\,dt\Big| \le$$

$$\le \mathrm{M} \int_0^x |\varphi_n(t) - \varphi_{n-1}(t)|\,dt \le \mathrm{CM} \int_0^x \frac{\mathrm{M}^{n-1} t^{n-1}}{(n-1)!}\,dt = \mathrm{C}\,\frac{\mathrm{M}^n x^n}{n!}.$$

Pour $x < 0$, le calcul est analogue. On a donc montré que, sur tout intervalle $[\mathrm{S}, \mathrm{T}] \subset \mathrm{I}$, la fonction $|\varphi_{n+1} - \varphi_n|$ est majorée par le terme général d'une série de fonctions normalement convergente (à savoir la série exponentielle $\mathrm{C} \sum (\mathrm{M}x)^n/(n!)$). On peut donc poser

$$\varphi(x) = y_0 + \sum_{n=0}^{+\infty} (\varphi_{n+1}(x) - \varphi_n(x)) = \lim_n \varphi_n(x).$$

La série des fonctions continues $x \to (\mathrm{F}(x, \varphi_{n+1}(x)) - \mathrm{F}(x, \varphi_n(x)))$ est elle aussi majorée par la série exponentielle, puisque $|\mathrm{F}(x, \varphi_{n+1}(x)) - \mathrm{F}(x, \varphi_n(x))| \le \mathrm{M}|\varphi_{n+1}(x) - \varphi_n(x)|$, donc cette série converge normalement sur tout intervalle $[\mathrm{S}, \mathrm{T}] \subset \mathrm{I}$. Or c'est la série dérivée de la série précédente, donc

$$\varphi'(x) = \mathrm{F}(x, y_0) + \sum_{n=1}^{+\infty} (\mathrm{F}(x, \varphi_{n+1}(x)) - \mathrm{F}(x, \varphi_n(x))) = \lim_n \mathrm{F}(x, \varphi_n(x)) = \mathrm{F}(x, \varphi(x))$$

pour tout $x \in \mathrm{I}$.

Remarque. Pour traiter le cas vectoriel, il suffit de remplacer les intégrales scalaires par des intégrales à valeurs vectorielles, et les majorations de valeur absolue par des majorations de norme.

Comment reconnaître les fonctions lipschitziennes?

1– Si $y \to h(y)$ est une fonction réelle dérivable sur un intervalle ouvert J, et si on a $|h'(y)| \leq \mathrm{M}$ pour tout $y \in \mathrm{J}$, la fonction h est M-lipschitzienne sur l'intervalle J : ceci résulte immédiatement de la formule des accroissements finis.

Par exemple, $y \to \sin(y)$ est 1-lipschitzienne sur $\mathbb{R}$.

2– Le graphe d'une fonction lipschitzienne peut faire des “coins” : par exemple, $y \to |y|$ est lipschitzienne. Exercice : montrer que le max de deux fonctions M-lipschitziennes est encore M-lipschitzien.

Un exemple : l'équation du pendule

Il s'agit de l'équation différentielle du second ordre

$$y'' = -\sin(y).$$

On considère que $y(t)$ représente l'angle à l'instant t entre la verticale et la tige (supposée rigide) d'un pendule. La valeur $y = 0$ correspond à la position d'équilibre basse, et $y = \pm\pi$ à l'équilibre instable en haut. En passant en dimension 2, on obtient une équation

$$\mathrm{Y}' = \mathrm{F}(\mathrm{Y})$$

avec

$$\mathrm{F}(x, y_1, y_2) = \begin{pmatrix} y_2 \\ -\sin(y_1) \end{pmatrix}.$$

Les fonctions coordonnées de F sont clairement lipschitziennes sur $\mathbb{R}$ tout entier, avec constante 1, donc il existe des solutions définies sur $\mathbb{R}$ tout entier pour toute donnée initiale (y_0, y_0') (position et vitesse) à l'instant $x = 0$. Pour fixer les idées on va s'intéresser au cas $y_0 = 0$, $y_0' \geq 0$. On observe la solution triviale $y = 0$ qui correspond à $y_0' = 0$.

On remarque que $y'y'' + \sin(y)y' = 0$ est une dérivée (c'est le fond du théorème de l'énergie cinétique). Il en résulte que

$$\frac{1}{2}y'^2 - \cos(y) = \mathrm{C} = \frac{1}{2}y_0'^2 - 1.$$

Cette remarque permet de ramener à une équation du premier ordre pas très sympathique. Tant que $y' > 0$, on a

$$y' = \sqrt{2\,(\cos(y) + \mathrm{C})}.$$

Si $y_0' > 0$ la fonction est d'abord croissante. On discute suivant plusieurs possibilités. Si $\mathrm{C} > 1$, la dérivée ne peut jamais s'annuler, elle reste donc toujours > 0 et en fait minorée par $\sqrt{2(\mathrm{C} - 1)} > 0$: le pendule fait des tours complets autour de l'axe. Si $\mathrm{C} < 1$, on a les mouvements d'oscillation usuels.

La solution “extraordinaire” : quand $\mathrm{C} = 1$ on obtient l'équation $y' = 2\cos(y/2)$ qui peut s'intégrer assez facilement,

$$\ln\tan(y/4 + \pi/4) = x, \quad y(x) = 4\,\mathrm{Arctan}(e^x) - \pi.$$

Théorème 5.3.3. (Cauchy-Lipschitz : existence locale) *On suppose* $\mathrm{F}(x, y)$ *continue et lipschitzienne en* y *dans un voisinage de* $(t_0, y^{(0)})$*. Il existe une solution* (I, φ)*, où* I *contient* t_0*, telle que* $\varphi(t_0) = y^{(0)}$*.*

Démonstration. Traitons le cas $d = 1$ pour simplifier. Puisque F est lipschitzienne en y dans un voisinage de (x_0, y_0), on peut trouver deux intervalles $]a, b[$ et $]c, d[$, centrés respectivement en x_0 et y_0, tels que F soit M-lipschitzienne en y sur $]a, b[\times]c, d[$. Posons pour tout $x \in]a, b[$ et $y \in \mathbb{R}$

$$\mathrm{G}(x, y) = \begin{cases} \mathrm{F}(x, c) \text{ si } y \leq c \\ \mathrm{F}(x, y) \text{ si } c < y < d \\ \mathrm{F}(x, d) \text{ si } y \geq d. \end{cases}$$

Il est facile de vérifier que pour chaque x fixé dans $]a, b[$, ce prolongement très simple est encore M-lipschitzien, c'est à dire que G est M-lipschitzienne en y sur $]a, b[\times\mathbb{R}$, ce qui permet d'appliquer le théorème d'unicité "semi-global" : d'après la proposition 5.3.2, il existe une solution φ définie sur $]a, b[$ et telle que

$$\varphi'(t) = \mathrm{G}(t, \varphi(t)),$$

$\varphi(t_0) = y^{(0)}$. La fonction φ étant continue, il existe un intervalle $\mathrm{I} \subset]a, b[$ contenant t_0 tel que $\varphi(t)) \in]c, d[$ pour tout $t \in \mathrm{I}$. Mais alors $\mathrm{G}(t, \varphi(t)) = \mathrm{F}(t, \varphi(t))$ pour tout $t \in \mathrm{I}$, donc (I, φ) est solution de $y' = \mathrm{F}(x, y)$.

Résumé. (Cauchy-Lipschitz : existence et unicité locale) *On suppose* $\mathrm{F}(x, y)$ *continue et lipschitzienne en* y *dans un voisinage de* $(t_0, y^{(0)})$*. Il existe un intervalle* I *contenant* t_0 *et une solution* (I, φ)*, telle que* $\varphi(t_0) = y^{(0)}$ *et telle que pour toute solution* (J, ψ) *vérifiant* $\psi(t_0) = y^{(0)}$*, on ait* $\psi(t) = \varphi(t)$ *pour tout* $t \in \mathrm{J} \cap \mathrm{I}$*.*

Exercice. Trouver toutes les solutions de l'équation $y' = y^2$.

Réduction théorique au premier ordre (homogène)

Supposons d'abord que (I, φ) soit solution de l'équation différentielle $y' = \mathrm{F}(x, y)$. Définissons $\psi : \mathrm{I} \to \mathbb{R}^{d+1}$ par $\psi(t) = (t, \varphi(t)) \in \mathbb{R}^{d+1}$. On voit que $\psi'(t) = \mathrm{G}(\psi(t))$, où on a défini la fonction G sur $\mathbb{R}^{d+1}$ par $\mathrm{G}(t, y) = (1, \mathrm{F}(t, y)) \in \mathbb{R}^{d+1}$. Réciproquement, si ψ est solution de $z' = \mathrm{G}(z)$ avec condition initiale $\psi(t_0) = (t_0, y^{(0)})$, on vérifie que $\psi(t)$ est de la forme $\psi(t) = (t, \varphi(t))$ et que φ vérifie $\varphi'(t) = \mathrm{F}(t, \varphi(t))$ pour tout $t \in \mathrm{I}$. Cette transformation montre donc que toute équation différentielle de la forme $y' = \mathrm{F}(x, y)$ peut se transformer en équation autonome $z' = \mathrm{G}(z)$, mais au prix du remplacement de $\mathbb{R}^d$ par $\mathbb{R}^{d+1}$ comme espace des valeurs.

Si on considère une équation différentielle d'ordre plus élevé, par exemple $y'' = \mathrm{F}(x, y, y')$, on posera $y_1(t) = \varphi(t) \in \mathbb{R}^d$, $y_2(t) = \varphi'(t) \in \mathbb{R}^d$, et on définira $\mathrm{G}(t, y_1, y_2)$ sur $\mathbb{R} \times \mathbb{R}^{2d}$, à valeurs dans $\mathbb{R}^{2d}$, par $\mathrm{G}(t, y_1, y_2) = (y_2, \mathrm{F}(t, y_1, y_2))$. Cette nouvelle transformation montre qu'on peut ramener ces équations d'ordre plus élevé à des équations d'ordre 1. On en déduit les résultats d'existence et d'unicité correspondants.

Equations différentielles linéaires à coefficients variables

Ce sont les équations de la forme

$$y'x) = \mathrm{A}(x)y(x) + g(x),$$

où $x \to \mathrm{A}(x)$ est une application continue d'un intervalle ouvert $\mathrm{I} \subset \mathbb{R}$ à valeurs matrices $d \times d$. Pour tout intervalle fermé borné $[\mathrm{S}, \mathrm{T}] \subset \mathrm{I}$, la fonction continue $x \to \|\mathrm{A}(x)\|$

est bornée sur $[\mathrm{S}, \mathrm{T}]$, disons par M, donc $\mathrm{F}(x, y) = \mathrm{A}(x)y$ est M-lipschitzienne en y sur $[\mathrm{S}, \mathrm{T}] \times \mathbb{R}^d$, ce qui garantit l'existence de solutions définies sur I tout entier d'après la proposition 5.3.2, pour toute donnée initiale $(t_0, y^{(0)}) \in \mathrm{I} \times \mathbb{R}^d$; expliquons ce dernier point ; choisissons une suite croissante d'intervalles $[\mathrm{S}_n, \mathrm{T}_n]$ contenant t_0 et qui recouvre I. Pour chaque n, il existe une solution φ_n définie sur $[\mathrm{S}_n, \mathrm{T}_n]$, et d'après le résultat d'unicité, on voit que la restriction de φ_{n+1} à $[\mathrm{S}_n, \mathrm{T}_n]$, qui est solution avec la même condition initiale, doit coïncider avec φ_n. On peut donc "recoller" ces différentes solutions pour définir une fonction φ sur I, qui sera solution.

ATTENTION. Soit $x \to \mathrm{B}(x)$ la primitive de la fonction matricielle $x \to \mathrm{A}(x)$, c'est à dire que B est la fonction matricielle dont les coefficients sont les primitives des coefficients respectifs de A. La dérivée de $x \to \mathrm{e}^{\mathrm{B}(x)}$ n'est pas en général égale à $\mathrm{A}(x)\,\mathrm{e}^{\mathrm{B}(x)}$, et on ne peut pas se contenter de recopier les formules du cas à coefficients constants.

Cependant, si $\mathrm{A}(x) = g(x)\mathrm{A} + h(x)\mathrm{I}_d$, où g et h sont deux fonctions réelles, toutes les matrices $\mathrm{A}(x)$ commutent et on n'a plus la difficulté précédente.

Résolution de l'équation avec second membre

Soit I un intervalle sur lequel est définie l'application $x \to \mathrm{A}(x)$, et soit $t_0 \in \mathrm{I}$; pour $j = 1, \ldots, d$, soit φ_j la solution de $y'(x) = \mathrm{A}(x)y(x)$ telle que $\varphi_j(t_0) = \mathbf{e}_j$ (le $j$ième vecteur de la base canonique de $\mathbb{R}^d$), et soit $\mathrm{C}(t)$ la matrice dont les colonnes sont $\varphi_1(t), \ldots, \varphi_d(t)$; on a alors $\mathrm{C}(t_0) = \mathrm{I}_d$, et on vérifie que $\mathrm{C}'(t) = \mathrm{A}(t)\mathrm{C}(t)$. La solution telle que $\varphi(t_0) = y^{(0)}$ est alors donnée par $\varphi(t) = \mathrm{C}(t)y^{(0)}$.

On vérifie que $\mathrm{C}(t)$ est inversible pour tout $t \in \mathrm{I}$. Sinon, il existerait $z \in \mathbb{R}^d$ tel que $\mathrm{C}(t)z = 0$, avec $z = (z_1, \ldots, z_d)$ non nul. La fonction $s \to \mathrm{C}(s)z$ est solution, et elle est nulle au point t. D'après l'unicité, elle doit être nulle sur I, donc $0 = \mathrm{C}(t_0)z = z$, contradiction.

Supposons que φ soit solution de l'équation avec second membre $y'(x) - \mathrm{A}(x)y(x) = g(x)$. Etudions $\psi(t) = \mathrm{C}(t)^{-1}\varphi(t)$. D'après le chapitre de calcul différentiel, on sait que la dérivée de $t \to \mathrm{C}(t)^{-1}$ au point t est égale à $-\mathrm{C}(t)^{-1}\mathrm{C}'(t)\mathrm{C}(t)^{-1}$, donc

$$\psi'(t) = -\mathrm{C}(t)^{-1}\mathrm{C}'(t)\mathrm{C}(t)^{-1}\varphi(t) + \mathrm{C}(t)^{-1}\varphi'(t) =$$

$$= -\mathrm{C}(t)^{-1}\mathrm{A}(t)\mathrm{C}(t)\mathrm{C}(t)^{-1}\varphi(t) + \mathrm{C}(t)^{-1}(\mathrm{A}(t)\varphi(t) + g(t)) = \mathrm{C}(t)^{-1}g(t),$$

ce qui permet théoriquement de calculer ψ, puis φ.

5.4. Solutions maximales

Dans tout le paragraphe on suppose que Γ est ouvert dans $\mathbb{R} \times \mathbb{R}^d$, on suppose la fonction $\mathrm{F} : (x, y) \to \mathrm{F}(x, y)$ continue sur Γ, et localement lipschitzienne en y sur Γ. Pour tout point (x_0, y_0) de Γ on peut trouver un petit intervalle ouvert I contenant x_0 et une solution φ définie sur I et telle que $\varphi(x_0) = y_0$. Jusqu'où peut-on "étendre" cette solution ? Cette question est beaucoup trop vague pour avoir une réponse ; on va plutôt essayer de préciser la question en définissant la notion de *solution maximale*, puis de donner un critère pour reconnaître les solutions maximales.

On dira que (I, φ) est une solution maximale de l'équation différentielle $y' = \mathrm{F}(x, y)$ si pour toute solution (J, ψ) telle que $\mathrm{I} \subset \mathrm{J}$ et $\psi(t) = \varphi(t)$ pour pour tout $t \in \mathrm{I}$, on a nécessairement $\mathrm{J} = \mathrm{I}$.

Lemme. *Soit* (I, φ) *une solution de l'équation* $y' = \mathrm{F}(x, y)$ *et soient* $t_0 \in \mathrm{I}$, $r > 0$ *; on suppose que* $\varphi(t_0) = y^{(0)}$ *et que* $\|\mathrm{F}(x, y)\| \le \mathrm{C}$ *pour tout* $(x, y) \in \mathrm{I} \times \mathrm{B}(y^{(0)}, r)$. *Alors*

– *pour tout t tel que $|t - t_0| < r/\mathrm{C}$, on a $\|\varphi'(t)\| \le \mathrm{C}$ et $\|\varphi(t) - y^{(0)}\| < r$*
– *si s_1, s_2 vérifient $|s_1 - t_0| < r/\mathrm{C}$ et $|s_2 - t_0| < r/\mathrm{C}$, on a*

$$\|\varphi(s_1) - \varphi(s_2)\| < \mathrm{C}\,|s_1 - s_2|.$$

Démonstration. S'il existe $t > t_0$ dans I tel que $\|\varphi(t) - y^{(0)}\| \ge r$, désignons par t_1 le plus petit nombre réel $> t_0$ tel que $\|\varphi(t_1) - y^{(0)}\| \ge r$. Par définition de t_1, on aura alors $\|\varphi(s) - y^{(0)}\| < r$ pour tout $s \in]t_0, t_1[$, donc $(s, \varphi(s)) \in \mathrm{I} \times \mathrm{B}(y^{(0)}, r)$, donc

$$\|\varphi'(s)\| = \|\mathrm{F}(s, \varphi(s))\| \le \mathrm{C}.$$

Par le théorème des accroissement finis, on en déduit que

$$\|\varphi(t_1) - y^{(0)}\| \le \mathrm{C}\,(t_1 - t_0),$$

et comme $\|\varphi(t_1) - y^{(0)}\| = r$ par continuité, il en résulte que $t_1 - t_0 \ge r/\mathrm{C}$; on procède de même à gauche de t_0, ce qui montre la première propriété de l'énoncé. La seconde propriété est conséquence immédiate de la première et de la majoration des accroissements finis (vectoriels) : pour tout t entre s_1 et s_2, on aura en effet $\|\varphi'(t)\| \le \mathrm{C}$.

Résumons en termes simples ce que dit le lemme précédent : si nous savons que notre vitesse reste $\le \mathrm{C}$ tant que notre distance à y_0 est $\le r$, nous ne pourrons pas sortir de la boule $\mathrm{B}(y_0, r)$ en un temps $\le r/\mathrm{C}$.

Lemme. *Si $(]a, b[, \varphi)$ est solution de l'équation différentielle $y' = \mathrm{F}(x, y)$ et s'il existe une suite (s_k) de points de $]a, b[$ tendant vers b telle que $(s_k, \varphi(s_k))$ tende vers un point $(b, z) \in \Gamma$, alors*

$$\lim_{t \to b, t < b} \varphi(t) = z.$$

Démonstration. On peut supposer que la suite (s_k) est croissante (en passant à une sous-suite). Puisque F est continue au point (b, z), il existe un ensemble de la forme $\mathrm{I} \times \mathrm{B} \subset \Gamma$ sur lequel F est bornée, disons par C, avec $\mathrm{B} = \mathrm{B}(z, \rho)$ pour un certain $\rho > 0$. Soit ε tel que $0 < \varepsilon < \rho$; pour $k \ge k_0$ assez grand, on aura $\varphi(s_k) \in \mathrm{B}(z, \varepsilon/2)$ et $b - s_k < \varepsilon/(2\mathrm{C})$. On applique le lemme précédent avec $\mathrm{I} =]s_{k_0 - 1}, b[$, $t_0 = s_{k_0}$, $y^{(0)} = \varphi(s_{k_0})$ et $r = \varepsilon/2$. On a $\|y^{(0)} - z\| < \varepsilon/2$. Pour tout $(t, y) \in \mathrm{I} \times \mathrm{B}(y^{(0)}, \varepsilon/2)$, on aura $(t, y) \in \mathrm{I} \times \mathrm{B}(z, \varepsilon)$, donc $\|\mathrm{F}(t, y)\| \le \mathrm{C}$. Pour tout $t \in [t_0, b[$ on aura $t - t_0 < \varepsilon/(2\mathrm{C}) = r/\mathrm{C}$, donc d'après le lemme précédent $\varphi(t)$ ne pourra pas sortir de la boule $\mathrm{B}(y^{(0)}, \varepsilon/2)$ pour $t \in [t_0, b[$, donc $\|\varphi(t) - z\| < \varepsilon$. On a donc démontré la convergence de $\varphi(t)$ vers z lorsque $t \to b$.

Théorème 5.4.1. *Si $(]a, b[, \varphi)$ est solution, et s'il existe une suite (t_k) de points de $]a, b[$ tendant vers b telle que $(t_k, \varphi(t_k))$ tende vers un point $(b, y) \in \Gamma$, la solution est prolongeable à un intervalle $]a, c[$, avec $c > b$.*

Dans le cas où Γ est de la forme $\mathbb{R} \times \mathbb{R}^d$, l'énoncé précédent signifie qu'une solution φ définie sur $]a, b[$ ne peut être maximale “du côté de b” que si $\|\varphi(t)\|$ tend vers $+\infty$ lorsque $t \to b$. En effet, on peut toujours dans le cas contraire trouver une suite (s_k) tendant vers b telle que $\varphi(s_k)$ converge vers un certain $z \in \mathbb{R}^d$ (utiliser Bolzano-Weierstrass).

Démonstration. D'après le théorème d'existence local il existe un petit intervalle $\mathrm{J} =]b - \varepsilon, b + \varepsilon[$ et une solution (J, ψ) de l'équation telle que $\psi(b) = y$. On peut supposer que J est assez petit pour qu'il existe un voisinage W de y tel que F soit M-lipschitzienne en

y sur $\mathrm{J}\times\mathrm{W}$ et $\psi(t)\in\mathrm{W}$ pour tout $t\in\mathrm{J}$. D'après le lemme on sait que pour K assez grand, on aura $\varphi(t)\in\mathrm{W}$ pour tout $t\in]t_{\mathrm{K}},b[$. Les deux trajectoires φ et ψ restent dans W pour s entre t et t_k, pour tout $k\geq\mathrm{K}$, donc

$$\|\varphi(t)-\psi(t)\|\leq \mathrm{e}^{\mathrm{M}|t-t_k|}\,\|\varphi(t_k)-\psi(t_k)\|,$$

ce qui donne $\varphi(t)=\psi(t)$ en passant à la limite quand $k\to+\infty$. On voit alors que ψ prolonge φ.

Exercice. Trouver toutes les solutions maximales de l'équation $y'=xy^2$.

Calcul des dérivées successives

Supposons que (I,φ) soit une solution de l'équation différentielle $y'=\mathrm{F}(x,y)$, et supposons $d=1$ pour commencer. Si F est de classe C^1, on voit que $\varphi'(x)=\mathrm{F}(x,\varphi(x))$ est dérivable et on peut calculer (en posant $\mathrm{M}_x=(x,\varphi(x))$)

$$\varphi''(x)=\frac{\partial\mathrm{F}}{\partial x}(\mathrm{M}_x)+\frac{\partial\mathrm{F}}{\partial y}(\mathrm{M}_x)\varphi'(x).$$

Si F est de classe C^2 et si on est courageux, on calcule φ''' selon le même principe.

Dans le cas $d>1$, on peut faire des calculs avec des dérivées partielles, qui deviennent rapidement épouvantanbles. En utilisant la notion de “différentielle partielle” par rapport à la variable $y\in\mathbb{R}^d$, on obtient en revanche des calculs formellement assez analogues au cas de la dimension $d=1$. En particulier

$$\varphi''(x)=\frac{\partial\mathrm{F}}{\partial x}(\mathrm{M}_x)+\frac{\partial\mathrm{F}}{\partial y}(\mathrm{M}_x)(\varphi'(x)).$$

5.5. Méthodes numériques. Runge-Kutta

On suppose que F est suffisamment dérivable (disons de classe C^4), et on rappelle que toute solution φ de l'équation différentielle $y'=\mathrm{F}(x,y)$ admet autant de dérivées que F, plus une.

Commençons par une méthode très rudimentaire. On veut décrire une approximation de l'évolution de la solution sur un petit intervalle $[t_0,t_0+\tau]$ (t_0 est une valeur fixée) de longueur $\tau>0$. Si $\varphi(t_0)=y_0$, on sait que $\varphi'(t_0)=\mathrm{F}(\varphi(t_0))=\mathrm{F}(y_0)$ donc en première approximation

$$\varphi(t_0+\tau)=\varphi(t_0)+\varphi'(t_0)\tau+\mathrm{O}(\tau^2)$$

ce qui conduit à poser comme première approximation de $\varphi(t_0+\tau)$

$$\widetilde{y}_0(\tau)=y_0+\tau\mathrm{F}(y_0).$$

On peut en déduire un algorithme (aux performances très médiocres) pour calculer des approximations de φ aux points successifs $t_k=t_0+k\tau$, pour $k=1,\ldots,\mathrm{N}$. On donne y_0 et ensuite pour tout $k\geq 0$ on pose

$$y_{k+1}=y_k+\tau\mathrm{F}(y_k)$$

(avec bien sûr l'idée que y_k doit approximer $\varphi(t_k)$).

On peut donner tout de suite une petite amélioration. Si $\widetilde{y}_1=y_0+\tau\mathrm{F}(y_0)$ est une approximation de $\varphi(t_0+\tau)$, $\mathrm{F}(\widetilde{y}_1)$ donne une approximation de $\varphi'(t_0+\tau)$. Si on approxime

$$\varphi(t_0+\tau)-\varphi(t_0)=\int_{t_0}^{t_0+\tau}\varphi'(u)\,du$$

par la formule des trapèzes, on trouve

$$\begin{aligned}\varphi(t_0+\tau)-\varphi(t_0) &= \frac{\tau}{2}\left(\varphi'(t_0)+\varphi'(t_0+\tau)\right)+\mathrm{O}(\tau^3)\\ &= \frac{\tau}{2}(\mathrm{F}(y_0)+\mathrm{F}(y_0+\tau\mathrm{F}(y_0)))+\mathrm{O}(\tau^3)\end{aligned}$$

d'où la nouvelle formule

$$y_{k+1}=y_k+\frac{\tau}{2}(\mathrm{F}(y_k)+\mathrm{F}(y_k+\tau\mathrm{F}(y_k))).$$

Ce procédé est un peu meilleur, mais il n'est pas utilisé dans la pratique car on peut faire beaucoup mieux pour pas beaucoup plus cher. C'est la méthode de Runge-Kutta.

On calcule pour passer de t_0 à $t_0+\tau$ les quantités auxiliaires suivantes

$$\begin{aligned}k_0 &= \mathrm{F}(t_0,y_0)\\ k_1(\tau) &= \mathrm{F}(t_0+\frac{\tau}{2},y_0+\frac{\tau}{2}\,k_0)\\ k_2(\tau) &= \mathrm{F}(t_0+\frac{\tau}{2},y_0+\frac{\tau}{2}\,k_1(\tau))\\ k_3(\tau) &= \mathrm{F}(t_0+\tau,y_0+\tau\,k_2(\tau))\end{aligned}$$

et on pose finalement

$$\widetilde{y}_0(\tau)=y_0+\frac{\tau}{6}\big(k_0+2k_1(\tau)+2k_2(\tau)+k_3(\tau)\big).$$

La méthode est d'ordre 4 (c'est à dire que l'erreur sur chaque petit pas est $\mathrm{O}(\tau^5)$), et simple à programmer.

La méthode de Runge-Kutta marche dans le cas vectoriel. Le calcul des dérivées de φ dans ce cas est un excellent exercice de calcul différentiel. Il en résulte qu'il suffit de vérifier la validité de Runge-Kutta dans le cas autonome vectoriel, pour les équations du premier ordre. La validité de la méthode vectorielle permet de déduire le cas non autonome, ou bien le cas des équations différentielles d'ordres plus élevés par réduction au cas autonome du premier ordre, quitte à augmenter la dimension.

Pour la vérification, il faut voir que l'expression de $\widetilde{y}_0(\tau)$ ci-dessus admet, en tant que fonction de τ, le même DL à l'ordre 4 au point $\tau=0$, dans la variable τ, que la vraie solution φ telle que $\varphi(t_0)=y_0$. C'est un exercice très pénible, surtout dans le cas vectoriel. Ecrivons le DL à l'ordre 3 de F au voisinage de $y_0\in\mathbb{R}^d$. Supposons que

$$\mathrm{F}(y_0+h)=z_0+z_1(h)+\frac{1}{2}z_2(h,h)+\frac{1}{6}z_3(h,h,h)+\mathrm{O}(\|h\|^4)$$

où z_1 est linéaire sur $\mathbb{R}^d$, z_2 bilinéaire symétrique sur $(\mathbb{R}^d)^2$ et z_3 trilinéaire symétrique sur $(\mathbb{R}^d)^3$. On écrira des DL à l'ordre 3 en τ

$$\begin{aligned}k_0 &= z_0,\\ k_1(\tau) &= z_0+\frac{\tau}{2}z_1(z_0)+\frac{\tau^2}{8}z_2(z_0,z_0)+\frac{\tau^3}{48}z_3(z_0,z_0,z_0)+\mathrm{O}(\tau^4),\\ k_2(\tau) &= z_0+\frac{\tau}{2}z_1(z_0)+\frac{\tau^2}{8}(2z_1(z_1(z_0))+z_2(z_0,z_0))+\\ &\quad+\frac{\tau^3}{48}(3z_1(z_2(z_0,z_0))+6z_2(z_0,z_1(z_0))+z_3(z_0,z_0,z_0))+\mathrm{O}(\tau^4),\\ k_3(\tau) &= z_0+\tau z_1(z_0)+\frac{\tau^2}{2}(z_1(z_1(z_0))+z_2(z_0,z_0))+\\ &\quad+\frac{\tau^3}{24}(6z_1(z_1(z_1(z_0)))+3z_1(z_2(z_0,z_0))+12z_2(z_0,z_1(z_0))+4z_3(z_0,z_0,z_0))+\mathrm{O}(\tau^4).\end{aligned}$$

On obtient alors

$$\frac{1}{6}\big(k_0+2k_1(\tau)+2k_2(\tau)+k_3(\tau)\big)=z_0+\frac{\tau}{2}z_1(z_0)+\frac{\tau^2}{6}\big(z_1(z_1(z_0))+z_2(z_0,z_0)\big)+$$
$$+\frac{\tau^3}{24}\big(z_3(z_0,z_0,z_0)+z_1(z_2(z_0,z_0))+3z_2(z_0,z_1(z_0))+z_1(z_1(z_1(z_0)))\big)+\mathrm{O}(\tau^4).$$

On peut ensuite calculer le DL à l'ordre 4 de $\varphi(t_0+\tau)$, et on constatera que $\widetilde{y}_0(\tau)-\varphi(t_0+\tau)=\mathrm{O}(\tau^5)$.

5.6. Dérivée par rapport à la condition initiale. Divers

Supposons pour simplifier que $\mathrm{F}(x,y)$ soit M-lipschitzienne en y sur $\mathrm{I}\times\mathbb{R}$, et soit $x_0\in\mathrm{I}$. Pour chaque donnée initiale $h\in\mathbb{R}$, il existe une unique solution définie sur I et telle que $y(x_0)=h$. Désignons par $\psi(x,h)$ la valeur de cette solution au point $x\in\mathrm{I}$. Si on sait que $\psi(x,h)$ est dérivable, on en déduit facilement une équation différentielle satisfaite par $\dfrac{\partial\psi}{\partial h}$, par dérivation sous l'intégrale. Mais ceci ne montre pas que ψ soit dérivable par rapport à h. Donnons tout de même cette équation différentielle. On part de

$$\psi(x,h)=h+\int_{x_0}^{x}\mathrm{F}(t,\psi(t,h))\,dt$$

qui donne par dérivation sous l'intégrale

$$\frac{\partial\psi}{\partial h}(x,h)=1+\int_{x_0}^{x}\frac{\partial\mathrm{F}}{\partial y}(t,\psi(t,h))\frac{\partial\psi}{\partial h}(t,h)\,dt$$

ce qui montre que $x\to\dfrac{\partial\psi}{\partial h}(x,h_0)$ vérifie l'équation différentielle

$$z'(x)=\frac{\partial\mathrm{F}}{\partial y}(x,\psi(x,h_0))\,z(x),$$

avec donnée initiale $z(x_0)=1$. Il s'agit d'une équation linéaire à coefficients variables, du premier ordre.

Inversion locale

Supposons que f soit une application de classe C^2 de $\mathbb{R}^d$ dans $\mathbb{R}^d$, telle que $(df)_x$ soit inversible pour tout $x\in\mathbb{R}^d$, et que $x\to(df_x)^{-1}$ soit lipschitzienne sur $\mathbb{R}^d$, à valeurs dans $\mathcal{L}(\mathbb{R}^d)$. Supposons aussi que $f(0)=0$. Soit z un point de $\mathbb{R}^d$; considérons l'équation différentielle à coefficients lipschitziens

$$y'=(df_y)^{-1}(z).$$

Si φ est solution, on voit que $\psi(t)=f(\varphi(t))$ vérifie l'équation $\psi'(t)=(df)_{\varphi(t)}(\varphi'(t))=z$ et la condition initiale $\psi(0)=0$, donc $z(t)=tz$, ce qui montre que $f(\varphi(1))=z$. On a ainsi prouvé que f est surjective. Compte tenu des autres hypothèses, on peut montrer que f est un difféomorphisme d'un voisinage de 0 dans $\mathbb{R}^d$ sur un autre voisinage de 0 (un *difféomeorphisme* d'un ouvert U sur un ouvert V est une bijection de classe C^1 de U sur V, dont la bijection inverse est également de classe C^1).

On pourrait raffiner un peu pour traiter le cas où $(df)_x$ est inversible uniquement dans un voisinage de 0, et se rapprocher ainsi du vrai énoncé du théorème d'inversion locale :

supposons que f soit une application de classe C^1 d'un ouvert U de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}^d$; soit a un point de U tel que $(df)_a$ soit inversible ; il existe un ouvert $\mathrm{U}_1 \subset \mathrm{U}$ contenant a et un ouvert V_1 contenant $\varphi(a)$ tels que φ soit un difféomorphisme de U_1 sur V_1.

L'équation de Newton

Il s'agit de la loi physique que l'on appelle la *loi de l'attraction universelle*. Cette loi dit que deux corps célestes A et B exercent l'un sur l'autre une force d'attraction, proportionnelle à l'inverse du carré de la distance de A à B, et dirigée suivant la droite qui joint A et B. Dans le problème du mouvement d'une planète autour du soleil, on considérera que le soleil, beaucoup plus lourd, est fixe, et on négligera aussi les interactions des planètes entre elles. Dans notre modèle mathématique simplifié, le soleil sera placé au point 0, et nous chercherons le mouvement d'un point M, représenté par $x(t) \in \mathbb{R}^3$, soumis à l'attraction du soleil. La loi fondamentale de la dynamique indique que l'accélération de M, c'est à dire $x''(t)$, est proportionnelle à la force exercée sur M, c'est à dire ici la force d'attraction du soleil. Comme nous ne sommes pas dans un cours de physique, nous ferons abstraction des constantes physiques (masse de M et du soleil, constante d'attraction universelle), et nous étudierons simplement l'équation

$$x'' = \frac{\overrightarrow{u}}{\|\overrightarrow{\mathrm{OM}}\|^2}$$

où u est le vecteur unitaire dirigé de M vers O. On peut donc écrire $\overrightarrow{u} = -x/\|x\|$, et l'équation devient

$$x'' = -\frac{x}{\|x\|^3}.$$

La loi des aires

Considérons le produit vectoriel $x(t) \wedge x'(t)$. Il est facile de voir que sa dérivée est donnée par

$$x'(t) \wedge x'(t) + x(t) \wedge x''(t) = 0$$

puisque x'' est colinéaire à x. Cela signifie que le vecteur $w = x \wedge x'$ est constant. On supposera $x(0) \wedge x'(0)$ non nul (sinon M ira tout droit s'écraser sur le soleil). Alors x reste constamment orthogonal à w, donc le mouvement s'effectue dans le plan orthogonal à w. On va donc restreindre l'étude à $\mathbb{R}^2$ compte tenu de cette remarque fondamentale.

On va représenter le problème en coordonnées polaires, en prenant pour origine 0 le “soleil”, et en désignant par r la distance de 0 au point M. Le mouvement en fonction du temps t est donné par $x(t) = r(\theta(t))\cos(\theta(t))$ et $y(t) = r(\theta(t))\sin(\theta(t))$. On supprimera la dépendance en t pour alléger l'écriture, et on posera $\varphi(\theta) = 1/r$, considéré comme fonction de θ. Les coordonnées de $\overrightarrow{\mathrm{OM}}$ sont donc $(\cos(\theta(t))/\varphi(\theta(t)), \sin(\theta(t))/\varphi(\theta(t)))$. La loi des aires nous indique que $r^2\theta' = c$ est constant, donc $\theta' = c\varphi^2$. Le vecteur vitesse est égal à

$$\overrightarrow{v} = -c\begin{pmatrix}\cos\theta \\ \sin\theta\end{pmatrix}\varphi' + \begin{pmatrix}-\sin\theta \\ \cos\theta\end{pmatrix}c\varphi,$$

et l'accélération à

$$\overrightarrow{a} = (-c\varphi''\theta' - c\varphi\theta')\begin{pmatrix}\cos\theta \\ \sin\theta\end{pmatrix} + (-c\varphi'\theta' + c\varphi'\theta')\begin{pmatrix}-\sin\theta \\ \cos\theta\end{pmatrix},$$

et on sait que

$$\overrightarrow{a} = (-\varphi^2)\begin{pmatrix}\cos\theta\\ \sin\theta\end{pmatrix},$$

donc $c^2(\varphi'' + \varphi) = 1$. On obtient ainsi

$$\varphi(\theta) = \frac{1}{c^2} + \mathrm{A}\sin\theta + \mathrm{B}\cos\theta$$

qui ne donne pas encore le mouvement, mais l'équation de la trajectoire. Moyennant une rotation on peut écrire cette équation sous la forme $\varphi(\theta) = \tau(1 + b\sin\theta)$. Suivant la valeur de b, on obtient une ellipse (si $|b| < 1$), une parabole (si $|b| = 1$) ou une hyperbole si $|b| > 1$. Continuons le calcul dans le cas elliptique. On ne va pas arriver à trouver θ en fonction de t, mais plutôt t en fonction de θ sur une période, quand θ varie de $-\pi$ à π. On a

$$\frac{d\theta}{dt} = \tau^2(1 + b\sin\theta)^2$$

donc

$$\int \frac{d\theta}{(1 + b\sin\theta)^2} = \tau^2 t + \mathrm{C}te,$$

ce qui conduit à

$$\tau^2 t = \frac{2}{(1-b^2)^{3/2}}\,\mathrm{Arctan}\Big(\frac{b + \tan(\theta/2)}{\sqrt{1-b^2}}\Big) + \frac{b\cos\theta}{(1-b^2)(1+b\sin\theta)}.$$

On en déduit la période T du mouvement,

$$\mathrm{T} = \frac{2\pi}{\tau^2(1-b^2)^{3/2}}.$$

Index

Index des notations