Algèbre bilinéaire et analyse de Fourier

Cours rédigé par Jean-Pierre Demailly

Université Joseph Fourier, Grenoble

Module MAT244, Année universitaire 2012/2013

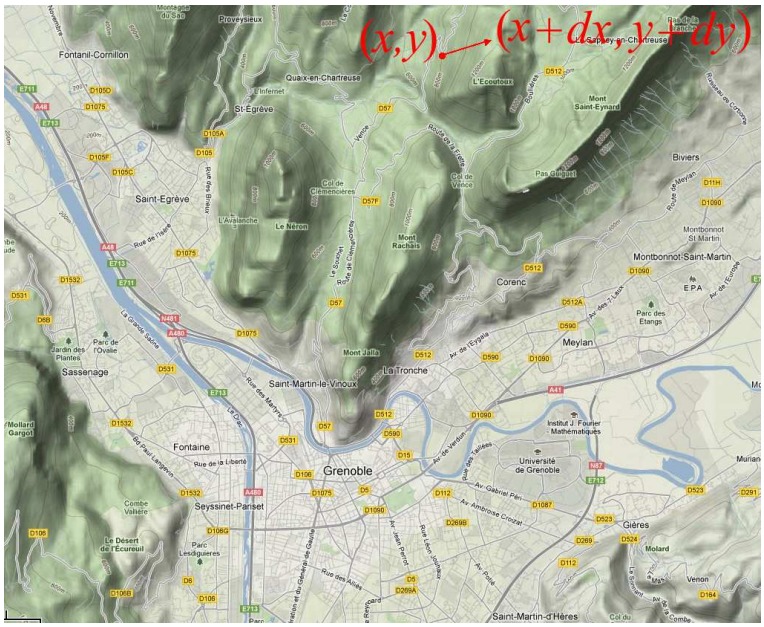
Table des matières

0. Motivations 2
1. Rappels et compléments d'algèbre linéaire 3
2. Formes bilinéaires. 12
3. Orthogonalité par rapport à une forme bilinéaire symétrique 23
4. Formes sesquilinéaires 34
5. Normes et distances, méthode des moindres carrés 44
6. Endomorphismes symétriques, anti-symétriques, orthogonaux et unitaires 49
7. Coniques et quadriques 59
8. Un bref aperçu de la vie de Fourier 69
9. L'équation de la chaleur 70
10. Séries de Fourier, introduction 74
11. Notions de base sur les séries numériques et les séries de fonctions 79
12. Séries de Fourier, théorèmes fondamentaux de convergence 85

Dans tout le cours, $\mathbb{K}$ désignera un corps commutatif tel que $\mathbb{Q}$, $\mathbb{R}$ ou $\mathbb{C}$ (dans toutes les parties purement algébriques cela pourrait être aussi un corps commutatif quelconque dans lequel $2 = 1 + 1 \neq 0$, mais nous n'aurons pas besoin ici de considérer ce cas plus général).

0. Motivations

Supposons que nous soyons amenés à nous promener en montagne, en nous repérant à partir d'une carte d'état-major.



On effectue un petit déplacement depuis le point (x, y) jusqu'au point $(x + dx, y + dy)$ en supposant (dx, dy) suffisamment petit pour que la pente du terrain n'ait pas le temps de changer sensiblement (ce n'est pas nécessairement le cas sur notre dessin, mais la flèche n'aurait pas été visible !) Le problème est de calculer la distance parcourue.

Sur la carte, le déplacement effectué est $\sqrt{dx^2 + dy^2}$ d'après le théorème de Pythagore, mais ceci ne tient absolument pas compte du fait que l'on se déplace sur un terrain en pente. En réalité, l'altitude z varie comme une fonction $z = f(x, y)$ du point (x, y) repéré sur la carte, et la longueur du déplacement effectué est donc

$$ds = \sqrt{dx^2 + dy^2 + dz^2},$$

du moins si on suppose que l'on s'est déplacé en ligne droite (sur une distance suffisamment faible, on peut considérer que c'est le cas). Par différentiation, on a

$$dz = f'_x\, dx + f'_y\, dy$$

où f'_x, f'_y sont les dérivées partielles au point (x, y). On trouve donc

$$ds^2 = dx^2 + dy^2 + (f'_x\, dx + f'_y\, dy)^2 = (1 + f'^2_x)dx^2 + (1 + f'^2_y)dy^2 + 2f'_x f'_y\, dx\, dy.$$

Si $(u, v) = (dx, dy)$ est le vecteur déplacement, on voit que la distance parcourue s'exprime comme $\sqrt{q(u, v)}$ où $q(u, v)$ est une expression de la forme

$$q(u, v) = au^2 + bv^2 + c\, uv.$$

C'est ce qu'on appelle une *forme quadratique* de 2 variables. Plus généralement, on est amené à considérer des espaces de dimension plus grande, par exemple en Physique on introduit *l'espace-temps* de dimension 4, avec ses coordonnées (x, y, z, t) où t est le temps. Dans ce cas, une forme quadratique jouant un rôle important en théorie de la relativité restreinte est la *forme quadratique de Lorentz*

$$q(dx, dy, dz, dt) = dx^2 + dy^2 + dz^2 - c^2\, dt^2$$

où c est la vitesse de la lumière. La théorie de la relativité généralisée consiste en l'étude de l'espace-temps courbé par la présence de la matière ; dans ce cas, de même que dans l'expression de ds^2 pour un parcours vallonné en montagne, on peut avoir des termes dx^2, $dx\,dy$, ... , $dx\,dt$, ... , dt^2 dont les coefficients dépendent eux-mêmes de (x, y, z, t) !

L'un des buts de ce cours est une étude systématique des formes quadratiques et de leurs propriétés. Cette étude est fortement liée à celle des applications et formes linéaires, c'est pourquoi nous commencerons par des rappels généraux d'algèbre linéaire.

1. Rappels et compléments d'algèbre linéaire

Un **$\mathbb{K}$-espace vectoriel** est un ensemble E muni d'une loi de composition interne notée $+$

$$E \times E \to E, \quad (x, y) \mapsto x + y,$$

et d'une loi de composition externe notée $\cdot$ (le $\cdot$ étant d'ailleurs très souvent omis)

$$\mathbb{K} \times E \to E, \quad (\lambda, x) \mapsto \lambda \cdot x,$$

appelée ici *multiplication par un scalaire*, satisfaisant aux propriétés suivantes :

(1) (associativité de $+$) $x + (y + z) = (x + y) + z$ pour tous $x, y \in E$
(2) (commutativité de $+$) $x + y = y + x$ pour tous $x, y \in E$
(3) (élément neutre) il existe un élément 0_E tel que $0_E + x = x + 0_E = x$ pour tout $x \in E$.
(4) pour tout $x \in E$, il existe un élément $x' \in E$ tel que $x + x' = x' + x = 0_E$ (cet élément x' est alors unique, appelé opposé de x, et il est noté $-x$)
(5) $1 \cdot x = x$ pour tout $x \in E$
(6) $(\lambda\mu) \cdot x = \lambda \cdot (\mu \cdot x)$ pour tous $\lambda, \mu \in \mathbb{K}$, $x \in E$
(7) $\lambda \cdot (x + y) = \lambda \cdot x + \lambda \cdot y$ pour tous $x, y \in E$, $\lambda \in \mathbb{K}$
(8) $(\lambda + \mu) \cdot x = \lambda \cdot x + \mu \cdot x$ pour tous $x \in E$, $\lambda, \mu \in \mathbb{K}$.

Exemples 1.1. l'ensemble $\mathbb{K}^n$ des n-uplets $(x_1, \ldots, x_n)$ d'éléments de $\mathbb{K}$; l'ensemble $\mathbb{K}[X]$ des polynômes à coefficients dans $\mathbb{K}$; l'ensemble $\mathrm{M}_n(\mathbb{K})$ des matrices carrées d'ordre n ; l'ensemble $C^0(I, \mathbb{K})$ des fonctions continues $f : I \to \mathbb{K}$; l'ensemble des suites réelles ou complexes.

Un **sous-espace vectoriel** F de E est un sous-ensemble de E **non vide**, stable par addition et multiplication par un scalaire : $\forall x, y \in F$ on a $x+y \in F$, et $\forall \lambda \in \mathbb{K}$, $\forall x \in F$, on a $\lambda \cdot x \in F$. De façon équivalente, c'est un sous-ensemble de E **non vide** et stable par combinaisons linéaires : $\forall \lambda, \mu \in \mathbb{K}$ et $\forall x, y \in F$ on $\lambda \cdot x + \mu \cdot y \in K$. Dans ce cas, F est lui même un $\mathbb{K}$-espace vectoriel pour les lois $+$ et $\cdot$ induites par E.

Remarque 1.2. Un sous-espace vectoriel F de E contient toujours 0_E, car si $x \in F$, alors $0 \cdot x = 0_E \in F$.

Exemple 1.3. Un plan d'équation $ax + by + cz = 0$ $(a, b, c \in \mathbb{R})$ est un sous-espace vectoriel de $\mathbb{R}^3$.

Contre-exemple 1.4. L'ensemble défini par $x - y + 2z = 3$ n'est pas un sous-espace vectoriel de $\mathbb{R}^3$.

Si F_1, F_2, ..., F_N sont des sous-espaces vectoriels de E, alors **l'intersection**

$$\bigcap_{i=1}^{N} F_i = F_1 \cap \ldots \cap F_N$$

est un sous-espace vectoriel de E, mais en général **la réunion**

$$\bigcup_{i=1}^{N} F_i = F_1 \cup \ldots \cup F_N$$

n'en est pas un (considérer par exemple deux droites concourantes dans un plan).

Soit E un $\mathbb{K}$-espace vectoriel. On appelle famille (finie ou infinie) de vecteurs de E une collection $\mathcal{S} = (s_i)_{i\in I}$ de vecteurs de E, non nécessairement distincts, numérotés par des indices i dans un certain ensemble I (lorsque la famille est finie de cardinal m, on choisit en général $I = \{1, 2, \ldots, m\}$). Une **combinaison linéaire** d'éléments de la famille $\mathcal{S}$ est un vecteur de E de la forme

$$\sum_{i\in I} \lambda_i s_i,$$

où $(\lambda_i)_{i\in I}$ est une famille de scalaires n'ayant qu'un nombre fini de coefficients $\lambda_i \neq 0$ (de sorte que la somme se réduit en fait à une somme finie) ; on dit qu'une telle famille de scalaires est *presque nulle* (noter que cette condition est toujours satisfaite si I est fini).

Le **sous-espace vectoriel de E engendré par $\mathcal{S}$** est l'ensemble $\mathrm{Vect}(\mathcal{S})$ des combinaisons linéaires d'éléments de $\mathcal{S}$. Autrement dit,

$$\mathrm{Vect}(\mathcal{S}) = \Big\{ \sum_{i\in I} \lambda_i s_i \,;\ \lambda_i \in \mathbb{K},\ \lambda_i \neq 0 \text{ en nombre fini} \Big\}.$$

On dit qu'une famille $\mathcal{S}$ de vecteurs de E est **génératrice** (ou **engendre** E) si tout vecteur $v \in E$ est une combinaison linéaire d'éléments de $\mathcal{S}$ (autrement dit si $E = \mathrm{Vect}(\mathcal{S})$).

Exemples 1.5.

(1) Un plan de $\mathbb{R}^3$ est engendré par deux vecteurs non colinéaires de ce plan.
(2) Les n vecteurs $(1, 0, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, 1)$ engendrent $\mathbb{K}^n$.
(3) $\mathbb{K}[X]$ est engendré par la famille infinie $(X^n)_{n\in\mathbb{N}}$.
(4) Une famille quelconque $\mathcal{S}$ est toujours par définition une famille génératrice de $\mathrm{Vect}(\mathcal{S})$.

On dit qu'une famille $\mathcal{S} = (s_i)_{i\in I}$ de vecteurs de E est **libre** si pour toute famille de scalaires $(\lambda_i)_{i\in I}$ presque nulle on a

$$\sum_{i\in I} \lambda_i s_i = 0 \Rightarrow \lambda_i = 0 \text{ pour tout } i \in I.$$

Cela aussi revient à dire qu'aucun élément de $\mathcal{S}$ n'est combinaison linéaires des autres.

Exemples 1.6.

(1) Une famille contenant 0 n'est **jamais** libre.
(2) Une famille contenant deux vecteurs identiques n'est **jamais** libre, puisqu'on peut écrire $1 \cdot v + (-1) \cdot v = 0$.
(3) Si $v_1, v_2 \in E$, une famille (v_1, v_2) est libre si et seulement si v_1 et v_2 sont non nuls et non colinéaires.
(4) La famille de fonctions continues (f_1, f_2, f_3) de $\mathbb{R}$ dans $\mathbb{R}$ telle que
$$f_1(x) = 1, \quad f_2(x) = \cos(2x), \quad f_2(x) = \cos^2(x)$$
n'est pas libre (**pourquoi** ?)

(5) La famille $(1, 0, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, 1)$ de vecteurs de $\mathbb{K}^n$ est libre.
(6) La famille (f, g) de fonctions $f(x) = \cos(x)$, $g(x) = \sin(x)$ sur $\mathbb{R}$ est libre.

On dit qu'une famille de vecteurs $\mathcal{B} = (e_i)_{i\in I}$ est une **base** de E si elle est à la fois libre et génératrice. Cela revient à dire que tout vecteur x de E s'écrit de manière **unique** comme combinaison linéaire $x = \sum_{i\in I} x_i e_i$ d'éléments de $\mathcal{B}$ (la somme n'ayant qu'un nombre fini de termes $x_i \neq 0$).

Exemples 1.7.

(1) La famille de vecteurs $(1, 0, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, 1)$ forme une base de $\mathbb{K}^n$, appelée **base canonique**.
(2) La famille $(1, X, \ldots, X^n)$ forme une base de l'espace vectoriel, noté $\mathbb{K}[X]_n$, des polynômes à coefficients dans $\mathbb{K}$ de degré au plus n. On a donc $\dim_{\mathbb{K}} \mathbb{K}[X]_n = n + 1$.
(3) Plus généralement, si $P_j \in \mathbb{K}[X]$ est un polynôme de degré exactement k (c'est-à-dire de coefficient de X^j non nul), alors $(P_0, P_1, \ldots, P_n)$ est une base de $\mathbb{K}[X]_n$. On démontre en effet facilement par récurrence sur n qu'il s'agit d'une famille libre, respectivement d'une famille génératrice.
(4) Si $\mathcal{S}$ est une famille libre de E, alors c'est une base de $\mathrm{Vect}(\mathcal{S})$.

On dit que E est **de dimension finie** s'il existe une famille génératrice de cardinal fini. Dans ce cas, E possède au moins une base, et toutes les bases sont de même cardinal, appelé la **dimension** de E sur $\mathbb{K}$, notée $\dim_{\mathbb{K}} E$; on omettra parfois l'indice $\mathbb{K}$ dans cette notation s'il n'y a pas d'ambiguité. (Nous ne redémontrerons pas ici ces résultats fondamentaux qui font l'objet du cours d'introduction à l'algèbre linéaire).

Si E n'est pas de dimension finie, on peut démontrer aussi que E admet une base (la démonstration utilise des raisonnements plus avancés de théorie des ensembles, en particulier "l'axiome du choix", et elle sera admise).

Théorème 1.8 (autres propriétés fondamentales)**.**

(1) Toute famille libre $\mathcal{S}$ de E **peut se compléter** *en une base de E, et a donc un nombre d'éléments inférieur ou égal à $\dim_{\mathbb{K}} E$.*
(2) De toute famille génératrice $\mathcal{S}$ de E **on peut extraire** *une base de E, et $\mathcal{S}$ doit donc avoir un nombre d'éléments supérieur ou égal à $\dim_{\mathbb{K}} E$.*
(3) Si E est de dimension finie, il en est de même pour tout sous-espace vectoriel F, et on a

$$\dim_{\mathbb{K}} F \leq \dim_{\mathbb{K}} E,$$

avec égalité si et seulement si $F = E$.
(4) Si $n = \dim_{\mathbb{K}} E$ est finie, une famille libre ou génératrice ayant exactement n éléments est nécessairement une base.

On suppose maintenant que E est de dimension finie n. Soit $\mathcal{B} = (e_1, \ldots, e_n)$ une base de E. Alors, pour tout $x \in E$, on peut écrire

$$x = x_1 e_1 + \ldots + x_n e_n, \qquad x_i \in \mathbb{K}.$$

La matrice colonne

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

s'appelle la matrice des **coordonnées** de x dans la base $\mathcal{B}$. Il convient de distinguer soigneusement le vecteur x de la matrice X qui le représente (et qui d'ailleurs dépend de la base $\mathcal{B}$ choisie).

Attention ! Une base est une famille ordonnée ! Si on change l'ordre des vecteurs, on obtient une nouvelle base, et les coordonnées x_i sont permutées.

Soit $\mathcal{B}' = (e'_1, \dots, e'_n)$ une autre base de E. Comment calculer les coordonnés de $x \in E$ dans la base $\mathcal{B}'$ lorsqu'on les connait dans la base $\mathcal{B}$?

Soit $P \in \mathrm{M}_n(\mathbb{K})$ la matrice de passage de $\mathcal{B}$ à $\mathcal{B}'$, c'est-à-dire la matrice carrée $P = (p_{ij})$ dont les colonnes successives

$$\begin{pmatrix} p_{11} \\ \vdots \\ p_{n1} \end{pmatrix}, \ \dots \ , \begin{pmatrix} p_{1n} \\ \vdots \\ p_{nn} \end{pmatrix}$$

sont les matrices de coordonnées des vecteurs $e'_1, \dots, e'_n$ de $\mathcal{B}'$ dans l'ancienne base $\mathcal{B}$. Alors par définition on a $e'_j = \sum_{i=1}^n p_{ij} e_i$. Si

$$X' = \begin{pmatrix} x'_1 \\ \vdots \\ x'_n \end{pmatrix}$$

désignent les coordonnées du vecteur x dans $\mathcal{B}'$, on a

$$x = \sum_{j=1}^n x'_j e'_j = \sum_{j=1}^n x'_j \left(\sum_{i=1}^n p_{ij} e_i \right) = \sum_{i=1}^n \left(\sum_{j=1}^n p_{ij} x'_j \right) e_i,$$

et par conséquent les anciennes coordonnées sont liées aux nouvelles par la relation

$$x_i = \sum_{j=1}^n p_{ij} x'_j \quad \Longleftrightarrow \quad X = PX' \quad \Longleftrightarrow \quad X' = P^{-1}X.$$

On notera que l'unicité des coordonnées entraîne que l'application $X' \mapsto X = PX'$ est bijective, donc la matrice P doit être inversible (sinon, il y a erreur, à moins que la famille $\mathcal{B}'$ considérée ne soit pas une base !)

Pour retenir la formule : se souvenir que si la matrice P exprime la nouvelle base $\mathcal{B}'$ par rapport à l'ancienne, alors la formule $X = PX'$ donne au contraire les anciennes coordonnées par rapport aux nouvelles.

Exemple 1.9. Supposons, dans un espace vectoriel E de dimension 3 muni d'une base (i, j, k), que l'on effectue le changement de coordonnées

$$\begin{cases} x' = 2x - y + z \\ y' = -x + 2y + 4z \\ z' = -4x + y + z \end{cases}$$

où (x, y, z) désignent les coordonnées dans la base (i, j, k). À quelle base correspondent ces nouvelles coordonnées ? Pour le trouver, on résout le système ci-dessus de la forme $X' = AX$, ce qui donne une solution unique (vérification laissée au lecteur)

$$\begin{cases} x = -\frac{1}{9}x' + \frac{1}{9}y' - \frac{1}{3}z' \\ y = -\frac{5}{6}x' + \frac{1}{3}y' - \frac{1}{2}z' \\ z = \ \frac{7}{18}x' + \frac{1}{9}y' + \frac{1}{6}z'. \end{cases}$$

Le changement de coordonnées est bien bijectif, les nouvelles coordonnées sont associées à la base (i', j', k') définie par la matrice de passage $P = A^{-1}$:

$$P = \begin{pmatrix} -\frac{1}{9} & \frac{1}{9} & -\frac{1}{3} \\ -\frac{5}{6} & \frac{1}{3} & -\frac{1}{2} \\ \frac{7}{18} & \frac{1}{9} & \frac{1}{6} \end{pmatrix} \qquad \text{soit} \quad \begin{cases} i' = -\frac{1}{9}i - \frac{5}{6}j + \frac{7}{18}k \\ j' = \ \frac{1}{9}i + \frac{1}{3}j + \frac{1}{9}k \\ k' = -\frac{1}{3}i - \frac{1}{2}j + \frac{1}{6}k. \end{cases}$$

Sommes et sommes directes de sous-espaces vectoriels. Soient $F_1, \ldots, F_m$ des sous-espaces vectoriels de E. La **somme** des sous-espaces $F_1, \ldots, F_m$ est le sous-espace vectoriel $F_1 + \ldots + F_m$ de E défini par

$$F_1 + \ldots + F_m = \big\{ v = v_1 + \ldots + v_m \,;\ v_i \in F_i \big\}$$

(il est très facile de vérifier que c'est bien un sous-espace vectoriel). Si on prend une base $\mathcal{B}_i$ de F_i et une famille $\mathcal{B}$ qui est la réunion des bases $\mathcal{B}_i$, il est facile de voir que c'est une famille génératrice (mais non nécessairement une base) de $F_1 + \ldots + F_m$. On a donc en général

$$\dim_{\mathbb{K}}(F_1 + \ldots + F_m) \leq \dim_{\mathbb{K}} F_1 + \dim_{\mathbb{K}} F_2 + \ldots + \dim_{\mathbb{K}} F_m.$$

L'inégalité est stricte s'il on prend par exemple pour E un plan vectoriel réel, et $F_1 = D_1$, $F_2 = D_2$, $F_3 = D_3$ trois droites deux à deux distinctes de ce plan. Dans ce cas, si e_i est un vecteur directeur de D_i, on voit que (e_1, e_2, e_3) est un système générateur de $E = D_1 + D_2 + D_3$, mais ce n'est pas une famille libre.

On dit que $F_1, \ldots, F_m$ sont en **somme directe** si pour tout $v_1 \in F_1, \ldots, v_m \in F_m$, on a

$$v_1 + \ldots + v_m = 0 \quad \Rightarrow \quad v_1 = \ldots = v_m = 0.$$

Par différence de deux décompositions donnant le même vecteur v

$$v = v_1 + \ldots + v_m = v_1' + \ldots + v_m' \qquad \text{avec } v_i, v_i' \in F_i,$$

on a $0 = (v_1 - v_1') + \ldots + (v_m - v_m')$ et donc $v_i - v_i' = 0$, soit $v_i' = v_i$; on voit ainsi que l'écriture d'une somme $v = v_1 + \ldots + v_m$ avec $v_i \in F_i$ est **unique**, et cette propriété caractérise les sommes directes (prendre $v = 0$ et $v_i' = 0$).

Dans cette situation, on dit que $F = F_1 + \ldots + F_m$ est la **somme directe** des sous-espaces $F_1, \ldots, F_m$ et on écrit

$$F = F_1 \oplus \ldots \oplus F_m.$$

De l'unicité de la décomposition on déduit facilement que l'on obtient une base $\mathcal{B}$ de F en prenant la réunion de bases $\mathcal{B}_i$ des F_i. On a donc bien ici

$$\dim_{\mathbb{K}}(F_1 \oplus \ldots \oplus F_m) = \dim_{\mathbb{K}} F_1 + \dim_{\mathbb{K}} F_2 + \ldots + \dim_{\mathbb{K}} F_m,$$

et en dimension finie cette propriété caractérise les sommes directes.

Si $m = 2$, on vérifie immédiatement que F_1 et F_2 sont en somme directe si et seulement si on a $F_1 \cap F_2 = \{0\}$ (en revanche l'exemple ci-dessus de 3 droites D_i d'un plan montre que $D_1 + D_2 + D_3$ n'est pas en somme directe, bien que $D_1 \cap D_2 = D_2 \cap D_3 = D_1 \cap D_3 = \{0\}$.)

Exemple 1.10. Si $e_1, \ldots, e_n$ est une base de E, alors

$$E = \mathbb{K}e_1 \oplus \cdots \oplus \mathbb{K}e_n.$$

Par exemple

$$\mathbb{R}^3 = \mathbb{R}(1,0,0) \oplus \mathbb{R}(0,1,0) \oplus \mathbb{R}(0,0,1).$$

Applications linéaires. *Soient E, E' deux $\mathbb{K}$-espaces vectoriels. Une* **application $\mathbb{K}$-linéaire** *de E dans E' est une application $f : E \to E'$ vérifiant les 2 propriétés :*

(1) $f(v_1 + v_2) = f(v_1) + f(v_2)$ pour tous $v_1, v_2 \in E$,

(2) $f(\lambda v) = \lambda f(v)$ *pour tous* $\lambda \in \mathbb{K}$, $v \in E$.

Il est équivalent de vérifier (1) *et* (2) *ou la propriété équivalente de transformation par* f *des combinaisons linéaires :*

(3) $f(\lambda_1 v_1 + \lambda_2 v_2) = \lambda_1 f(v_1) + \lambda_2 f(v_2)$ *pour tous* $\lambda_1, \lambda_2 \in \mathbb{K}$, $v_1, v_2 \in E$.

Dans ce cas, on a nécessairement $f(0) = 0$ (comme on le voit en faisant $\lambda = 0$ dans l'axiome (2)).

Le **noyau** de f, noté $\operatorname{Ker} f$, est l'ensemble

$$\operatorname{Ker} f = \big\{ v \in E \,;\; f(v) = 0 \big\} \subset E.$$

C'est un sous-espace vectoriel de E.

L'image de f, notée $\operatorname{Im} f$, est l'ensemble

$$\operatorname{Im} f = \big\{ f(v), v \in E \big\} \subset E'.$$

C'est un sous-espace vectoriel de E'.

Notation 1.11. *On notera* $\mathcal{L}_{\mathbb{K}}(E; E')$ *l'ensemble des applications* $\mathbb{K}$*-linéaires de* E *dans* E' (*et on se permettra souvent d'omettre le corps* $\mathbb{K}$ *s'il n'y a pas d'ambiguité possible*).

Si $\mathcal{B} = (e_1, \ldots, e_p)$ est une base de E et $\mathcal{B}' = (e'_1, \ldots, e'_n)$ une base de E', toute application linéaire $f \in \mathcal{L}(E; E')$ s'exprime sous la forme

$$x = x_1 e_p + \ldots + x_n e_p \longmapsto x' = f(x) = x_1 f(e_1) + \ldots + x_p f(e_p).$$

Si l'on pose $f(e_j) = \sum_{1 \le i \le n} a_{ij} e'_i$ dans la base (e'_i) de E', $a_{ij} \in \mathbb{K}$, on a

$$x' = \sum_{j=1}^{p} x_j \Big(\sum_{i=1}^{n} a_{ij} e'_i \Big) = \sum_{i=1}^{n} \Big(\sum_{j=1}^{p} a_{ij} x_j \Big) e'_i.$$

La matrice colonne des coordonnées X' de x' dans $\mathcal{B}'$ s'exprime donc par

$$x'_i = \sum_{j=1}^{p} a_{ij} x_j \,, \quad \text{soit} \quad X' = AX, \quad \text{où} \quad A = (a_{ij})_{1 \le i \le p,\ 1 \le j \le n} = \operatorname{Mat}_{\mathcal{B},\mathcal{B}'}(f)$$

est par définition la matrice de f relativement aux bases $\mathcal{B}$ et $\mathcal{B}'$; c'est une matrice à n lignes et p colonnes à coefficients dans $\mathbb{K}$, avec $n = \dim_{\mathbb{K}} E'$ et $p = \dim_{\mathbb{K}} E$.

Formule générale de changement de base. Supposons que l'on change simultanément les bases de E et de E' : dans E, on remplace la base $\mathcal{B} = (e_j)$ par une base $\widetilde{B} = (\widetilde{e}_j)$ définie par la matrice de passage P, et dans E' la base $\mathcal{B}' = (e'_i)$ par une base $\widetilde{B}' = (\widetilde{e}'_i)$ définie par une matrice de passage P'. Étant donné une application linéaire $f \in \mathcal{L}_{\mathbb{K}}(E; E')$, la question est de trouver la nouvelle matrice

$$\widetilde{A} = \operatorname{Mat}_{\widetilde{\mathcal{B}},\widetilde{\mathcal{B}}'}(f) \quad \text{en fonction de} \quad A = \operatorname{Mat}_{\mathcal{B},\mathcal{B}'}(f).$$

Avec des notations évidentes, on a

$$X' = AX, \quad X = P\widetilde{X}, \quad X' = P'\widetilde{X}'.$$

Ceci donne

$$\widetilde{X}' = (P')^{-1} X' = (P')^{-1} AX = (P')^{-1} AP \widetilde{X}.$$

On voit donc que la nouvelle matrice de f est donnée par

$$\operatorname{Mat}_{\widetilde{\mathcal{B}},\widetilde{\mathcal{B}}'}(f) = \widetilde{A} = (P')^{-1} AP,$$

si P et P' sont les matrices de passage dans E et E' respectivement.

Théorème 1.12 (Théorème du rang)**.** *Soient E, E' deux $\mathbb{K}$-espaces vectoriels, et soit $f : E \to E'$ une application linéaire. Si E est de dimension finie, alors* $\operatorname{Ker} f$ *et* $\operatorname{Im} f$ *sont de dimension finie et on a*

$$\dim_{\mathbb{K}} \operatorname{Ker} f + \dim_{\mathbb{K}} \operatorname{Im} f = \dim_{\mathbb{K}} E.$$

Démonstration. Comme $\operatorname{Ker} f$ est un sous-espace vectoriel de E, il est nécessairement de dimension finie. Soit $(e_1, \ldots, e_p)$ une base de $\operatorname{Ker} f$, avec $p = \dim_{\mathbb{K}} \operatorname{Ker} f$. On la complète en une base $(e_1, \ldots, e_n)$ de E, où $n = \dim_{\mathbb{K}} E \geq p$. On a $f(e_1) = \ldots = f(e_p) = 0$ puisque ces vecteurs e_i sont dans $\operatorname{Ker} f$. L'image $\operatorname{Im} f$ est par définition l'ensemble des vecteurs images $w = f(x_1e_1 + \ldots + x_ne_n)$. Comme f est linéaire, il vient

$$w = f(x_1e_1 + \ldots + x_ne_n) = x_{p+1}f(e_{p+1}) + \ldots + x_nf(e_n) = f(x_{p+1}e_{p+1} + \ldots + x_ne_n),$$

et on voit déjà que la famille $\mathcal{G} = (f(e_{p+1}), \ldots, f(e_n))$ est une famille génératrice de $\operatorname{Im} f$. Montrons que c'est une base : il reste à voir que $\mathcal{G}$ est libre. Pour cela, supposons $w = 0$. Alors $v = x_{p+1}e_{p+1} + \ldots + x_ne_n \in \operatorname{Ker} f$, et comme $(e_1, \ldots, e_p)$ est une base de $\operatorname{Ker} f$, il existe des scalaires $v_1, \ldots, v_p \in \mathbb{K}$ tels que

$$v = x_{p+1}e_{p+1} + \ldots + x_ne_n = v_1e_1 + \ldots + v_pe_p.$$

Maintenant, comme $(e_1, \ldots, e_n)$ est une base de E, on en conclut que $x_{p+1} = \ldots = x_n = 0$ (et aussi $v_1 = \ldots = v_p = 0$). Ceci entraîne que $\mathcal{G}$ est bien libre. On a donc $\dim_{\mathbb{K}} \operatorname{Im} f = n - p$ et

$$\dim_{\mathbb{K}} \operatorname{Ker} f + \dim_{\mathbb{K}} \operatorname{Im} f = p + (n - p) = n = \dim_{\mathbb{K}} E.$$

□

Remarque complémentaire. Si $S = \operatorname{Vect}(e_{p+1}, \ldots, e_n)$, alors on a par construction la somme directe

$$E = \operatorname{Ker} f \oplus S,$$

et la restriction $f_{|S} : S \to \operatorname{Im} f$ est une bijection ("isomorphisme" d'espaces vectoriels), envoyant la base $(e_{p+1}, \ldots, e_n)$ de S sur la base $\mathcal{G} = (f(e_{p+1}), \ldots, f(e_n))$ de $\operatorname{Im} f$.

Définition 1.13. Le rang d'une application linéaire $f : E \to E'$ est par définition

$$\operatorname{rang}_{\mathbb{K}}(f) = \dim_{\mathbb{K}} \operatorname{Im} f$$

Il existe plusieurs méthodes pour calculer le rang, l'une d'elles est de calculer $\operatorname{Ker} f$ et sa dimension, puis d'appliquer le théorème du rang. Une autre est d'observer que pour toute base $\mathcal{B} = (\varepsilon_1, \ldots, \varepsilon_n)$ de E, la famille $\mathcal{G} = (f(\varepsilon_1), \ldots, f(\varepsilon_n))$ constituée par les vecteurs colonnes de $\operatorname{Mat}_{\mathcal{B}}(f)$ est une famille génératrice de $\operatorname{Im} f$. On cherche alors à éliminer les vecteurs qui sont combinaisons linéaires des autres pour en extraire une base de $\operatorname{Im} f$; on remarquera que ces combinaisons linéaires $\sum \lambda_j f(\varepsilon_j) = 0$ s'obtiennent précisément en cherchant les vecteurs $x = \lambda_j \varepsilon_j$ tels que $f(x) = 0$.

Formes linéaires. Une **forme linéaire** sur E est par définition une application linéaire de E dans $\mathbb{K}$.

Lemme 1.14. *Soit E un $\mathbb{K}$-espace vectoriel de dimension n, et soit $f : E \to \mathbb{K}$ une forme linéaire non nulle. Alors* $\dim_{\mathbb{K}} \operatorname{Ker} f = n - 1$.

Démonstration. D'après le théorème du rang, il suffit de montrer que $\dim_{\mathbb{K}} \operatorname{Im} f = 1$. En fait, on va montrer que $\operatorname{Im} f = \mathbb{K}$, ce qui entraînera que $\dim_{\mathbb{K}} \operatorname{Im} f = \dim_{\mathbb{K}} \mathbb{K} = 1$.

Puisque f est supposée non nulle, il existe $x_0 \in E$ tel que $f(x_0) \neq 0$. Soit maintenant $\lambda \in \mathbb{K}$. On a

$$f\Big(\frac{\lambda}{f(x_0)} x_0\Big) = \frac{\lambda}{f(x_0)} f(x_0) = \lambda,$$

et donc $\lambda \in \operatorname{Im} f$. Ceci étant vrai pour tout $\lambda \in \mathbb{K}$, on obtient $\operatorname{Im} f = \mathbb{K}$, ce qui achève la démonstration. □

Remarque 1.15. Le résultat peut être approché de manière plus calculatoire comme suit : soit $(e_1, \ldots, e_n)$ une base E. Posons $a_i = f(e_i) \in \mathbb{K}$. Alors pour tout $x = x_1e_1 + \ldots + x_ne_n$

$$f(x) = x_1f(e_1) + \ldots + x_nf(e_n) = a_1x_1 + \ldots + a_nx_n.$$

On a donc

$$f(x) = 0 \iff a_1x_1 + \ldots + a_nx_n = 0.$$

Puisque f est non nulle, un des a_i est non nul, et donc

$$f(x) = 0 \iff x_i = -\Big(\frac{a_1}{a_i}x_1 + \ldots + \frac{a_{i-1}}{a_i}x_{i-1} + \frac{a_{i+1}}{a_i}x_{i+1} + \ldots + \frac{a_n}{a_i}x_n\Big)$$

$$\iff \begin{pmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ \vdots \\ -\frac{a_1}{a_i} \\ \vdots \\ 0 \end{pmatrix} + \ldots + x_{i-1} \begin{pmatrix} \vdots \\ 1 \\ -\frac{a_{i-1}}{a_i} \\ \vdots \\ 0 \end{pmatrix} + x_{i+1} \begin{pmatrix} 0 \\ \vdots \\ -\frac{a_{i+1}}{a_i} \\ 1 \\ \vdots \end{pmatrix} + \ldots + x_n \begin{pmatrix} 0 \\ \vdots \\ -\frac{a_n}{a_i} \\ \vdots \\ 1 \end{pmatrix}$$

où chaque matrice colonne du membre de droite n'a que deux coefficients non nuls au plus. Ces $(n-1)$ matrices colonnes sont les coordonnées d'une base de $\operatorname{Ker} f$ relativement à la base $(e_1, \ldots, e_n)$ de E.

Espace dual. *Si E est un $\mathbb{K}$-espace vectoriel, on appelle* **dual** *de E le $\mathbb{K}$-espace vectoriel $E^* = \mathcal{L}_{\mathbb{K}}(E; \mathbb{K})$ des formes linéaires sur E.*

Suupposons que E soit de dimension finie n, muni d'une base $(e_1, \ldots, e_n)$, et soit $\ell \in E^*$ une forme linéaire. Alors pour $x = x_1e_1 + \ldots + x_ne_n$ on a

$$\ell(x) = x_1\ell(e_1) + \ldots + x_n\ell(e_n) = a_1x_1 + \ldots + a_nx_n$$

avec $a_i = \ell(e_i)$. Si on utilise comme base de $\mathbb{K}$ la base canonique (1), la matrice de ℓ relativement aux bases $(e_1, \ldots, e_n)$ de E et (1) de $\mathbb{K}$ est la matrice ligne $A = (a_1, \ldots, a_n)$. En identifiant les scalaires aux matrices 1×1 on a

$$\ell(x) = (a_1, \ldots, a_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = AX$$

où X est la matrice colonne de x dans $(e_1, \ldots, e_n)$. On introduit maintenant les **formes linéaires coordonnées**, notée e_i^*, telles que

$$e_i^*(x) = x_i = (0, \ldots, 0, 1, 0, \ldots, 0) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \text{(le 1 étant en position } i\text{)}.$$

On voit aussitôt que ces formes sont linéairement indépendantes, que $e_i^*(e_j) = \delta_{ij}$ (symbole de Kronecker, égal à 1 si $i = j$ et 0 si $i \neq j$), et que

$$\ell(x) = a_1e_1^*(x) + \ldots + a_ne_n^*(x), \qquad \text{soit} \quad \ell = a_1e_1^* + \ldots + a_ne_n^*.$$

Il en résulte que $(e_1^*, \ldots, e_n^*)$ est une base de E^*. On l'appelle la **base duale** de la base $(e_1, \ldots, e_n)$. On notera que les coordonnées de ℓ dans la base $(e_1^*, \ldots, e_n^*)$ sont précisément les coefficients $a_i = \ell(e_i)$. Il résulte aussi de ce qui précède qu'on a toujours

$$\dim_{\mathbb{K}} E^* = n = \dim_{\mathbb{K}} E.$$

Endomorphismes, sous-espace stables, valeurs propres et vecteurs propres.

Un cas très important est le cas où $E' = E$ (les espaces de départ et d'arrivée coïncident), on dit alors qu'une application linéaire $f \in \mathcal{L}_{\mathbb{K}}(E;E)$ est un **endomorphisme** et on note

$$\mathrm{End}_{\mathbb{K}}(E) = \mathcal{L}_{\mathbb{K}}(E;E).$$

Dans ce cas, on dit qu'un sous-espace vectoriel $S \subset E$ est **stable** si $f(S) \subset S$ (on notera que cette propriété n'a pas de sens si $f \in \mathcal{L}_{\mathbb{K}}(E;E')$ et que E et E' sont sans rapport l'un avec l'autre). On parlera de **droite stable** ou de **plan stable** si $\dim_{\mathbb{K}} S = 1$, resp. $\dim_{\mathbb{K}} S = 2$.

Supposons en particulier que D soit une droite stable, et soit $v \neq 0$ un vecteur directeur de D. Comme $D = \{tv\,;\ t \in \mathbb{K}\}$ et que $f(tv) = tf(v)$, on a $f(D) \subset D$ si et seulement si $f(v) \in D$, c'est-à-dire si

$$\exists \lambda \in \mathbb{K} \quad \text{tel que} \quad f(v) = \lambda v.$$

Un vecteur $v \neq 0$ vérifiant cette propriété s'appelle un **vecteur propre** de f, et on dit que le scalaire $\lambda \in \mathbb{K}$ est la **valeur propre** associée à v (comme $f(tv) = t\,f(v) = \lambda(tv)$, la valeur propre associée à un vecteur colinéaire $\tilde{v} = tv$ est aussi égale à λ). **Ne jamais oublier que l'on doit avoir** $v \neq 0$ **!**

Recherche des valeurs propres et vecteurs propres. Soit $(e_1, \ldots, e_n)$ une base de E, supposé de dimension finie n, et soit $A = \mathrm{Mat}_{\mathcal{B},\mathcal{B}}(f)$. Rechercher les vecteurs propres de f revient à résoudre l'équation

$$AX = \lambda X, \quad X \neq 0 \quad \Longleftrightarrow \quad (A - \lambda I)X = 0, \quad X \neq 0.$$

Autrement dit, l'endomorphisme $f - \lambda\,\mathrm{Id}_E$ de matrice $A - \lambda I$ doit avoir un noyau non réduit à $\{0\}$ (puisqu'il contient le vecteur v de matrice $X \neq 0$), ce qui impose que $\det(A - \lambda I) = 0$. Or, d'après les propriétés du déterminant, $\chi_A(\lambda) = \det(A - \lambda I)$ est un polynôme de degré n à coefficients dans le corps $\mathbb{K}$, de terme de plus haut degré $(-1)^n\lambda^n$. Le polynôme $\chi_A(\lambda)$est appelé **polynôme caractéristique** de la matrice A.

On peut vérifier que ce polynôme **ne dépend pas** de la base $\mathcal{B}$ dans laquelle la matrice de f est exprimée : étant donné une nouvelle base $\widetilde{B}$ définie par la matrice de passage P, la nouvelle matrice de f est donnée par $\widetilde{A} = P^{-1}AP$ et on a

$$\begin{aligned} \det(\widetilde{A} - \lambda I) &= \det(P^{-1}AP - \lambda I) = \det(P^{-1}(A - \lambda I)P) \\ &= (\det(P))^{-1}\det(A - \lambda I)\det(P) = \det(A - \lambda I). \end{aligned}$$

On parlera donc aussi de polynôme caractéristique de l'endomorphisme f, et on notera $\chi_f(\lambda) = \chi_A(\lambda)$. En développant le déterminant, on voit que

$$\chi_A(\lambda) = (-1)^n\lambda^n + (-1)^{n-1}\,\mathrm{Tr}(A)\lambda^{n-1} + \ldots + \det(A)$$

où $\mathrm{Tr}(A) = \sum_{i=1}^{n} a_{ii}$ est la **trace** de A. Les coefficients $\mathrm{Tr}(A)$ et $\det(A)$ ne dépendent eux aussi que de f (mais pas de la base $\mathcal{B}$), on posera donc $\mathrm{Tr}(f) = \mathrm{Tr}(A)$ et $\det(f) = \det(A)$.

Lorsque l'on est sur le corps $\mathbb{K} = \mathbb{R}$, il peut arriver que l'équation $\chi_f(\lambda) = 0$ n'ait aucune racine, mais sur $\mathbb{K} = \mathbb{C}$ on sait (théorème de d'Alembert-Gauss) que le polynôme $\chi_f(\lambda)$ de degré n **admet exactement** n **racines** $\lambda_1, \ldots, \lambda_n$, lorsque celles-ci sont comptées avec multiplicités (il peut y avoir des racines multiples, et éventuellement, il peut n'y avoir qu'une seule racine λ_1 de multiplicité n). Les vecteurs propres associés à la valeur propre λ_j se calculent en résolvant explicitement le système linéaire

$$(A - \lambda_j I)X = 0.$$

On obtient ainsi un sous-espace vectoriel S_j formé de vecteurs v tels que $f(v) = \lambda_j v$, appelé **espace propre** associé à la valeur propre λ_j. En conséquence :

Théorème 1.16. *Si $f \in \mathcal{L}_{\mathbb{C}}(E;E)$ est un endomorphisme sur un espace vectoriel E de dimension finie $n \geq 1$ sur le corps des complexes, alors f possède au moins une droite stable D (la droite $D = \mathbb{C}v$ associée à un vecteur propre).*

On remarquera que ce résultat est faux sur le corps $\mathbb{K} = \mathbb{R}$, il suffit de prendre pour f une rotation d'angle α dans un plan, avec $\alpha \neq 0, \pi$ modulo 2π. Néanmoins, sur $\mathbb{K} = \mathbb{R}$, s'il existe une valeur propre réelle, le résultat précédent est correct. Sinon, il existe certainement une racine complexe $\lambda = \alpha + i\beta$ du polynôme caractéristique, et la résolution du système linéaire associé donne une matrice colonne complexe $Z = X + iY$ non nulle solution de $AZ = \lambda Z$, c'est-à-dire

$$AX + iAY = (\alpha + i\beta)(X + iY) \Longleftrightarrow AX = \alpha X - \beta Y, \quad AY = \beta X + \alpha Y.$$

Si $\lambda \notin \mathbb{R}$, les vecteurs colonnes X et Y ne peuvent être $\mathbb{R}$-colinéaires sinon on aurait disons $X \neq 0$, $Y = tX$ et $Z = X + itX = (1 + it)X$ serait $\mathbb{C}$-colinéaire à X, de sorte que $X = (1 + it)^{-1}Z$ serait aussi vecteur propre de A pour la valeur propre λ (mais c'est contradictoire avec la relation $AX = \lambda X$ où A, X sont réels et λ non réel ; le raisonnement est le même si $Y \neq 0$ et $X = tY$). Considérons alors le plan vectoriel $P \subset E$ engendré par les vecteurs u, v de coordonnées X, Y. On a les relations

$$f(u) = \alpha u - \beta v \in P, \quad f(v) = \beta u + \alpha v \in P \Longrightarrow f(P) \subset P$$

et on voit que P est un plan stable de f. On peut énoncer :

Théorème 1.17. *Si $f \in \mathcal{L}_{\mathbb{R}}(E;E)$ est un endomorphisme sur un espace vectoriel E de dimension finie $n \geq 1$ sur le corps des réels, alors f possède au moins une droite stable D (s'il y a une valeur propre réelle) ou un plan stable P (si la matrice associée possède une valeur propre complexe non réelle).*

2. Formes bilinéaires.

Soient E, E', F des $\mathbb{K}$-espaces vectoriels, et soit

$$\varphi : E \times E' \to F, \quad (x, y) \mapsto \varphi(x, y)$$

une application de $E \times E'$ dans F.

Définition 2.1. *On dit que φ est* **une application bilinéaire** *si φ est linéaire en chacune des variables. Autrement dit φ est une application bilinéaire si pour tous $x_1, x_2, x \in E$, $y_1, y_2, y \in E'$, et tous $\lambda, \mu \in \mathbb{K}$, on a*

$$\varphi(x_1 + x_2, y) = \varphi(x_1, y) + \varphi(x_2, y), \quad \varphi(\lambda x, y) = \lambda\varphi(x, y) \quad (\textit{linéarité en } x),$$
$$\varphi(x, y_1 + y_2) = \varphi(x, y_1) + \varphi(x, y_2), \quad \varphi(x, \mu y) = \mu\varphi(x, y) \quad (\textit{linéarité en } y).$$

En prenant $\lambda = \mu = 0$, on voit qu'on a nécessairement

$$\varphi(0, y) = \varphi(x, 0) = 0 \text{ pour tous } x, y \in E.$$

Une autre façon équivalente de formuler les axiomes de la bilinéarité est de vérifier l'unique identité de "distributivité"

$$\varphi(\lambda_1 x_1 + \lambda_2 x_2, \mu_1 y_1 + \mu_2 y_2) = \lambda_1\mu_1\varphi(x_1, y_1) + \lambda_1\mu_2\varphi(x_1, y_2) + \lambda_2\mu_1\varphi(x_2, y_1) + \lambda_2\mu_2\varphi(x_2, y_2),$$

(mais cette forme plus concise n'est pas nécessairement la plus pratique).

Définition 2.2. *Une* **forme bilinéaire** *est une application bilinéaire $\varphi : E \times E' \to \mathbb{K}$, c'est-à-dire que l'espace d'arrivée $F = \mathbb{K}$ est le corps des scalaires.*

Dans la plus grande partie de ce qui suit, on ne considérera que le cas où $E = E'$, c'est-à-dire le cas où les vecteurs x, y sont pris dans le **même espace** vectoriel E. On dit alors que

- $\varphi : E \times E \to F$ est **symétrique** si $\varphi(y, x) = \varphi(x, y)$ pour tous $x, y \in E$.
- $\varphi : E \times E \to F$ est **anti-symétrique** si $\varphi(y, x) = -\varphi(x, y)$ pour tous $x, y \in E$.

Exemples 2.3.

(1) L'application

$$\varphi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \quad (x, y) = ((x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n)) \mapsto x \cdot y = \sum_{i=1}^{n} x_i y_i$$

est une forme bilinéaire symétrique. Lorsque $n = 2$ ou 3, on retrouve le produit scalaire usuel des vecteurs.

(2) Identifions ici $\mathbb{R}^3$ aux vecteurs colonnes de dimension 3. L'application

$$\varphi : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3, \quad (x, y) = \left(\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \right) \mapsto x \wedge y = \begin{pmatrix} x_2 y_3 - x_3 y_2 \\ x_3 y_1 - x_1 y_3 \\ x_1 y_2 - x_2 y_1 \end{pmatrix}$$

est une application bilinéaire anti-symétrique (mais ce n'est pas une *forme* bilinéaire).

(3) L'application

$$\varphi : \mathrm{M}_{n\times n}(\mathbb{R}) \times \mathrm{M}_{n\times n}(\mathbb{R}) \to \mathrm{M}_{n\times n}(\mathbb{R}), \quad (M, N) \mapsto [M, N] = MN - NM$$

est une application bilinéaire anti-symétrique.

(4) L'application

$$\varphi : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}, \quad ((x_1, x_2), (y_1, y_2)) \mapsto x_1 x_2 + 2 x_1 y_2$$

n'est pas bilinéaire. En effet, posons $x = (x_1, x_2)$, $y = (y_1, y_2)$. On a

$$\varphi(\lambda x, y) = (\lambda x_1)(\lambda x_2) + 2(\lambda x_1) y_2 = \lambda^2 x_1 x_2 + 2\lambda x_1 y_2 \neq \lambda \varphi(x, y) = \lambda(x_1 x_2 + 2 x_1 y_2)$$

en général (prendre par exemple $x_1 = x_2 = y_1 = y_2 = 1$, $\lambda = 2$).

(5) Soit $E = C^0([a, b], \mathbb{R})$ l'ensemble des fonctions continues sur l'intervalle $[a, b]$, à valeurs dans $\mathbb{R}$. Alors l'application $\varphi : E \times E \to \mathbb{R}$ définie par

$$\varphi(f, g) = \int_a^b f(x) g(x)\, dx$$

est une *forme $\mathbb{R}$-bilinéaire.*

(6) Lorsque $E = C^0([a, b], \mathbb{C})$ est l'espace des fonctions continues à valeurs complexes, la formule du (5) définit cette fois une *forme $\mathbb{C}$-bilinéaire* $\varphi : E \times E \to \mathbb{C}$. Rappelons la définition de l'intégrale d'une fonction f à valeurs complexes : il existe alors deux fonctions $f_1, f_2 : [a, b] \to \mathbb{R}$ telles que $f = f_1 + i f_2$ et on pose

$$\int_a^b f(x)\, dx = \int_a^b f_1(x)\, dx + i \int_a^b f_2(x)\, dx \in \mathbb{C}.$$

Cette intégrale jouit de propriétés similaires à celles de l'intégrale d'une fonction à valeurs réelles et en particulier, il est facile de vérifier que $f \mapsto \int_a^b f(x)\, dx$ est $\mathbb{C}$-linéaire et que

$$\int_a^b \overline{f(x)}\, dx = \overline{\int_a^b f(x)\, dx} = \int_a^b f_1(x)\, dx - i \int_a^b f_2(x)\, dx.$$

Écriture matricielle d'une forme bilinéaire. Soit E un $\mathbb{K}$-espace vectoriel de dimension finie n, soit $\mathcal{B} = (e_1, \dots, e_n)$ une base de E, et soit $\varphi : E \times E \to \mathbb{K}$ une forme bilinéaire. Étant donnés des vecteurs

$$x = \sum_{i=1}^{n} x_i e_i \in E, \quad y = \sum_{i=1}^{n} y_i e_i \in E,$$

les propriétés de bilinéarité de φ permettent d'écrire

$$\varphi(x,y) = \varphi\Big(\sum_{i=1}^{n} x_i e_i, y\Big) = \sum_{i=1}^{n} x_i\, \varphi(e_i, y) = \sum_{i=1}^{n} x_i\, \varphi\Big(e_i, \sum_{j=1}^{n} y_j e_j\Big) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j\, \varphi(e_i, e_j).$$

Si l'on introduit les coefficients $c_{ij} = \varphi(e_i, e_j) \in \mathbb{K}$, ceci devient simplement

$$\varphi(x,y) = \sum_{1 \le i,j \le n} c_{ij} x_i y_j.$$

Inversement, toute expression de cette forme avec des coefficients $c_{ij} \in \mathbb{K}$ quelconques définit bien une forme bilinéaire de $E \times E$ dans $\mathbb{K}$.

Définition 2.4. *Si E est un $\mathbb{K}$-espace vectoriel de dimension finie n, $\mathcal{B} = (e_1, \dots, e_n)$ une base de E, et $\varphi : E \times E \to \mathbb{K}$ une forme bilinéaire, on appelle* **matrice représentative** *de φ dans la base $\mathcal{B}$ la matrice $n \times n$ de ses coefficients*:

$$\mathrm{Mat}_{\mathcal{B}}(\varphi) = C = (c_{ij}) = (\,\varphi(e_i, e_j)\,)_{1 \le i,j \le n}.$$

Rappelons que si $M = (c_{ij})$ est une matrice $n \times p$ (à n lignes et p-colonnes), on appelle **transposée** de M, notée M^t, la matrice $p \times n$ telle que

$$M^t = (\widetilde{m}_{ij}), \qquad \widetilde{m}_{ij} = m_{ji}$$

obtenue en transformant les lignes et colonnes et vice-versa. Il est évident que $(M^t)^t = M$. Rappelons aussi que si M, N sont des matrices $n \times p$ et $p \times r$ quelconques, alors

$$(MN)^t = N^t M^t.$$

Il est commode d'introduire ici les matrices colonnes

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

représentant les coordonnées des vecteurs $x, y \in E$ dans la base $\mathcal{B}$. On constate alors que l'on a l'égalité

$$\varphi(x,y) = \sum_{1 \le i,j \le n} c_{ij} x_i y_j = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} c_{11} & \dots & c_{1n} \\ \vdots & & \vdots \\ c_{n1} & \dots & c_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = X^t C\, Y$$

si l'on identifie les scalaires et les matrices 1×1 (ce que nous ferons toujours désormais). On notera que la matrice C est uniquement déterminée par φ dans toute base $\mathcal{B} = (e_i)$, puisque l'on doit avoir $C = (c_{ij})$ avec $c_{ij} = \varphi(e_i, e_j)$.

Si $\varphi : E \times E \to \mathbb{K}$ est une forme bilinéaire, sa **transposée** φ^t est la forme définie par

$$\varphi^t(x,y) = \varphi(y,x), \quad \forall x, y \in E\,;$$

elle est encore bilinéaire. Comme $\varphi^t(e_i, e_j) = \varphi(e_j, e_i)$, on voit aussitôt que

$$\mathrm{Mat}_{\mathcal{B}}(\varphi) = C \quad \Longrightarrow \quad \mathrm{Mat}_{\mathcal{B}}(\varphi^t) = C^t.$$

L'énoncé suivant est évident.

Théorème 2.5. *Soit E un espace vectoriel de dimension finie n, $\mathcal{B} = (e_1, \dots, e_n)$ une base de E, $\varphi : E \times E \to \mathbb{K}$ une forme bilinéaire et $C = \mathrm{Mat}_{\mathcal{B}}(\varphi)$ sa matrice. On a*

(1) φ symétrique $\Leftrightarrow \varphi^t = \varphi \Leftrightarrow C^t = C \Leftrightarrow c_{ji} = c_{ij}$ pour tous $1 \leq i, j \leq n$.

(2) φ anti-symétrique $\Leftrightarrow \varphi^t = -\varphi \Leftrightarrow C^t = -C \Leftrightarrow c_{ji} = -c_{ij}$ pour tous $1 \leq i, j \leq n$.

Dans le cas anti-symétrique, on notera qu'on a nécessairement $c_{ii} = 0$ sur la diagonale, tandis que dans le cas symétrique les coefficients diagonaux c_{ii} sont quelconques.

Exemples 2.6.

(1) L'application

$$\varphi : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}, \quad ((x_1, x_2), (y_1, y_2)) \mapsto x_1y_1 + x_2y_2 + 3x_1y_2 - x_2y_1$$

est une forme bilinéaire, et sa matrice représentative dans la base canonique de $\mathbb{R}^2$ est

$$C = \begin{pmatrix} 1 & 3 \\ -1 & 1 \end{pmatrix}.$$

Cette forme n'est *ni symétrique ni anti-symétrique.*

(2) Considérons l'application

$$\varphi : \mathbb{R}[X]_2 \times \mathbb{R}[X]_2 \to \mathbb{R}, \quad (P, Q) \mapsto P(1)Q(0).$$

On peut vérifier directement que φ est bilinéaire, mais un calcul explicite en coordonnées le montrera aussi. Pour cela, considérons la base $(1, X, X^2)$ de $\mathbb{R}_2[X]$. On écrit alors

$$P = x_1 + x_2X + x_3X^2, \quad Q = y_1 + y_2X + y_3X^2.$$

On a alors $\varphi(P, Q) = (x_1 + x_2 + x_3)y_1 = x_1y_1 + x_2y_1 + x_3y_1$. Donc φ est bilinéaire et sa matrice représentative dans la base $(1, X, X^2)$ est

$$\mathrm{Mat}_{(1,X,X^2)}(\varphi) = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

Cette forme n'est *ni symétrique ni anti-symétrique.*

(3) Considérons l'application

$$\varphi : \mathbb{R}[X]_2 \times \mathbb{R}[X]_2 \to \mathbb{R}, \quad (P, Q) \mapsto \int_0^1 P(t)Q(t)\, dt.$$

On voit que $\varphi(X^a, X^b) = \int_0^1 t^{a+b} dt = \frac{1}{a+b+1}$, donc

$$\mathrm{Mat}_{(1,X,X^2)}(\varphi) = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}.$$

Il s'agit ici d'une *forme bilinéaire symétrique.*

Remarque 2.7. Si $\varphi : E \times E \to \mathbb{K}$ est une forme bilinéaire quelconque, on peut écrire

$$\varphi = \frac{1}{2}(\varphi + \varphi^t) + \frac{1}{2}(\varphi - \varphi^t),$$

et on remarque que $\sigma = \frac{1}{2}(\varphi + \varphi^t)$ est symétrique, tandis que $\alpha = \frac{1}{2}(\varphi - \varphi^t)$ est anti-symétrique :

$$\sigma^t = \frac{1}{2}(\varphi^t + \varphi) = \sigma, \qquad \alpha^t = \frac{1}{2}(\varphi^t - \varphi) = -\alpha.$$

Inversement, si on a une décomposition $\varphi = \sigma + \alpha$ avec σ symétrique et α anti-symétrique, alors $\varphi^t = \sigma - \alpha$ et par conséquent $\sigma = \frac{1}{2}(\varphi + \varphi^t)$ et $\alpha = \frac{1}{2}(\varphi - \varphi^t)$, de sorte que cette

décomposition est unique. De façon équivalente, toute matrice $C = (c_{ij})$ se décompose de manière unique en une somme $C = S + A$ d'une matrice symétrique $S = (s_{ij})$ et d'une matrice anti-symétrique $A = (a_{ij})$ par les formules

$$S = \frac{1}{2}(C + C^t), \quad s_{ij} = \frac{1}{2}(c_{ij} + c_{ji}), \qquad A = \frac{1}{2}(C - C^t), \quad a_{ij} = \frac{1}{2}(c_{ij} - c_{ji}).$$

De façon plus abstraite, on peut dire aussi que l'espace $\mathrm{Bil}_{\mathbb{K}}(E)$ des formes bilinéaires est somme directe de ses sous-espaces $\mathrm{Sym}_{\mathbb{K}}(E)$, $\mathrm{Antisym}_{\mathbb{K}}(E)$ des formes bilinéaires symétriques (resp. anti-symétriques) :

$$\mathrm{Bil}_{\mathbb{K}}(E) = \mathrm{Sym}_{\mathbb{K}}(E) \oplus \mathrm{Antisym}_{\mathbb{K}}(E).$$

L'espace $\mathrm{Bil}_{\mathbb{K}}(E)$ est isomorphe à l'espace $\mathrm{M}_{n\times n}(\mathbb{K})$ des matrices (c_{ij}) carrées $n \times n$ à coefficients dans $\mathbb{K}$, donc

$$\dim_{\mathbb{K}} \mathrm{Bil}_{\mathbb{K}}(E) = n^2.$$

La matrice $C = (c_{ij})$ d'une forme $\varphi \in \mathrm{Antisym}_{\mathbb{K}}(E)$ est déterminée par les coefficients situés au dessus de la diagonale (soit $j > i$), ceux situés sous la diagonale étant alors donnés par $c_{ji} = -c_{ij}$. Une base est obtenue en prenant les matrices A_{ij} dont tous les coefficients sont nuls, sauf ceux d'indices ij et $j\,i$ $(j > i)$, égaux respectivement à 1 et -1. La dimension est donc égale au nombre d'éléments situés strictement au dessus de la diagonale, soit

$$\dim_{\mathbb{K}} \mathrm{Antisym}_{\mathbb{K}}(E) = \frac{n^2 - n}{2} = \frac{n(n-1)}{2}.$$

De même, une base des matrices symétriques est obtenue en prenant les matrices diagonales D_i n'ayant qu'un seul coefficient non nul égal à 1 en position ii, et les matrices S_{ij} ayant un 1 en position ij et $j\,i$ $(j > i)$, et 0 partout ailleurs. Au total

$$\dim_{\mathbb{K}} \mathrm{Sym}_{\mathbb{K}}(E) = \frac{n(n-1)}{2} + n = \frac{n(n+1)}{2}.$$

Formes quadratiques. *Si $\varphi : E \times E \to \mathbb{K}$ est une forme bilinéaire, on appelle forme quadratique q associée à φ l'application*

$$q : E \to \mathbb{K}, \quad x \mapsto q(x) = \varphi(x, x).$$

[Le but des formes bilinéaires symétriques est de fournir une généralisation du produit scalaire $x \cdot y$, celui des formes quadratiques est de généraliser le carré scalaire $x \cdot x$.]

On notera que pour tout scalaire $\lambda \in \mathbb{K}$ on a

$$\varphi(\lambda x, \lambda x) = \lambda^2 \varphi(x, x) \quad \text{donc} \quad q(\lambda x) = \lambda^2 q(x)$$

(d'où la terminologie de "forme quadratique"). En particulier l'application q **n'est pas linéaire**, et évidemment, elle n'est pas non plus bilinéaire. Si $C = (c_{ij}) = (\varphi(e_i, e_j))$ est la matrice des coefficients de φ dans une base $\mathcal{B} = (e_i)_{1\le i\le n}$ de E, on a bien sûr

$$q(x) = \varphi(x, x) = \sum_{1\le i,j\le n} c_{ij} x_i x_j,$$

q s'exprime donc comme un **polynôme homogène** du second degré dans les variables $x_1, x_2, \ldots, x_n$. Réciproquement, par définition, une telle expression est bien une forme quadratique associée à une forme bilinéaire.

Si φ est *anti-symétrique*, la relation $\varphi(y, x) = -\varphi(x, y)$ donne $\varphi(x, x) = -\varphi(x, x)$ quand $y = x$, donc $q(x) = \varphi(x, x) = 0$. Par conséquent, si l'on décompose $\varphi = \sigma + \alpha$ en la somme

d'une forme bilinéaire symétrique σ et d'une forme bilinéaire anti-symétrique α, on trouve simplement

$$q(x) = \varphi(x,x) = \sigma(x,x).$$

Ceci montre que l'on peut toujours se ramener au cas où φ est symétrique, quitte à la remplacer par sa partie symétrique $\sigma = \frac{1}{2}(\varphi+\varphi^t)$. On **supposera donc** la plupart du temps que φ est une forme bilinéaire **symétrique**, c'est-à-dire que $c_{ji} = c_{ij}$.

Identité du parallélogramme et formule de polarisation. On suppose ici que

$$q(x) = \varphi(x,x)$$

est la forme quadratique associée à une forme bilinéaire *symétrique* $\varphi : E \times E \to \mathbb{K}$. Alors pour tous vecteurs $x, y \in E$ on trouve

$$q(x+y) = \varphi(x+y, x+y) = \varphi(x,x) + \varphi(x,y) + \varphi(y,x) + \varphi(y,y)$$

par bilinéarité, ce qui, compte tenu de la symétrie de φ, donne les formules

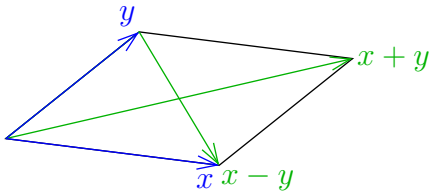
$$\begin{aligned} q(x+y) &= q(x) + 2\varphi(x,y) + q(y), \\ q(x-y) &= q(x) - 2\varphi(x,y) + q(y), \end{aligned}$$

la deuxième ligne se déduisant de la première en remplaçant y par $-y$. On en déduit alors par addition et soustraction :

$$\begin{aligned} &q(x+y) + q(x-y) = 2\big(q(x)+q(y)\big) && (\textit{identité du parallélogramme}), \\ &q(x+y) - q(x-y) = 4\varphi(x,y) && (\textit{formule de polarisation}). \end{aligned}$$

Dans le cas du produit scalaire usuel, l'identité du parallélogramme peut s'interpréter en disant que la somme des carrés des longueurs des diagonales d'un parallélogramme est égale à la somme des carrés des longueurs des côtés :

$$\|x+y\|^2 + \|x-y\|^2 = 2(\|x\|^2 + \|y\|^2) \ :$$



L'identité de polarisation, quant à elle, peut se récrire

$$\varphi(x,y) = \frac{1}{4}\big(q(x+y) - q(x-y)\big),$$

on dit aussi que φ est la **forme polaire** de q. On a une formule analogue permettant de retrouver φ à partir de q, donnée par l'expression à trois termes

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

Une conséquence de ces formules est le résultat suivant.

Théorème 2.8. *Soit* $\mathrm{Quad}_{\mathbb{K}}(E)$ *l'ensemble des formes quadratiques du* $\mathbb{K}$*-espace vectoriel* E*. C'est un* $\mathbb{K}$*-espace vectoriel et on a une bijection linéaire*

$$\mathrm{Sym}_{\mathbb{K}}(E) \longrightarrow \mathrm{Quad}_{\mathbb{K}}(E), \quad \varphi \longmapsto q \quad \textit{où } q(x) = \varphi(x,x)$$

(*la surjectivité résulte du fait que l'on peut toujours prendre* φ *symétrique, et l'injectivité des formules de polarisation calculant* φ *en fonction de* q*). En particulier*

$$\dim_{\mathbb{K}} \mathrm{Quad}_{\mathbb{K}}(E) = \dim_{\mathbb{K}} \mathrm{Sym}_{\mathbb{K}}(E) = \frac{n(n+1)}{2}.$$

Exemple 2.9. Considérons, dans $E = \mathbb{R}^3$ muni de sa base canonique $\mathcal{B} = (e_1, e_2, e_3)$, la forme quadratique

$$q(x_1, x_2, x_3) = -x_1^2 + 3x_1x_3 + 2x_2^2 - 5x_2x_3 - 4x_3^2, \quad x = (x_1, x_2, x_3) \in \mathbb{R}^3.$$

La question est de trouver la forme polaire φ et sa matrice dans la base $\mathcal{B}$. Dans cet exemple, en considérant un par un les termes carrés x_i^2 et les termes x_ix_j, $i \neq j$, on voit que la forme polaire φ est donnée par

$$\varphi((x_1, x_2, x_3), (y_1, y_2, y_3)) = -x_1y_1 + \frac{3}{2}(x_1y_3 + x_3y_1) + 2x_2y_2 - \frac{5}{2}(x_2y_3 + x_3y_2) - 4x_3y_3.$$

On notera le "dédoublement" des termes x_ix_j en $\frac{1}{2}(x_iy_j + x_jy_i)$ pour $i \neq j$, tandis que les carrés x_i^2 donnent des termes déjà symétriques x_iy_i. En effet, il est évident que φ est bilinéaire symétrique et que $\varphi(x, x) = q(x)$ (et on sait par ce qui précède que la forme polaire est unique). On en déduit

$$\mathrm{Mat}_{\mathcal{B}}(\varphi) = \begin{pmatrix} -1 & 0 & 3/2 \\ 0 & 2 & -5/2 \\ 3/2 & -5/2 & -4 \end{pmatrix}.$$

On n'oubliera pas que la matrice trouvée *doit être symétrique* !

Décomposition en sommes de carrés de formes linéaires. On observe d'abord que si $\ell_1, \ldots, \ell_r \in E^*$ sont des formes linéaires, et $a_1, \ldots, a_r \in \mathbb{K}$ des scalaires, la "somme de carrés"

$$q(x) = \sum_{j=1}^{r} a_j \ell_j(x)^2, \quad x \in E$$

est une forme quadratique, de forme bilinéaire symétrique associée

$$\varphi(x, y) = \sum_{j=1}^{r} a_j \ell_j(x)\ell_j(y), \quad x, y \in E.$$

Inversement, on va démontrer que toute forme quadratique peut être mise sous cette forme, grâce à une méthode due à Gauss (Carl Friedrich Gauss, 1777-1855, l'un des très grands mathématiciens de cette époque, auteur de nombreux travaux d'arithmétique, de géométrie et d'astronomie...)

Théorème 2.10. *Soit E un $\mathbb{K}$-espace vectoriel de dimension finie et q une forme quadratique sur E. Alors il existe une famille de formes linéaires* **indépendantes** *$\ell_1, \ldots, \ell_r \in E^*$ et des scalaires $a_1, \ldots, a_r \in \mathbb{K}$, $a_j \neq 0$ tels que*

$$q(x) = \sum_{j=1}^{r} a_j \ell_j(x)^2, \quad \forall x \in E$$

(*"décomposition en carrés de formes linéaires indépendantes"*).

Avant de donner la preuve générale, nous allons donner deux exemples.

Exemple 2.11. Considérons sur $E = \mathbb{Q}^3$ la forme quadratique

$$q(x, y, z) = -x^2 + 3xy + 2y^2 - 5xz + 3z^2 + 8yz, \quad (x, y, z) \in \mathbb{Q}^3.$$

On commence par regrouper tous les termes contenant l'une des variables, par exemple x :

$$-x^2 + 3xy - 5xz.$$

L'idée est d'écrire ceux-ci comme le début d'un carré, auquel on va retrancher les termes qui ne contient pas x :

$$-x^2+3xy-5xz=-\Big(x-\frac{3}{2}y+\frac{5}{2}z\Big)^2+\frac{9}{4}y^2+\frac{25}{4}z^2-\frac{15}{2}yz.$$

En incorporant ceux-ci dans $q(x,y,z)$, on obtient le carré déjà trouvé et des termes qui ne *contiennent plus la variable* x :

$$q(x,y,z)=-\Big(x-\frac{3}{2}y+\frac{5}{2}z\Big)^2+\frac{17}{4}y^2+\frac{37}{4}z^2+\frac{1}{2}yz.$$

On procède maintenant de même avec les termes restants qui contiennent l'une des autres variables, par exemple y :

$$\frac{17}{4}y^2+\frac{1}{2}yz=\frac{17}{4}\Big(y^2+\frac{2}{17}yz\Big)=\frac{17}{4}\Big(y+\frac{1}{17}z\Big)^2-\frac{1}{68}z^2.$$

En définitive, si $v=(x,y,z)\in\mathbb{Q}^3$, on obtient

$$\begin{aligned} q(x,y,z) &= -\Big(x-\frac{3}{2}y+\frac{5}{2}z\Big)^2+\frac{17}{4}\Big(y+\frac{1}{17}z\Big)^2+\frac{157}{17}z^2 \\ &= a_1\ell_1(v)^2+a_2\ell_2(v)^2+a_3\ell_3(v)^2, \end{aligned}$$

avec

$$\begin{aligned} a_1&=-1, & \ell_1(v)&=x-\frac{3}{2}y+\frac{5}{2}z,\\ a_2&=\frac{17}{4}, & \ell_2(v)&=y+\frac{1}{17}z,\\ a_3&=\frac{157}{17}, & \ell_3(v)&=z. \end{aligned}$$

On observera que les formes linéaires ℓ_1, ℓ_2, ℓ_3 sont forcément indépendantes car si on a $\lambda_1\ell_1+\lambda_2\ell_2+\lambda_3\ell_3=0$, on trouve successivement $\lambda_1=0$ du fait que ℓ_1 est la seule des trois formes à contenir x, puis $\lambda_2=0$ car ℓ_2 contient y mais pas ℓ_3, et enfin $\lambda_3=0$.

Exemple 2.12. Lorsque la forme quadratique ne contient aucun terme carré, on ne peut procéder comme ci-dessus et il faut utiliser une technique différente. Considérons par exemple la forme quadratique sur $E=\mathbb{R}^4$ définie par

$$q(x,y,z,t)=3xy+5yz+7zt-4yt+9xt,\quad (x,y,z,t)\in\mathbb{R}^4.$$

Dans ce cas on choisit un "terme rectangle" formé de 2 variables, tel que $3xy$, et on essaie de regouper tous les termes contenant x ou y en un produit $(x+[...])(y+[...])$ où les $[...]$ ne font intervenir que les autres variables, c'est-à-dire ici z et t. Ceci donne

$$\begin{aligned} 3xy+5yz-4yt+9xt &= 3\Big(xy+\frac{5}{3}yz-\frac{4}{3}yt+3xt\Big)\\ &= 3\Big(x+\frac{5}{3}z-\frac{4}{3}t\Big)\Big(y+3t\Big)-15zt+12t^2, \end{aligned}$$

et donc

$$q(x,y,z,t)=3\Big(x+\frac{5}{3}z-\frac{4}{3}t\Big)\Big(y+3t\Big)-8zt+12t^2.$$

Les termes restants (ici $-8zt+12t^2$) ne doivent plus faire intervenir que les variables non encore traitées, soient z et t. On peut ramener le premier produit AB à une différence de deux carrés grâce à l'identité de polarisation élémentaire

$$AB=\frac{1}{4}(A+B)^2-\frac{1}{4}(A-B)^2.$$

Il vient

$$q(x,y,z,t) = \frac{3}{4}\Big(x+y+\frac{5}{3}z+\frac{5}{3}t\Big)^2 - \frac{3}{4}\Big(x-y+\frac{5}{3}z-\frac{13}{3}t\Big)^2 - 8zt + 12t^2.$$

Il reste à transformer les derniers termes en carrés, on trouve par exemple :

$$-8zt+12t^2 = 12\Big(t^2-\frac{2}{3}zt\Big) = 12\Big(t-\frac{1}{3}z\Big)^2 - \frac{4}{3}z^2.$$

En définitive on obtient la décomposition en carrés de formes linéaires

$$q(x,y,z,t) = \frac{3}{4}\Big(x+y+\frac{5}{3}z+\frac{5}{3}t\Big)^2 - \frac{3}{4}\Big(x-y+\frac{5}{3}z-\frac{13}{3}t\Big)^2 + 12\Big(t-\frac{1}{3}z\Big)^2 - \frac{4}{3}z^2.$$

et on peut vérifier que ces formes sont indépendantes dans $(\mathbb{R}^4)^*$.

Preuve et cas général de la méthode de Gauss. On travaille par récurrence sur la dimension n (si $n=1$, $q(x)=c_{11}x_1^2$ est déjà un carré et il n'y a rien à faire). En général, soit à décomposer en carrés une forme quadratique

$$q(x) = \sum_{1\le i,j\le n} c_{ij}x_ix_j, \quad x=(x_1,x_2,\dots,x_n)\in\mathbb{K}^n.$$

$(*)$ Si la somme contient un terme carré non nul, disons $c_{11}x_1^2$, $c_{11}\neq 0$, on regroupe tous les termes contenant x_1, c'est-à-dire

$$\begin{aligned} c_{11}x_1^2 + 2c_{12}x_1x_2 + \ldots + 2c_{1n}x_1x_n &= c_{11}\Big(x_1^2 + 2\frac{c_{12}}{c_{11}}x_1x_2 + \ldots + 2\frac{c_{1n}}{c_{11}}x_1x_n\Big) \\ &= c_{11}\Big(x_1+\frac{c_{12}}{c_{11}}x_2+\ldots+\frac{c_{1n}}{c_{11}}x_n\Big)^2 - q'(x_2,\dots,x_n) \end{aligned}$$

où $q'(x_2,\dots,x_n)$ ne contient plus x_1. On en déduit

$$q(x) = c_{11}\Big(x_1+\frac{c_{12}}{c_{11}}x_2+\ldots+\frac{c_{1n}}{c_{11}}x_n\Big)^2 + \widetilde{q}(x_2,\dots,x_n).$$

Par hypothèse de récurrence pour la dimension $n-1$, on obtient que $\widetilde{q}(x_2,\dots,x_n)$ s'exprime comme une somme de carrés de formes linéaires indépendantes dans les variables $x_2,\dots,x_n$:

$$\widetilde{q}(x_2,\dots,x_n) = a_2\ell_2(x')^2 + \ldots + a_r\ell_r(x')^2, \quad x'=(x_2,\dots,x_n).$$

À cette somme, il faut encore ajouter le carré initialement trouvé

$$a_1\ell_1(x)^2, \quad a_1=c_{11}, \quad \ell_1(x) = x_1+\frac{c_{12}}{c_{11}}x_2+\ldots+\frac{c_{1n}}{c_{11}}x_n.$$

On voit que les formes linéaires $x\mapsto \ell_1(x)$, $\ell_2(x'),\dots,\ell_r(x')$ sont encore indépendantes dans $(\mathbb{K}^n)^*$ car $\ell_1(x)$ est la seule forme qui fasse intervenir x_1, donc $\lambda_1\ell_1+\lambda_2\ell_2+\ldots+\lambda_r\ell_r=0$ implique $\lambda_1=0$ et par suite $\lambda_2\ell_2+\ldots+\lambda_r\ell_r=0$, d'où aussi $\lambda_2=\ldots=\lambda_r=0$.

$(**)$ Si $q(x)$ ne contient aucun terme carré mais contient un "terme rectangle" non nul, par exemple $2c_{12}x_1x_2$, $c_{12}\neq 0$, on regroupe tous les termes contenant x_1 ou x_2, c'est-à-dire

$$\begin{aligned} &2c_{12}x_1x_2 + 2c_{13}x_1x_3 + \ldots + 2c_{1n}x_1x_n \quad + \quad 2c_{23}x_2x_3 + \ldots + 2c_{2n}x_2x_n \\ &\quad = 2c_{12}\Big(x_1x_2 + \frac{c_{13}}{c_{12}}x_1x_3 + \ldots + \frac{c_{1n}}{c_{12}}x_1x_n + \frac{c_{23}}{c_{12}}x_2x_3 + \ldots + \frac{c_{2n}}{c_{12}}x_2x_n\Big) \\ &\quad = 2c_{12}\Big(x_1+\frac{c_{23}}{c_{12}}x_3+\ldots+\frac{c_{2n}}{c_{12}}x_n\Big)\Big(x_2+\frac{c_{13}}{c_{12}}x_3+\ldots+\frac{c_{1n}}{c_{12}}x_n\Big) - q'(x_3,\dots,x_n). \end{aligned}$$

On en déduit

$$q(x) = 2c_{12}\Big(x_1+\frac{c_{23}}{c_{12}}x_3+\ldots+\frac{c_{2n}}{c_{12}}x_n\Big)\Big(x_2+\frac{c_{13}}{c_{12}}x_3+\ldots+\frac{c_{1n}}{c_{12}}x_n\Big) + \widetilde{q}(x_3,\dots,x_n).$$

Le produit $2c_{12}AB$ s'écrit comme $\frac{1}{2}c_{12}(A+B)^2 - \frac{1}{2}c_{12}(A-B)^2 = a_1\ell_1(x)^2 + a_2\ell_2(x)^2$ avec $a_1 = \frac{1}{2}c_{12}$, $a_2 = -\frac{1}{2}c_{12}$, $\ell_1(x) = x_1 + x_2 + (\ldots)$, $\ell_2 = x_1 - x_2 + (\ldots)$, tandis qu'on a par hypothèse de récurrence pour la dimension $n-2$ une décomposition en carrés de formes linéaires indépendantes

$$\widetilde{q}(x_3, \ldots, x_n) = a_3\ell_3(x'')^2 + \ldots + a_r\ell_r(x'')^2, \quad x'' = (x_3, \ldots, x_n).$$

Si $\lambda_1\ell_1 + \ldots + \lambda_r\ell_r = 0$, alors en regardant les termes qui contiennent x_1 et x_2, présents dans les seules formes ℓ_1 et ℓ_2, on voit que $\lambda_1 + \lambda_2 = 0$ et $\lambda_1 - \lambda_2 = 0$, donc $\lambda_1 = \lambda_2 = 0$, puis $\lambda_3 = \ldots = \lambda_r = 0$ par hypothèse de récurrence. Le théorème est démontré. En voici une conséquence importante.

Théorème 2.13. *Soit q une forme quadratique sur un $\mathbb{K}$-espace vectoriel E de dimension finie n. Alors il existe une base $(\widetilde{e}_1, \ldots, \widetilde{e}_n)$ dans laquelle q s'exprime en coordonnées comme*

$$q(x) = a_1\widetilde{x}_1^2 + \ldots + a_n\widetilde{x}_n^2, \quad a_j \in \mathbb{K},$$

de sorte que la forme polaire associée

$$\varphi(x, y) = a_1\widetilde{x}_1\widetilde{y}_1 + \ldots + a_n\widetilde{x}_n\widetilde{y}_n$$

admet dans cette base une matrice diagonale

$$\mathrm{Mat}_{(\widetilde{e}_j)}(\varphi) = \begin{pmatrix} a_1 & \ldots & 0 \\ \vdots & a_j & \vdots \\ 0 & \ldots & a_n \end{pmatrix}.$$

Démonstration. Supposons donnée une base $(e_1, \ldots, e_n)$ de E dans laquelle on exprime la matrice colonne $X = (x_j)$ des coordonnées d'un vecteur $x \in E$. On utilise une décomposition en carrés de formes linéaires indépendantes

$$q(x) = a_1\ell_1(x)^2 + \ldots + a_r\ell_r(x)^2, \quad a_j \in \mathbb{K},\ a_j \neq 0.$$

On sait qu'on peut compléter les formes indépendantes ℓ_j en une base $(\ell_1, \ldots, \ell_n)$ de E^* [de manière pratique, dans la méthode de Gauss, on se fatigue en général le moins possible en ajoutant juste les coordonnées $\ell_j(x) = x_{i_j}$ qui n'ont pas encore été choisies dans les étapes successives du processus de récurrence]. On complète aussi les coefficients a_j en posant $a_{r+1} = \ldots = a_n = 0$ si $r < n$. Le changement de coordonnées

$$\widetilde{x}_1 = \ell_1(x), \ldots, \widetilde{x}_n = \ell_n(x)$$

est alors bijectif, du type $\widetilde{X} = LX$ où les lignes de L sont les matrices lignes des formes ℓ_j. En calculant $P = L^{-1}$ on obtient la matrice de passage vers une base $(\widetilde{e}_j)$ dans laquelle les coordonnées sont précisément les $(\widetilde{x}_j)$. Le résultat s'ensuit. □

Formule de changement de base pour les formes bilinéaires. Soit $\varphi : E \times E \to \mathbb{K}$ une forme bilinéaire, $\mathcal{B} = (e_1, \ldots, e_n)$ une base de E (supposé de dimension finie n), et $C = \mathrm{Mat}_{\mathcal{B}}(\varphi)$ la matrice de φ dans $\mathcal{B}$. On effectue un changement de base

$$\mathcal{B} = (e_j) \mapsto \widetilde{\mathcal{B}} = (\widetilde{e}_j) \quad \text{donné par une matrice de passage } P.$$

Si X, Y (resp. $\widetilde{X}$, $\widetilde{Y}$) désignent les matrices colonnes de vecteurs x, y dans les bases respectives $\mathcal{B}$, $\widetilde{\mathcal{B}}$ nous avons

$$\varphi(x, y) = X^t C Y, \qquad X = P\widetilde{X}, \quad Y = P\widetilde{Y}.$$

Ceci donne

$$\varphi(x, y) = (\widetilde{X}^t P^t) C (P\widetilde{Y}) = \widetilde{X}^t (P^t C P)\widetilde{Y},$$

par conséquent la nouvelle matrice $\widetilde{C} = \mathrm{Mat}_{\widetilde{\mathcal{B}}}(\varphi)$ est donnée par la formule

$$\widetilde{C} = P^t C P.$$

On peut maintenant reformuler le Théorème 2.13 comme suit.

Théorème 2.14. *Soit q une forme quadratique sur un $\mathbb{K}$-espace vectoriel E de dimension finie n, de matrice C relativement à une base $(e_1, \dots, e_n)$ de E. Alors il existe une base $(\widetilde{e}_1, \dots, \widetilde{e}_n)$ donnée par une matrice de passage P telle que*

$$\widetilde{C} = P^t C P = \textit{matrice diagonale } \begin{pmatrix} a_1 & \dots & 0 \\ \vdots & a_j & \vdots \\ 0 & \dots & a_n \end{pmatrix}.$$

Pour cela, on cherche une décomposition en carré $q(x) = \sum_{j=1}^r a_j \ell_j(x)^2$ en somme de carrés de formes linaires indépendantes, on complète en une base $(\ell_1, \ell_2, \dots, \ell_n)$ de E^* et on calcule $P = L^{-1}$ où L est la matrice dont les lignes sont les matrices des formes linaires ℓ_j.

Remarque 2.15. Lorsque le corps de base est $\mathbb{K} = \mathbb{Q}$ on s'arrête en général à ce point. Lorsque $\mathbb{K} = \mathbb{R}$, on peut écrire $a_j = \varepsilon_j |a_j|$ avec $\varepsilon_j = \pm 1$, et donc

$$a_j \ell_j(x)^2 = \varepsilon_j \Big(\sqrt{|a_j|}\, \ell_j(x) \Big)^2, \quad 1 \leq j \leq r.$$

Les formes $\big(\sqrt{|a_j|}\big)_{1 \leq j \leq r}$ sont zncore indépendantes puisques $a_j \neq 0$, donc en remplaçant ℓ_j par $\widehat{\ell}_j = \sqrt{|a_j|}\ell_j$, $1 \leq j \leq r$ et en complétant en une base $(\widehat{\ell}_1, \dots \widehat{\ell}_n)$ de E^* on obtient maintenant une décomposition en carrés

$$q(x) = \varepsilon_1 \widehat{\ell}_1(x)^2 + \dots + \varepsilon_r \widehat{\ell}_r(x)^2,$$

ce qui donne une matrice diagonale de rang r

$$\widehat{C} = \begin{pmatrix} \varepsilon_1 & \dots & \dots & 0 \\ \vdots & \varepsilon_r & & \vdots \\ \vdots & & 0 & \vdots \\ 0 & \dots & \dots & 0 \end{pmatrix}$$

n'ayant comme coefficients diagonaux que $\varepsilon_j = \pm 1$ ou 0. Sur le corps $\mathbb{K} = \mathbb{C}$, les coefficients a_j admettent toujours des racines carrées $\sqrt{a_j}$ et on peut écrire

$$q(x) = \widehat{\ell}_1(x)^2 + \dots + \widehat{\ell}_r(x)^2, \qquad \widehat{\ell}_j(x) = \sqrt{a_j}\, \ell_j(x),$$

ce qui conduit à la matrice de rang r

$$\widehat{C} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & 1 & \vdots \\ 0 & \dots & 0 \end{pmatrix}.$$

On montre pour terminer que le rang r ne dépend pas de la décomposition en carrés (et donc du changement de base effectué pour diagonaliser). Pour cela on utilise le lemme suivant – en mathématiques, un **lemme** est un petit théorème préparatoire.

Lemme 2.16. *Soit $u : E \to F$, $v : F \to G$, $w : G \to H$ des applications linéaires. Si u et w sont des isomorphismes, alors*

$$\mathrm{rang}(w \circ v \circ u) = \mathrm{rang}(v).$$

De manière analogue, pour des matrices A, B, C multipliables, si A et C sont des matrices inversibles, on a

$$\operatorname{rang}(ABC) = \operatorname{rang}(B).$$

Démonstration. Il suffit de démontrer le résultat dans le cas des applications linéaires. Par définition $\operatorname{rang}(v) = \dim v(F)$ et de même

$$\operatorname{rang}(w \circ v \circ u) = \dim w \circ v \circ u(E).$$

Or, comme u est par hypothèse un isomorphisme, on a en particulier que u est surjective, donc $u(E) = F$. De même, l'hypothèse que w soit un isomorphisme implique que w est injective, donc la restriction $w : v(F) \to w(v(F)$ est injective. Comme cette restriction est évidemment surjective, on en déduit que $w : v(F) \to w(v(F)$ est un isomorphisme, donc

$$\dim w \circ v \circ u(E) = \dim w(v(F)) = \dim v(F) = \operatorname{rang}(v).$$

Le lemme en résulte. On observera que le résultat est vrai en fait plus généralement si u (resp. C) est surjective et w (resp. A) injective. □

Théorème 2.17. *Si φ est une forme quadratique et C sa matrice dans une base quelconque, alors $r = \operatorname{rang}(C)$ ne dépend pas de la base choisie pour calculer C. Le rang r est aussi le nombre de carrés apparaissant dans une décomposition quelconque en carrés de formes linéaires indépendantes.*

Démonstration. Si on effectue un changement de base donné par une matrice de passage P, la nouvelle matrice de φ devient $\widetilde{C} = P^t C P$. Comme P et P^t sont inversibles, le lemme implique bien que $\operatorname{rang}(\widetilde{C}) = \operatorname{rang}(C)$. D'autre part, trouver une décomposition en somme de carrés de formes linéaires indépendantes revient à rendre la matrice C diagonale, et dans ce cas le rang est bien le nombre de coefficients diagonaux non nuls. □

3. Orthogonalité par rapport à une forme bilinéaire symétrique

Soit $\varphi : E \times E \to \mathbb{K}$ une forme bilinéaire symétrique et $q(x) = \varphi(x, x)$ la forme quadratique associée.

Définition 3.1. *On dit que deux vecteurs $x, y \in E$ sont (φ-)orthogonaux lorsque $\varphi(x, y) = 0$, et on écrit alors $x \perp y$ (ou $x \perp_\varphi y$ s'il y a ambiguité).*

Définition 3.2. *Si $F \subset E$ est un sous-espace vectoriel de E, on note $F^\perp$ (ou au besoin $F^{\perp_\varphi}$) l'ensemble des vecteurs y de E qui sont orthogonaux à tous les vecteurs x de F, c'est-à-dire*

$$F^\perp = \big\{ y \in E \,;\ \forall x \in F,\ \varphi(x, y) = 0 \big\}.$$

On écrit aussi $y \perp F$ (ou $F \perp y$) pour exprimer que y est perpendiculaire à tous les éléments de F, c'est-à-dire que $y \in F^\perp$.

Il est immédiat de verifier en utilisant la bilinéarité de φ que $F^\perp$ est bien toujours un sous-espace vectoriel de E ; en effet, si $y_1, y_2 \in F^\perp$ et si $\lambda_1, \lambda_2 \in \mathbb{K}$, alors $\lambda_1 y_1 + \lambda_2 y_2 \in F^\perp$.

Remarque 3.3. Cette notion générale ne correspond pas nécessairement à l'intuition usuelle de ce que sont les vecteurs orthogonaux. Ainsi, si on choisit $\varphi = 0$ (forme bilinéaire nulle), tous les vecteurs sont orthogonaux entre eux, et on a donc $F^\perp = E$ pour tout sous-espace F ! Pour sortir du cas $\varphi = 0$ (qui n'est à vrai dire pas très intéressant...), on peut considérer l'espace $E = \mathbb{K}[X]$ des polynômes à coefficients dans $\mathbb{K}$ et la forme bilinéaire symétrique non nulle $\varphi(P, Q) = P(0)Q(0)$. Dans ce cas, on constatera que le polynome $P = X$ est orthogonal à tout polynôme Q, c'est-à-dire que $X \in E^\perp$. On vérifiera facilement ici que $E^\perp$ consiste en l'ensemble des polynômes P tels que $P(0) = 0$.

Calcul de l'orthogonal d'un sous-espace vectoriel F. Supposons que F soit de dimension finie p. On choisit alors une base $(a_1, \dots, a_p)$ de F. Nous affirmons que

$$F^\perp = \{y \in E\,;\ y \perp a_1, \dots, y \perp a_p\} = \{y \in E\,;\ \varphi(a_1, y) = \dots = \varphi(a_p, y) = 0\}.$$

En effet, si $y \in F^\perp$, il est clair que y est orthogonal à chacun des vecteurs $a_1, \dots, a_p \in F$. Réciproquement, si $y \perp a_j$, $1 \leq j \leq p$, on voit facilement que y est aussi orthogonal à tout vecteur $x = \lambda_1 a_1 + \dots + \lambda_p a_p \in F$, puisque

$$\varphi(a_j, y) = 0 \quad \Longrightarrow \quad \varphi(x, y) = \sum_{j=1}^{p} \lambda_j \varphi(a_j, y) = 0.$$

En pratique les conditions $\varphi(a_1, x) = \dots = \varphi(a_p, y) = 0$ se traduisent matriciellement par le système d'équations linéaires

$$(A_1)^t C Y = 0,\ \dots\ ,\ (A_p)^t C Y = 0$$

où $C = \mathrm{Mat}_{\mathcal{B}}(\varphi)$, $Y = \mathrm{Mat}_{\mathcal{B}}(y)$, $A_j = \mathrm{Mat}_{\mathcal{B}}(a_j)$. D'après ce qui précède, ces équations déterminent le sous-espace $F^\perp$.

Théorème et définition 3.4. *On note* $\operatorname{Ker}\varphi = E^\perp$ *l'ensemble des vecteurs* $y \in E$ *orthogonaux à tous les autres. Si* E *est de dimension finie, muni d'une base* $\mathcal{B}$*, et si* $C = \mathrm{Mat}_{\mathcal{B}}(\varphi)$*,* $\operatorname{Ker}\varphi$ *est l'ensemble des vecteurs* y *de* E *dont la matrice colonne* Y *dans la base* $\mathcal{B}$ *vérifie* $CY = 0$ (*autrement dit,* $\operatorname{Ker}\varphi$ *correspond en coordonnées au noyau de la matrice* C).

Démonstration. Remarquons d'abord que pour deux matrices colonnes $X = (x_j)$ et $Z = (z_j)$, on a $X^t Z = \sum_{j=1}^n x_j z_j$ et par suite $X^t Z = 0$ pour tout X si et seulement si $Z = 0$. Par conséquent

$$\varphi(x, y) = X^t C Y = X^t (CY) = 0 \quad \text{pour tout } x \in E \quad \Longleftrightarrow \quad Z = CY = 0.$$

C'est exactement l'affirmation du théorème. □

Définition 3.5. *On dit que la forme bilinéaire* φ *est* **dégénérée** *si* $\operatorname{Ker}\varphi \neq \{0\}$*, et* **non dégénérée** *si* $\operatorname{Ker}\varphi = \{0\}$.

En dimension finie, il suffit de calculer le déterminant $\det C$ de la matrice dans une base. Dire que φ est non dégénérée revient à dire $\det C \neq 0$ ou encore que $\operatorname{rang}\varphi = n = \dim E$, ou encore que le nombre de carrés intervenant dans une décomposition en carrés de formes linaires indépendantes est égal à la dimension de l'espace. Outre le noyau, une autre notion utile est celle de *vecteur isotrope*.

Définition 3.6. *On dit qu'un vecteur* x *est isotrope pour la forme quadratique* q (*ou pour la forme bilinéaire associée* φ) *si*

$$q(x) = \varphi(x, x) = 0$$

autrement dit si le vecteur x *est orthogonal à lui-même. On notera* $\mathrm{Isotrope}(q)$ (*ou parfois* $\mathrm{Isotrope}(\varphi)$) *l'ensemble des vecteurs isotropes de* E *relativement à* q.

La propriété $q(\lambda x) = \lambda^2 q(x)$ montre que tout multiple λx d'un vecteur isotrope est encore isotrope. Dans un espace vectoriel un sous-ensemble invariant par multiplication par les scalaires s'appelle un cône (de sommet 0). On parlera donc du **cône des vecteurs isotropes**.

Exemple 3.7. Prenons $E = \mathbb{R}^2$ et $q(x, y) = x^2 - y^2$. La forme bilinéaire associée est

$$\varphi((x, y), (x', y')) = xx' - yy', \qquad \text{et on a } \ C = \mathrm{Mat}_{\mathcal{B}}(\varphi) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

dans la base canonique $\mathcal{B}$ de $\mathbb{R}^2$. On a $\det C = -1 \neq 0$, donc φ est non dégénérée, $\operatorname{Ker}\varphi = \{0\}$. En revanche, l'ensemble des vecteurs isotropes $\operatorname{Isotrope}(q)$ est obtenu en écrivant $q(x,y) = x^2 - y^2 = 0$, c'est donc la réunion des deux diagonales $y = x$ et $y = -x$ dans le plan Oxy. On notera que $\operatorname{Isotrope}(q)$ n'est pas un sous-espace vectoriel.

Dans ce cas, il est facile de déterminer les orthogonaux des sous-espaces vectoriels F de $E = \mathbb{R}^2$. On a en effet $\dim_{\mathbb{R}} F = 0, 1$ ou 2. Si $\dim_{\mathbb{R}} F = 0$ alors $F = \{0\}$ et $F^\perp = E = \mathbb{R}^2$. Si $\dim_{\mathbb{R}} F = 2$, alors $F = E = \mathbb{R}^2$ et $F^\perp = E^\perp = \operatorname{Ker}\varphi = \{0\}$. Il reste à traiter le cas où F est une droite. Or, si F est la droite de vecteur directeur (a,b), on a

$$(x,y) \in F^\perp \iff \varphi((a,b),(x,y)) = ax - by = 0.$$

On voit donc que $F^\perp$ est la droite de vecteur directeur (b,a). On prendra garde au fait que l'on peut tout à fait avoir $F^\perp = F$! En fait, c'est le cas précisément si le vecteur (a,b) est isotrope, c'est-à-dire $b = \pm a$, ce qui correspond aux deux diagonales du plan Oxy.

Exemple 3.8. On prend ici $E = \mathbb{R}^4$ (l'espace-temps d'Einstein) et on considère la forme quadratique dite de Lorentz

$$q(x,y,z,t) = x^2 + y^2 + z^2 - c^2t^2$$

où $c > 0$ est la vitesse de la lumière. Il s'agit d'une forme quadratique *non dégénérée* (la matrice est diagonale à coefficients non nuls), c'est-à-dire que $\operatorname{Ker}\varphi = \{0\}$. L'ensemble des vecteurs isotropes est constitué au temps t des vecteurs de la sphère $x^2 + y^2 + z^2 = c^2t^2$ de rayon $R = c|t|$, autrement dit il s'agit d'une sphère qui grossit exactement à la vitesse de la lumière. Dans ce cas, le cône des vecteurs isotropes est souvent appelé le "cône de lumière". L'intérieur du cône de lumière, c'est-à-dire les vecteurs tels que $x^2+y^2+z^2 < c^2t^2$ peut-être interprété comme les points de l'espace-temps qui peuvent être reliés au big-bang (pris comme point origine) par des phénomènes se propageant à vitesse inférieure à celle de la lumière. Les points $x^2 + y^2 + z^2 > c^2t^2$ de l'espace-temps sont "au delà" de l'horizon accessible par un lien de causalité aux phénomènes issus du big-bang (si toutefois la vitesse limite est bien celle de la lumière...)

Remarque 3.9. Si φ est une forme bilinéaire symétrique, on a toujours

$$\operatorname{Ker}\varphi \subset \operatorname{Isotrope}(\varphi),$$

en effet $\operatorname{Ker}\varphi$ est constitué des vecteurs x tels que $\varphi(x,y) = 0$ pour tout y, on trouve donc en particulier $\varphi(x,x) = 0$ en prenant $y = x$. Les exemples précédents montrent que l'on n'a pas égalité en général.

Proposition 3.10. *Pour tout sous-espace vectoriel F de E, on a $(F^\perp)^\perp \supset F$.*

En effet, tout vecteur x de F est orthogonal à tout vecteur de $F^\perp$ par définition, donc $x \in (F^\perp)^\perp$. En revanche, l'égalité n'a pas toujours lieu, comme le montre l'exemple suivant.

Exemple 3.11. Prenons $E = \mathbb{R}^3 \ni (x,y,z)$ et $q(x,y,z) = x^2 - y^2$. La forme bilinéaire associée est

$$\varphi((x,y,z),(x',y',z')) = xx' - yy', \qquad \text{et on a } \; C = \operatorname{Mat}_{\mathcal{B}}(\varphi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

dans la base canonique $\mathcal{B}$ de $\mathbb{R}^3$. Le noyau (calculé en prenant le noyau de la matrice C) est donné par les équations $x = y = 0$, soit

$$\operatorname{Ker}\varphi = \mathbb{R}(0,0,1),$$

il s'agit donc d'une forme quadratique dégénérée. L'ensemble Isotrope(q) est constitué de la réunions des deux plans $x = y$ et $x = -y$. Calculons ici l'orthogonal de la droite

$$D = \mathbb{R}(1, 2, 1).$$

La condition d'orthogonalité $(x, y, z) \perp (1, 2, 1)$ s'écrit

$$\varphi((x, y, z), (1, 2, 1)) = x - 2y = 0 \quad \Leftrightarrow \quad x = 2y.$$

Il s'agit donc d'un plan P admettant pour base les vecteurs $(2, 1, 0)$ et $(0, 0, 1)$. Calculons maintenant $(D^\perp)^\perp = P^\perp$. L'orthogonal $P^\perp$ est l'ensembles des vecteurs (x, y, z) qui sont orthogonaux à la fois à $(2, 1, 0)$ et $(0, 0, 1)$, ce qui donne les équations $2x-y = 0$ et $0x+0y = 0$. La second est toujours vérifiée, et on voit que $(D^\perp)^\perp$ est en fait un plan P', à savoir le plan d'équation $2x-y = 0$. Ce plan P' contient la droite D, mais on a bien sûr $P' = ((D^\perp)^\perp \neq D$.

On va voir que cette difficulté ne se produit pas lorsque la forme bilinéaire φ est non dégénérérée. Rappelons un lemme utile d'algèbre linéaire.

Lemme 3.12. *Soient $\ell_1, \ldots, \ell_p \in E^*$ des formes linéaires indépendantes sur un $\mathbb{K}$-espace vectoriel de dimension finie E. alors le sous-espace*

$$S = \big\{x \in E\,;\ \ell_1(x) = \ldots = \ell_p(x) = 0\big\}$$

est de codimension p, c'est-à-dire que

$$\dim_{\mathbb{K}} S = \dim_{\mathbb{K}} E - p.$$

Démonstration. Complétons $(\ell_1, \ldots, \ell_p)$ en une base $(\ell_1, \ldots, \ell_n)$ de E^*, avec $n = \dim_{\mathbb{K}} E = \dim_{\mathbb{K}} E^*$. On peut effectuer le changement de coordonnées bijectif

$$\begin{cases} \widetilde{x}_1 = \ell_1(x), \\ \ldots \\ \widetilde{x}_n = \ell_n(x). \end{cases}$$

Soit L la matrice dont les lignes sont les coefficients des formes linéaires ℓ_j. Ce changement de coordonnées $\widetilde{X} = LX$ correspond à une base $(\widetilde{e}_1, \ldots, \widetilde{e}_n)$ donnée par la matrice de passage $P = L^{-1}$. Dans ces coordonnées, S est le sous-espace défini par les équations

$$\widetilde{x}_1 = \ldots = \widetilde{x}_p = 0,$$

c'est donc le sous-espace engendré par $(\widetilde{e}_{p+1}, \ldots, \widetilde{e}_n)$, et on voit que $\dim_{\mathbb{K}} S = n - p$. □

Théorème 3.13. *Soit φ une forme bilinéaire symétrique* non dégénérée *sur un espace vectoriel E de dimension finie. Alors pour tout sous-espace vectoriel F de E on a*

$$\dim_{\mathbb{K}} F^\perp = \dim_{\mathbb{K}} E - \dim_{\mathbb{K}} F.$$

Démonstration. Fixons une base $(b_1, \ldots, b_p)$ de F. Alors $F^\perp$ est défini par l'annulation des p formes linéaires

$$\ell_1(x) = \varphi(x, b_1), \ldots, \ell_p(x) = \varphi(x, b_p).$$

Montrons qu'elles sont indépendantes : si $\lambda_1, \ldots, \lambda_p \in \mathbb{K}$ sont des scalaires tels que $\lambda_1\ell_1 + \ldots + \lambda_p\ell_p = 0$, alors on a

$$0 = \lambda_1\ell_1(x) + \ldots + \lambda_p\ell_p(x) = \lambda_1\varphi(x, b_1) + \ldots + \lambda_p\varphi(x, b_p) = \varphi(x, \lambda_1 b_1 + \ldots + \lambda_p b_p)$$

pour tout $x \in E$, autrement dit $\lambda_1 b_1 + \ldots + \lambda_p b_p \in \operatorname{Ker}\varphi$. Or $\operatorname{Ker}\varphi = \{0\}$ par hypothèse, donc $\lambda_1 b_1 + \ldots + \lambda_p b_p = 0$ et par suite $\lambda_1 = \ldots = \lambda_p = 0$. Le lemme précédent montre que $F^\perp$ est de codimension p, c'est-à-dire que

$$\dim_{\mathbb{K}} F^\perp = \dim_{\mathbb{K}} E - p = \dim_{\mathbb{K}} E - \dim_{\mathbb{K}} F.$$ □

Notons que ceci s'applique en particulier au produit scalaire usuel sur $\mathbb{R}^n$. Une conséquence importante est la suivante.

Théorème 3.14. *Soit φ une forme bilinéaire symétrique* non dégénérée *sur un espace vectoriel E de dimension finie. Alors pour tout sous-espace vectoriel F de E on a*

$$(F^\perp)^\perp = F.$$

Démonstration. On sait déjà que $(F^\perp)^\perp \supset F$. De plus le théorème précédent nous dit que

$$\dim_{\mathbb{K}}(F^\perp)^\perp = \dim_{\mathbb{K}} E - \dim_{\mathbb{K}} F^\perp = \dim_{\mathbb{K}} E - (\dim_{\mathbb{K}} E - \dim_{\mathbb{K}} F) = \dim_{\mathbb{K}} F,$$

ce qui implique bien $(F^\perp)^\perp = F$. □

Attention ! Même si φ est non dégénérée, il n'est pas vrai en général que $E = F \oplus F^\perp$. Ainsi, si $q(x, y) = x^2 - y^2$, le vecteur $(1, 1)$ est isotrope et la droite $D = \mathbb{R}(1, 1)$ est orthogonale à elle même. On a en fait $D^\perp = D$, donc D, $D^\perp$ ne sont pas en somme directe.

Pour que $F \oplus F^\perp = E$ il faut et il suffit que $F \cap F^\perp = \{0\}$, en effet on a dans ce cas

$$\dim_{\mathbb{K}} F + F^\perp = \dim_{\mathbb{K}} F \oplus F^\perp = \dim_{\mathbb{K}} F + \dim_{\mathbb{K}} F^\perp = \dim E.$$

Définition 3.15. *Soit E un espace vectoriel muni d'une forme bilinéaire symétrique. On dit qu'une base $(b_1, \ldots, b_n)$ de E est*

- *orthogonale si $b_i \perp b_j$ pour $i \neq j$; ceci équivaut à dire que $\varphi(b_i, b_j) = 0$ pour $i \neq j$, ou encore que la matrice $\mathrm{Mat}_{(b_1,\ldots,b_n)}(\varphi)$ est diagonale.*
- *orthonormée si de plus $q(b_i) = \varphi(b_i, b_i) = +1$; ceci équivaut à dire que $\mathrm{Mat}_{(b_1,\ldots,b_n)}(\varphi) = I_n$ (matrice unité d'ordre n).*

On notera que d'après la méthode de Gauss, une forme bilinéaire symétrique φ quelconque admet toujours des bases orthogonales. Si φ est non dégénérée et si $\mathbb{K} = \mathbb{R}$, on peut se ramener à $q(b_j) = \varphi(b_j, b_j) = \pm 1$ (en divisant si nécessaire b_j par $\sqrt{|q(b_j)|}$) ; en revanche sur $\mathbb{C}$, on peut toujours se ramener dans ce cas à $q(b_j) = 1$:

Théorème 3.16. *Soit E de dimension n sur le corps $\mathbb{K}$. Si $\mathbb{K} = \mathbb{R}$, une forme bilinéaire φ non dégénérée admet toujours une base orthogonale $(b_1, ..., b_n)$ telle que $q(b_j) = \pm 1$. Si $\mathbb{K} = \mathbb{C}$, une forme bilinéaire φ non dégénérée admet toujours une base orthonormée $(b_1, ..., b_n)$.*

Formes quadratiques semi-positives et définies positives. Comme la notion de positivité implique l'utilisation d'une relation d'ordre, on supposera ici que $\mathbb{K} = \mathbb{R}$ (bien que les notions fassent sens aussi lorsque $\mathbb{K} = \mathbb{Q}$).

Définition 3.17. *On dit qu'une forme quadratique q (ou la forme bilinéaire symétrique φ associée) est* semi-positive *sur l'espace E si pour tout $x \in E$ on a*

$$q(x) = \varphi(x, x) \geq 0.$$

On peut alors associer à q la semi-norme

$$\|x\| = \sqrt{q(x)} = \sqrt{\varphi(x, x)} \in \mathbb{R}_+.$$

On notera alors que pour tout scalaire $\lambda \in \mathbb{R}$

$$\|\lambda x\| = \sqrt{\lambda^2 q(x)} = |\lambda|\, \|x\|.$$

D'autre part, on a par définition

$$\mathrm{Isotrope}(q) = \big\{x \in E\,;\ q(x) = 0\big\} = \big\{x \in E\,;\ \|x\| = 0\big\}.$$

La notion de forme quadratique semi-négative est définie de même (mais on s'abstient alors d'introduire la norme).

Définition 3.18. *On dit que q est* définie *si q ne possède pas de vecteur isotrope $x \neq 0$.*

On forme quadratique définie est nécessairement non dégénérée (puisque tout vecteur du noyau est isotrope).

Exemples 3.19. • Sur $\mathbb{R}^n$, la forme quadratique $q(x_1 \ldots, x_n) = -(x_1^2 + \ldots + x_n^2)$ est définie.

• Sur $\mathbb{R}^3$, la forme quadratique $q(x, y, z) = x^2 + (y+z)^2$ est semi-positive. Le cône Isotrope(q) est constitué des vecteurs $(x, y, z) \in \mathbb{R}^3$ tels que $x = 0$ et $y + z = 0$, c'est-à-dire la droite vectorielle $\mathbb{R}(0, 1, -1)$. La forme q est non définie.

• Sur $\mathbb{Q}^2$, la forme quadratique $q(x, y) = x^2 - 2y^2$ est définie. En effet $q(x, y) = 0$ équivaut à $x = \pm\sqrt{2}\, y$, ce qui est impossible avec x, y rationnels non nuls, puisque $\sqrt{2}$ est irrationnel. Bien que q soit définie, on remarquera qu'elle n'est ni semi-positive ni semi-négative.

• sur le corps $\mathbb{K} = \mathbb{R}$, une forme quadratique définie est soit définie positive, soit définie négative : en effet, elle doit être non dégénérée, et si elle admet après changement de base une décomposition en carrés de formes linéaires

$$q(x_1, \ldots, x_n) = \varepsilon_1 \widetilde{x}_1^2 + \ldots + \varepsilon_n \widetilde{x}_n^2$$

avec $\varepsilon_j = \pm 1$, alors tous les signes doivent être les mêmes, sinon si $\varepsilon_j = 1 \neq \varepsilon_k = -1$, on obtient un vecteur isotrope non nul en prenant $x_j = x_k = 1$ et les autres coordonnées nulles.

Conformément aux définitions qui précèdent, on utilise la terminologie suivante.

Définition 3.20. On dit qu'une forme quadratique q sur un expace vectoriel réel E est

• définie positive si $q(x) > 0$ pour tout vecteur $x \in E$, $x \neq 0$,

• définie négative si $q(x) < 0$ pour tout vecteur $x \in E$, $x \neq 0$.

Si q est définie > 0, la semi-norme $\|x\| = \sqrt{q(x)}$ est une vraie norme, c'est-à-dire que

$$\|x\| = 0 \iff q(x) = 0 \iff x = 0$$

Si q est seulement semi-positive, on a bien sûr

$$\|x\| = 0 \iff x \in \text{Isotrope}(q).$$

Définition 3.21. *On appelle* espace euclidien *un espace vectoriel réel muni d'une forme quadratique q définie positive, et de sa forme bilinéaire symétrique associée φ. On la note en général $(x, y) \mapsto \langle x, y\rangle = \varphi(x, y)$, et on appelle $\langle x, y\rangle$ le* produit scalaire.

Exemple 3.22. Soit $E = C^0([a, b], \mathbb{R})$ l'ensemble des fonctions continues sur $[a, b]$ muni de la forme bilinéaire symétrique

$$\varphi(f, g) = \langle f, g\rangle = \int_a^b f(x)\, g(x)\, dx.$$

On vérifie facilement que (E, φ) est un espace euclidien (de dimension infinie).

Inégalité de Cauchy-Schwarz. C'est une inégalité fondamentale qui majore le produit scalaire en fonction de la (semi)-norme associée.

Théorème 3.23 (Inégalité de Cauchy-Schwarz)**.** *Soit q une forme quadratique semi-positive sur un $\mathbb{R}$-espace vectoriel E et φ la forme bilinéaire symétrique associée. Alors pour tous vecteurs $x, y \in E$ on a*

$$|\varphi(x, y)| \leq \|x\|\, \|y\| = \sqrt{q(x)}\, \sqrt{q(y)}.$$

Dans un espace euclidien, on écrira cette inégalité sous la forme

$$\big|\langle x, y\rangle\big| \leq \|x\|\,\|y\|.$$

Démonstration. On considère le polynôme de degré ≤ 2 en la variable $\lambda \in \mathbb{R}$

$$P(\lambda) = q(\lambda x + y) = \varphi(\lambda x + y, \lambda x + y) = q(x)\lambda^2 + 2\varphi(x, y)\lambda + q(y).$$

D'après l'hypothèse de semi-positivité de q, nous avons $P(\lambda) \geq 0$ pour tout $\lambda \in \mathbb{R}$.

Si $q(x) = 0$ (c'est-à-dire si x est isotrope), nous avons $P(\lambda) = 2\varphi(x, y)\lambda + q(y)$ et cette fonction affine ne peut-être positive pour tout $\lambda \in \mathbb{R}$ que si elle est constante, c'est-à-dire $\varphi(x, y) = 0$. L'inégalité est donc bien vraie dans ce cas ($0 \leq 0$).

Si $q(x) > 0$, il s'agit d'un vrai trinôme $P(\lambda)$ du second degré, et son discriminant Δ est nécessairement ≤ 0, sinon $\Delta > 0$ entraînerait l'existence de deux racines réelles $\lambda_1 < \lambda_2$ et $P(\lambda) = q(x)(\lambda - \lambda_1)(\lambda - \lambda_2)$ serait strictement négatif sur $]\lambda_1, \lambda_2[$. On en déduit

$$\Delta = 4\varphi(x, y)^2 - 4q(x)q(y) \leq 0 \quad \Longrightarrow \quad \varphi(x, y)^2 \leq q(x)q(y).$$

L'inégalité de Cauchy-Schwarz s'en déduit en prenant les racines carrées. On peut également vérifier l'inégalité directement en développant le carré de la norme de $\lambda x + \mu y$:

$$0 \leq q(\lambda x + \mu y) = \lambda^2 q(x) + 2\lambda\mu\,\varphi(x, y) + \mu^2 q(y),$$

soit

$$0 \leq \|\lambda x + \mu y\|^2 = \lambda^2\|x\|^2 + 2\lambda\mu\,\varphi(x, y) + \mu^2\|y\|^2.$$

Si l'on prend $\lambda = \|y\|$ et $\mu = \pm\|x\|$, on trouve

$$0 \leq \|x\|^2\|y\|^2 \pm 2\|x\|\,\|y\|\,\varphi(x, y) + \|x\|^2\|y\|^2 = 2\|x\|\,\|y\|\Big(\|x\|\,\|y\| \pm \varphi(x, y)\Big).$$

Si $\|x\| \neq 0$ et $\|y\| \neq 0$ on peut diviser par $\|x\|\,\|y\|$, ce qui donne $\|x\|\,\|y\| \pm \varphi(x, y) \geq 0$. Si $\|x\| = 0$, le choix $\lambda = \pm\varepsilon$, $\mu = 1$ donne $|\varphi(x, y)| \leq (\varepsilon/2)\,\|y\|^2$ pour tout $\varepsilon > 0$, donc $|\varphi(x, y)| = 0$, et on raisonne de même si $\|y\| = 0$ en prenant $\lambda = 1$ et $\mu = \pm\varepsilon$. □

Corollaire 3.24. *Si q est une forme quadratique semi-positive, on a* $\mathrm{Ker}(q) = \mathrm{Isotrope}(q)$.

En effet on sait déjà que $\mathrm{Ker}(q) \subset \mathrm{Isotrope}(q)$. Inversement, si x est isotrope, on a $\|x\| = \sqrt{q(x)} = 0$, donc $\varphi(x, y) = 0$ pour tout $y \in E$ par Cauchy-Schwarz, ce qui signifie que $x \in \mathrm{Ker}(q)$, et donc $\mathrm{Isotrope}(q) \subset \mathrm{Ker}(q)$.

Discutons maintenant le cas d'égalité, en supposant q définie positive. Le deuxième raisonnement direct montre que l'inégalité de Cauchy-Schwarz est une égalité lorsque

$$q(\|y\|x \pm \|x\|y) = 0,$$

ce qui implique $\|y\|x \pm \|x\|y = 0$ et donc x, y colinéaires si $\|x\| \neq 0$ ou $\|y\| \neq 0$ (mais si l'un des vecteurs x, y est nul, ils forment de toute manière une famille liée). Réciproquement, si x, y sont colinéaires avec par exemple $y = \lambda x$, $\lambda \in \mathbb{R}$, on a

$$\|x\|\,\|y\| = |\lambda|\,\|x\|^2, \qquad \varphi(x, y) = \lambda\,\varphi(x, x) = \lambda\|x\|^2 = \pm|\lambda|\,\|x\|^2 = \pm\|x\|\,\|y\|$$

suivant le signe de λ. On peut donc énoncer :

Théorème 3.25 (Cas d'égalité de l'inégalité de Cauchy-Schwarz)**.** *Si q est une forme quadratique définie positive de forme polaire associée φ, l'égalité $\varphi(x, y) = +\|x\|\,\|y\|$ (resp. $\varphi(x, y) = -\|x\|\,\|y\|$) se produit si et seulement si x, y sont* colinéaires de même sens (*resp.* colinéaires de sens opposés.)

Produit scalaire et angles non orientés. Si E est un espace euclidien de dimension finie n sur $\mathbb{R}$, on sait qu'il existe toujours une base orthonormée $(e_1, \dots, e_n)$, de sorte que $\langle e_i, e_j \rangle = \delta_{ij}$. Dans ce cas l'application

$$\mathbb{R}^n \to E, \qquad (x_1, \dots, x_n) \mapsto x = x_1 e_1 + \dots + x_n e_n$$

définit un isomorphisme $E \simeq \mathbb{R}^n$ tel que

$$\|x\|^2 = x_1^2 + \dots + x_n^2, \qquad \langle x, y \rangle = x_1 y_1 + \dots + x_n y_n$$

pour tous vecteurs x, y de E de coordonnées $(x_1, \dots, x_n)$ et $(y_1, \dots, y_n)$. Une observation importante (même si elle est assez évidente !) est que les coordonnées x_j se calculent comme des produits scalaires :

$$x_j = \langle e_j, x \rangle, \qquad 1 \leq j \leq n.$$

D'autre part, pour deux vecteurs x, $y \neq 0$, le rapport $\frac{\langle x,y \rangle}{\|x\| \, \|y\|} \in [-1, 1]$ ne change pas lorsqu'on multiplie x ou y par un scalaire $\lambda > 0$. Par analogie avec le cas usuel du plan euclidien, on définit l'angle non orienté $\theta = \widehat{(x, y)}$ par

$$\cos \theta = \frac{\langle x, y \rangle}{\|x\| \, \|y\|} \, .$$

Cette relation donne un angle $\theta \in [0, \pi]$ unique. On obtient alors la formule classique

$$\langle x, y \rangle = \|x\| \, \|y\| \, \cos \theta.$$

Procédé d'orthonormalisation de Gram-Schmidt. Si F est un sous-espace de dimension finie p d'un espace euclidien (E, φ), la restriction $\varphi_{|F \times F}$ du produit scalaire à F en fait un espace euclidien $(F, \varphi_{|F \times F})$. Étant donné une base $(a_1, \dots, a_p)$ de F, le but du procédé d'orthonormalisation de Gram-Schmidt est de fabriquer une base orthonormée $(b_1, \dots, b_p)$ de F. Notons $F_j = \mathrm{Vect}(a_1, \dots, a_j)$ l'espace vectoriel engendré par les j premiers vecteurs de base, de sorte que $\dim_{\mathbb{R}} F_j = j$ et $F_p = F$. On calcule les vecteurs b_j par récurrence sur j, en commençant par

$$b_1 = \frac{1}{\|a_1\|} a_1,$$

et on procède en sorte qu'à chaque étape j on ait

$$F_j = \mathrm{Vect}(a_1, \dots, a_j) = \mathrm{Vect}(b_1, \dots, b_j) \quad \text{et donc} \quad F_j = F_{j-1} \oplus \mathbb{R} a_j = F_{j-1} \oplus \mathbb{R} b_j.$$

Supposons $b_1, \dots, b_{j-1}$ déjà calculés. Le point est que le vecteur a_j n'est pas nécessairement orthogonal à $b_1, \dots, b_{j-1}$, on le "redresse" donc en posant

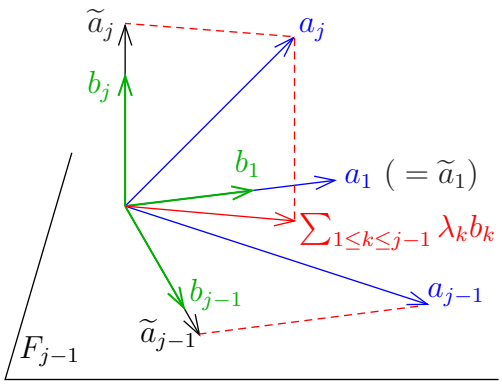
$$\widetilde{a}_j = a_j - (\lambda_1 b_1 + \dots + \lambda_{j-1} b_{j-1}), \qquad \lambda_k \in \mathbb{R}.$$

La condition d'orthogonalité $\widetilde{a}_j \perp b_k$ donne

$$\langle b_k, \widetilde{a}_j \rangle = \langle b_k, a_j \rangle - \lambda_k = 0 \qquad \text{d'où} \quad \lambda_k = \langle b_k, a_j \rangle, \quad 1 \leq k \leq j-1.$$

Ceci conduit à la formule

$$(*) \qquad \widetilde{a}_j = a_j - \sum_{k=1}^{j-1} \langle b_k, a_j \rangle b_k.$$



Le vecteur $\widetilde{a}_j$ ne peut être nul, sinon a_j serait dans le sous-espace $F_{j-1} = \mathrm{Vect}(b_1, \dots, b_{j-1}) = \mathrm{Vect}(a_1, \dots, a_{j-1})$ ce qui est absurde, et on peut donc rendre ce vecteur de norme 1 en posant

$$(**) \qquad b_j = \frac{1}{\|\widetilde{a}_j\|}\widetilde{a}_j = \frac{1}{\sqrt{\langle \widetilde{a}_j, \widetilde{a}_j\rangle}}\,\widetilde{a}_j.$$

Les formules $(*)$ et $(**)$ permettent le calcul des vecteurs $\widetilde{a}_j$ et b_j par récurrence (avec $\widetilde{a}_1 = a_1$). On observera que l'on a bien par construction

$$\mathrm{Vect}(b_1, \dots, b_j) = F_{j-1} + \mathbb{R}b_j = F_{j-1} + \mathbb{R}\widetilde{a}_j = F_{j-1} + \mathbb{R}a_j = F_j.$$

du fait que b_j est colinéaire à $\widetilde{a}_j$ et que $\widetilde{a}_j$ diffère de a_j par un vecteur de F_{j-1}. En effectuant le calcul de $(*)$ et $(**)$ pour tous $j = 1, 2, \dots, p$, on obtient bien la base orthonormée $(b_1, \dots, b_p)$ voulue.

Remarque 3.26. On peut combiner les formules $(*)$ et $(**)$ pour obtenir une formule de récurrence directe pour les vecteurs $\widetilde{a}_j$: on prend $\widetilde{a}_1 = a_1$, puis

$$(***) \qquad \widetilde{a}_j = a_j - \sum_{k=1}^{j-1} \frac{\langle \widetilde{a}_k, a_j\rangle}{\langle \widetilde{a}_k, \widetilde{a}_k\rangle}\,\widetilde{a}_k.$$

L'intérêt de cette formule (entièrement équivalente aux précédentes) est qu'elle évite tout calcul de racines carrées. On obtient ainsi une base orthogonale $(\widetilde{a}_1, \dots, \widetilde{a}_p)$ de F. On obtient ensuite par la formule de renormalisation $(**)$ une base orthonormée $(b_1, \dots, b_p)$.

Exemple 3.27. On identifie ici $\mathbb{R}^3$ à l'espace des matrices colonnes 3×1, que l'on munit de son produit scalaire canonique, et on considère la base

$$a_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad a_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad a_3 = \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}.$$

Appliquons le procédé de Gram-Schmidt à cette base afin d'obtenir une base orthonormée (b_1, b_2, b_3). On trouve

$$\widetilde{a}_1 = a_1, \qquad b_1 = \frac{1}{\|a_1\|}a_1 = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix},$$

puis

$$\widetilde{a}_2 = a_2 - \langle b_1, a_2\rangle b_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - \Big(\frac{1}{\sqrt{2}}\Big)^2 \times 1 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}, \qquad b_2 = \frac{1}{\sqrt{6}}\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$$

(le facteur $\frac{1}{2}$ se simplifie dans le calcul de b_2), et enfin

$$\widetilde{a}_3 = a_3 - \langle b_1, a_3\rangle b_1 - \langle b_2, a_3\rangle b_2 = \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix} - \Big(\frac{1}{\sqrt{2}}\Big)^2 \times 1 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - \Big(\frac{1}{\sqrt{6}}\Big)^2 \times (-5) \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}.$$

On notera (pour éviter des calculs fastidieux) que les racines carrées prises pour calculer les normes sont toujours élevées au carré dans les lignes de calculs ultérieures ; le facteur (-5) qui intervient par exemple dans l'évaluation du terme $\langle b_2, a_3\rangle b_2$ n'est autre que le produit scalaire du premier vecteur $a_3 = \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}$ par le vecteur $\begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$ qui intervient dans b_2. On

obtient finalement

$$\widetilde{a}_3 = \begin{pmatrix} 0 - \frac{1}{2} + \frac{5}{6} \\ 1 - \frac{1}{2} - \frac{5}{6} \\ -2 - 0 + \frac{5}{3} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ -\frac{1}{3} \\ -\frac{1}{3} \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}, \qquad b_3 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}.$$

Exemple 3.28. L'application

$$\varphi : \mathbb{R}[X]_2 \times \mathbb{R}[X]_2 \to \mathbb{R}, \qquad (P, Q) \mapsto \int_{-1}^{1} P(t)Q(t)\, dt$$

est bilinéaire symétrique. De plus, comme l'intégrale sur $[-1, 1]$ d'un monôme de degré impair est nulle, on voit que $1 \perp_\varphi X$ et $X \perp_\varphi X^2$. En revanche, 1 et X^2 ne sont pas φ-orthogonaux, puisque l'on a $\varphi(1, X^2) = \dfrac{2}{3}$; la base standard $(P_0, P_1, P_2) = (1, X, X^2)$ n'est donc pas φ-orthogonale. Cependant, comme $P_0 \perp P_1$, la méthode de Gram-Schmidt appliquée à (P_0, P_1, P_2) amène à une base φ-orthogonale $(\widetilde{P}_0, \widetilde{P}_1, \widetilde{P}_2)$ avec

$$\widetilde{P}_0 = P_0 = 1, \quad \widetilde{P}_1 = P_1 = X, \quad \widetilde{P}_2 = P_2 - \lambda P_0 = X^2 - \lambda,$$

et comme $\int_{-1}^{1} \widetilde{P}_0(t)\widetilde{P}_2(t)\, dt = \int_{-1}^{1} (t^2 - \lambda)\, dt = \frac{2}{3} - 2\lambda$, on voit qu'il faut prendre $\lambda = \frac{1}{3}$ pour avoir aussi $\widetilde{P}_0 \perp_\varphi \widetilde{P}_2$. Une calcul de normes fournit la base orthonormée $Q_i = \|\widetilde{P}_i\|^{-1}\widetilde{P}_i$ correspondante :

$$\|1\|^2 = 2, \qquad \|X\|^2 = \frac{2}{3}, \qquad \|X^2 - 1/3\|^2 = \frac{8}{45},$$

d'où

$$Q_0 = \frac{1}{\sqrt{2}}, \qquad Q_1 = \frac{\sqrt{3}}{\sqrt{2}}\, X, \qquad Q_2 = \frac{\sqrt{45}}{\sqrt{8}}\, (X^2 - 1/3).$$

Calcul de la projection orthogonale sur un sous-espace F de E. La détermination d'une base orthonormée $(b_1, \dots, b_p)$ de F par la méthode de Gram-Schmidt permet aussi d'obtenir l'expression en coordonnées de la projection orthogonale sur F. On notera que dans un espace euclidien E on a toujours $F \cap F^\perp = \{0\}$, car un vecteur $x \in F \cap F^\perp$ est orthogonal à lui-même, ce qui n'est possible par hypothèse que si $x = 0$. Pour tout $x \in E$ on peut écrire

$$x = x' + x'' \qquad \text{avec} \quad x' = \sum_{j=1}^{p} \langle b_j, x\rangle\, b_j \in F, \quad x'' = x - x' = x - \sum_{j=1}^{p} \langle b_j, x\rangle\, b_j \in F^\perp$$

(en effet il est facile de constater que $\langle b_j, x''\rangle = 0$ de sorte que $x'' \perp F$). On voit donc que

$$E = F \oplus F^\perp,$$

ceci étant valable même si E est de dimension infinie, à condition que F soit de dimension finie. Les projections orthogonales $\pi_F(x)$ et $\pi_{F^\perp}(x)$ sur F et $F^\perp$ sont données par

$$\pi_F(x) = \sum_{j=1}^{p} \langle b_j, x\rangle\, b_j \quad \in F, \qquad \pi_{F^\perp}(x) = x - \sum_{j=1}^{p} \langle b_j, x\rangle\, b_j \quad \in F^\perp.$$

Remarque 3.29. On observera que le procédé d'orthogonalisation de Gram-Schmidt revient à prendre $\widetilde{a}_j = a_j - \pi_{F_{j-1}}(a_j)$ où $\pi_{F_{j-1}}(a_j)$ est la projection orthogonale de a_j sur F_{j-1}. En termes des vecteurs $\widetilde{a}_j$, les projections s'expriment (sans racines carrées) par

$$\pi_F(x) = \sum_{j=1}^{p} \frac{\langle \widetilde{a}_j, x\rangle}{\langle \widetilde{a}_j, \widetilde{a}_j\rangle}\, \widetilde{a}_j \quad \in F, \qquad \pi_{F^\perp}(x) = x - \sum_{j=1}^{p} \frac{\langle \widetilde{a}_j, x\rangle}{\langle \widetilde{a}_j, \widetilde{a}_j\rangle}\, \widetilde{a}_j \quad \in F^\perp.$$

La symétrie orthogonale par rapport à F est donnée quant à elle par

$$\sigma_F(x) = x' - x'' = 2x' - (x' + x'') = 2x' - x = 2\pi_F(x) - x$$

(soit encore $\sigma_F = 2\pi_F - \mathrm{Id}_E$), ce qui implique la formule

$$\sigma_F(x) = 2\Big(\sum_{j=1}^{p} \langle b_j, x\rangle\, b_j\Big) - x = 2\Big(\sum_{j=1}^{p} \frac{\langle \widetilde{a}_j, x\rangle}{\langle \widetilde{a}_j, \widetilde{a}_j\rangle}\Big) - x.$$

Signature et théorème d'inertie de Sylvester. Revenons maintenant à l'étude des formes quadratiques réelles de signe quelconque. Soit E un espace vectoriel réel de dimension n et q une forme quadratique de rang r sur E. On sait qu'on peut trouver une base orthogonale $\mathcal{B} = (b_1, \dots, b_n)$ pour q de sorte que

$$\mathrm{Mat}_{\mathcal{B}}(q) = \begin{pmatrix} a_1 & \dots & \dots & 0 \\ \vdots & a_r & & \vdots \\ \vdots & & 0 & \vdots \\ 0 & \dots & \dots & 0 \end{pmatrix}, \qquad a_j = q(b_j).$$

Le noyau de q est $\mathrm{Ker}(q) = \mathrm{Vect}(b_{r+1}, \dots, b_n)$, sa dimension est $\dim_{\mathbb{R}} \ker(q) = n - r$. On introduit les sous-espaces

$$F_+ = \mathrm{Vect}\Big(b_j\,;\ 1 \le j \le r,\ a_j = q(b_j) > 0\Big), \quad F_- = \mathrm{Vect}\Big(b_j\,;\ 1 \le j \le r,\ a_j = q(b_j) < 0\Big),$$

et on note

$$r_+ = \dim_{\mathbb{R}} F_+ = (\text{nombre de coefficients diagonaux } a_j > 0),$$
$$r_- = \dim_{\mathbb{R}} F_- = (\text{nombre de coefficients diagonaux } a_j < 0).$$

Nous avons par construction une somme directe

$$E = F_+ \oplus F_- \oplus \mathrm{Ker}(q), \qquad r_+ + r_- = r = n - \dim_{\mathbb{R}} \mathrm{Ker}(q)$$

avec q définie > 0 sur F_+, q définie < 0 sur F_- (et bien sûr $q = 0$ sur $\mathrm{Ker}(q)$). Il n'est pas évident à priori que les dimensions r_+ et r_- soient indépendantes de la base orthogonale choisie. C'est ce qu'affirme le théorème suivant.

Théorème 3.30 (Théorème d'inertie de Sylvester et définition de la signature)**.** *Soit E un $\mathbb{R}$-espace vectoriel de dimension finie n, et soit $q : E \to \mathbb{R}$ une forme quadratique. Soit $\mathcal{B} = (b_j)$ une base q-orthogonale et*

$$r_+ = \mathrm{card}\{j\,;\ q(b_j) > 0\} \quad \textit{et} \quad r_- = \mathrm{card}\{j\,;\ q(b_j) < 0\}.$$

Alors le couple (r_+, r_-) ne dépend pas de la base q-orthogonale choisie $\mathcal{B}$. De plus

$$r_+ + r_- = r = \mathrm{rang}(q) = n - \dim_{\mathbb{R}} \mathrm{Ker}(q).$$

On dit que le couple (r_+, r_-) est la signature *de q.*

Démonstration. Soit $\widetilde{\mathcal{B}} = (\widetilde{b}_j)$ une autre base q-orthogonale et soit

$$\widetilde{r}_+ = \mathrm{card}\{j\,;\ q(\widetilde{b}_j) > 0\} \ \text{ et } \ \widetilde{r}_- = \mathrm{card}\{j\,;\ q(\widetilde{b}_j) < 0\}.$$

Comme expliqué ci-dessus, nous avons deux décompositions en somme directe

$$E = F_+ \oplus F_- \oplus \mathrm{Ker}(q) = \widetilde{F}_+ \oplus \widetilde{F}_- \oplus \mathrm{Ker}(q).$$

Ceci donne en particulier pour les dimensions respectives

$$r_+ + r_- = \widetilde{r}_+ + \widetilde{r}_- = r, \qquad \dim_{\mathbb{R}} \mathrm{Ker}(q) = n - r.$$

Par construction $q > 0$ sur $F_+ \smallsetminus \{0\}$, tandis que $q \leq 0$ sur $\widetilde{F}_- \oplus \mathrm{Ker}(q)$. Les deux sous-espaces vectoriels F_+ et $\widetilde{F}_- \oplus \mathrm{Ker}(q)$ ne peuvent donc se rencontrer qu'en 0, ce qui fournit une somme directe

$$S = F_+ \oplus \widetilde{F}_- \oplus \mathrm{Ker}(q) \subset E.$$

En prenant les dimensions, on en déduit

$$\dim_{\mathbb{R}} S = r_+ + \widetilde{r}_- + (n - r) \leq n \quad \Longrightarrow \quad \widetilde{r}_- \leq r - r_+ = r_- \quad \Longrightarrow \quad \widetilde{r}_+ \geq r_+.$$

En échangeant les rôles de $\mathcal{B}$ et $\widetilde{\mathcal{B}}$ on obtient de même $r_+ \geq \widetilde{r}_+$ et $r_- \leq \widetilde{r}_-$, par suite $r_+ = \widetilde{r}_+$ et $r_- = \widetilde{r}_-$. □

Remarque 3.31. Pour calculer la signature d'une forme quadratique q, il suffit d'utiliser l'algorithme de Gauss pour écrire $q(x)$ sous la forme d'une somme de carrés

$$q(x) = a_1 \ell_1(x)^2 + \ldots + a_r \ell_r(x)^2$$

de formes linéaires indépendantes, et de compter les nombres r_+, r_- de coefficients a_j qui sont respectivement > 0 et < 0. Ainsi dans l'exemple 2.11, la signature est $(1, 2)$, dans 2.12 la signature est $(2, 2)$ et dans 3.11 la signature est $(1, 1)$. Un espace euclidien (E, q) de dimension n est de signature $(n, 0)$. En général, la forme quadratique est non dégénérée si et seulement si $r_+ + r_- = n = \dim_{\mathbb{R}} E$, et dégénérée si et seulement si $r_+ + r_- < n$.

4. Formes sesquilinéaires

Le but de la théorie des formes sesquilinéaires est principalement de généraliser le produit scalaire, la norme et l'orthogonalité au cas des espaces vectoriels complexes. Une grande partie de la théorie est très similaire à celle des formes bilinéaires symétriques. Nous nous attacherons donc surtout à expliquer les différences.

Considérons l'espace vectoriel complexe $E = \mathbb{C}^n$ muni de sa base canonique. On note

$$z = (z_1, \ldots, z_n) \in \mathbb{C}^n \quad \text{avec} \quad z_j = x_j + iy_j \in \mathbb{C}, \quad x_j, y_j \in \mathbb{R}.$$

La "norme canonique" de $\mathbb{C}^n$ est donnée par

$$\|z\|^2 = |z_1|^2 + \ldots + |z_n|^2, \qquad |z_j|^2 = x_j^2 + y_j^2,$$

soit encore $\|z\| = \sqrt{|z_1|^2 + \ldots + |z_n|^2}$. Pour un autre vecteur $w = (w_1, \ldots, w_n) \in \mathbb{C}^n$, nous sommes amenés à poser

$$\langle z, w \rangle = \overline{z}_1 w_1 + \ldots + \overline{z}_n w_n \in \mathbb{C},$$

ce qui donne alors l'expression attendue pour la norme :

$$\langle z, z \rangle = \|z\|^2.$$

Si λ, μ sont des scalaires complexes, on voit que

$$\langle \lambda z, w \rangle = \overline{\lambda} \langle z, w \rangle, \qquad \langle z, \mu w \rangle = \mu \langle z, w \rangle.$$

L'application $w \mapsto \langle z, w \rangle$ est donc $\mathbb{C}$-linéaire, mais $z \mapsto \langle z, w \rangle$ n'est pas $\mathbb{C}$-linéaire, elle est seulement *anti-linéaire* :

Définition 4.1. *On dit qu'une forme $\varphi : E \times E \to \mathbb{C}$ sur un espace vectoriel complexe E est* une forme sesquilinéaire *si*

• *$x \mapsto \varphi(x, y)$ est anti-linéaire pour $y \in E$ fixé, c'est-à-dire*

$$\varphi(\lambda x, y) = \overline{\lambda}\, \varphi(x, y), \qquad \varphi(x_1 + x_2, y) = \varphi(x_1, y) + \varphi(x_2, y)$$

pour tout $\lambda \in \mathbb{C}$ et tous $x, x_1, x_2, y \in E$,

• $y \mapsto \varphi(x,y)$ *est* $\mathbb{C}$*-linéaire pour* $x \in E$ *fixé, c'est-à-dire*

$$\varphi(x,\mu y) = \mu\,\varphi(x,y), \qquad \varphi(x,y_1+y_2) = \varphi(x,y_1)+\varphi(x,y_2)$$

pour tout $\mu \in \mathbb{C}$ *et tous* $x,y,y_1,y_2 \in E$.

Une autre formulation équivalente de la sesquilinéarité est de vérifier l'identité de distributivité

$$\varphi(\lambda_1 x_1+\lambda_2 x_2, \mu_1 y_1+\mu_2 y_2) = \overline{\lambda}_1\mu_1\varphi(x_1,y_1)+\overline{\lambda}_1\mu_2\varphi(x_1,y_2)+\overline{\lambda}_2\mu_1\varphi(x_2,y_1)+\overline{\lambda}_2\mu_2\varphi(x_2,y_2)$$

pour tous $\lambda_1,\lambda_2,\mu_1,\mu_2 \in \mathbb{C}$ et $x_1,x_2,y_1,y_2 \in E$.

Écriture matricielle d'une forme sesquilinéaire. On suppose ici E de dimension finie n sur $\mathbb{C}$, muni d'une base $(e_1,\dots,e_n)$. Étant donnés des vecteurs

$$x = \sum_{j=1}^n x_j e_j \in E, \quad y = \sum_{j=1}^n y_j e_j \in E,$$

l'hypothèse de sesquilinéarité de φ implique

$$\varphi(x,y) = \varphi\Big(\sum_{j=1}^n x_j e_j, y\Big) = \sum_{j=1}^n \overline{x}_j\,\varphi(e_j,y) = \sum_{j=1}^n \overline{x}_j\,\varphi\Big(e_j, \sum_{k=1}^n y_k e_k\Big) = \sum_{j=1}^n\sum_{k=1}^n \overline{x}_j y_k\,\varphi(e_j,e_k).$$

Si l'on introduit les coefficients $c_{jk} = \varphi(e_j,e_k) \in \mathbb{C}$, ceci devient simplement

$$\varphi(x,y) = \sum_{1\le j,k\le n} c_{jk}\overline{x}_j y_k.$$

Inversement, toute expression de cette forme avec des coefficients $c_{jk} \in \mathbb{C}$ quelconques définit bien une forme sesquilinéaire $\varphi : E\times E \to \mathbb{C}$.

Définition 4.2. *Si* E *est un* $\mathbb{C}$*-espace vectoriel de dimension finie* n, $\mathcal{B} = (e_1,\dots,e_n)$ *une base de* E, *et* $\varphi : E\times E \to \mathbb{C}$ *une forme sesquilinéaire, on appelle* **matrice représentative** *de* φ *dans la base* $\mathcal{B}$ *la matrice complexe* $n\times n$ *de ses coefficients*:

$$\mathrm{Mat}_{\mathcal{B}}(\varphi) = C = (c_{jk}) = (\,\varphi(e_j,e_k)\,)_{1\le j,k\le n}.$$

Si $M = (c_{ij})$ est une matrice $n\times p$ (à n lignes et p-colonnes), on appelle **matrice adjointe** de M, notée M^*, la matrice $p\times n$ telle que

$$M^* = \overline{M}^{\,t} = (\overline{m}_{kj})$$

obtenue en conjuguant les coefficients et en transposant. Il est évident que $(M^*)^* = M$. On prendra garde au fait que $M \mapsto M^*$ est **anti-linéaire** (et non pas linéaire) :

$$(\lambda M)^* = \overline{\lambda}\,M^*, \qquad (M_1+M_2)^* = M_1^*+M_2^*.$$

D'autre part, si M, N sont des matrices $n\times p$ et $p\times r$ quelconques, alors

$$(MN)^* = N^*M^*.$$

Comme dans le cas des formes bilinéaires, on introduit les matrices colonnes complexes

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

représentant les coordonnées des vecteurs $x,y \in E$ dans la base $\mathcal{B}$. On obtient alors

$$\varphi(x,y) = \sum_{1\le i,j\le n} c_{ij}\overline{x}_i y_j = \begin{pmatrix}\overline{x}_1 & \dots & \overline{x}_n\end{pmatrix}\begin{pmatrix} c_{11} & \dots & c_{1n} \\ \vdots & & \vdots \\ c_{n1} & \dots & c_{nn}\end{pmatrix}\begin{pmatrix} y_1 \\ \vdots \\ y_n\end{pmatrix} = X^*C\,Y$$

La matrice C est uniquement déterminée par les relations $c_{jk} = \varphi(e_j, e_k)$. L'espace vectoriel $\mathrm{Sesq}(E)$ des formes sesquilinéaires est donc un espace vectoriel complexe de dimension n^2, isomorphe à l'espace des matrices carrées complexes $n \times n$

$$\mathrm{Sesq}(E) \simeq \mathrm{M}_{n\times n}(\mathbb{C}).$$

Par exemple, si $n = 2$, on a un espace de dimension complexe 4 ayant pour base les formes sesquilinéaires de matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Si $\varphi : E \times E \to \mathbb{C}$ est une forme sesquilinéaire, on définit la forme sesquilinéaire **adjointe** φ^* comme étant la forme

$$\varphi^*(x, y) = \overline{\varphi(y, x)}, \quad \forall x, y \in E.$$

La forme φ^* est encore sesquilinéaire : en effet on a

$$\varphi^*(\lambda x, y) = \overline{\varphi(y, \lambda x)} = \overline{\lambda\varphi(y, x)} = \overline{\lambda}\,\varphi^*(x, y),$$
$$\varphi^*(x, \mu y) = \overline{\varphi(\mu y, \lambda x)} = \overline{\mu}\varphi(y, x)} = \mu\,\varphi^*(x, y).$$

Comme $\varphi^*(e_j, e_k) = \overline{\varphi(e_k, e_j)}$, on voit aussi que $\mathrm{Mat}_{\mathcal{B}}(\varphi^*) = C^*$.

Définition 4.3. Soit E un espace vectoriel de dimension finie n, $\mathcal{B} = (e_1, \dots, e_n)$ une base de E, $\varphi : E \times E \to \mathbb{C}$ une forme sesquilinéaire et $C = \mathrm{Mat}_{\mathcal{B}}(\varphi)$ sa matrice. On dit que

(1) φ est hermitienne si $\varphi^* = \varphi \Leftrightarrow C^* = C \Leftrightarrow \overline{c}_{kj} = c_{jk}$ pour tous $1 \le j, k \le n$.

(2) φ est anti-hermitienne si $\varphi^* = -\varphi \Leftrightarrow C^* = -C \Leftrightarrow \overline{c}_{kj} = -c_{jk}$ pour tous $1 \le j, k \le n$.

On utilise aussi la terminologie de forme sesquilinéaire "symétrique" ou "anti-symétrique".

Exemple 4.4. Soit $E = C^0([a, b], \mathbb{C})$ l'espace (de dimension infinie) des fonctions continues $[a, b] \to \mathbb{C}$. Prenons également une fonction continue complexe $w \in E$. La forme

$$\varphi_w(f, g) = \int_a^b w(t)\,\overline{f(t)}\,g(t)\,dt$$

est une forme sesquilinéaire (vérification immédiate). On a

$$\varphi_w^*(f, g) = \overline{\varphi_w(g, f)} = \overline{\int_a^b w(t)\,\overline{g(t)}\,f(t)\,dt} = \int_a^b \overline{w(t)}\,\overline{f(t)}\,g(t)\,dt = \varphi_{\overline{w}}(f, g).$$

On voit que φ_w est une forme hermitienne si et seulement si la fonction w est réelle (la condition $\varphi_{\overline{w}}(f, g) = \varphi_w(f, g)$ entraîne facilement $w - \overline{w} = 0$ en soustrayant et en prenant $f(t) = 1$ et $g = w - \overline{w}$).

Une caractérisation alternative (qui n'existe pas dans le cas bilinéaire) est la suivante.

Proposition 4.5. *Une forme sesquilinéaire φ est hermitienne si et seulement si $\varphi(x, x)$ est réel pour tout $x \in E$.*

Démonstration. En effet, si $\varphi^*(x, y) = \overline{\varphi(y, x)} = \varphi(x, y)$ on obtient $\overline{\varphi(x, x)} = \varphi(x, x)$ pour $y = x$, donc $\varphi(x, x)$ est réel. Inversement, supposons que $\varphi(x, x)$ soit réel pour tout $x \in E$. Alors pour tous $x, y \in E$

$$\varphi(x, y) + \varphi(y, x) = \varphi(x + y, x + y) - \varphi(x - y, x - y) \in \mathbb{R},$$

ce qui entraîne

$$\varphi(x, y) - \overline{\varphi(y, x)} = \varphi(x, y) + \varphi(y, x) - \big(\varphi(y, x) + \overline{\varphi(y, x)}\big) \in \mathbb{R}.$$

En remplaçant y par iy on en déduit aussi

$$i(\varphi(x,y) - \overline{\varphi(y,x)}) = \varphi(x,iy) - \overline{\varphi(iy,x)} \in \mathbb{R}.$$

Si un nombre complexe a vérifie à la fois $a \in \mathbb{R}$ et $ia \in \mathbb{R}$, alors $a = 0$. On en déduit que $\varphi(x,y) - \overline{\varphi(y,x)} = 0$, donc φ est hermitienne. □

Remarque 4.6. Comme $(i\varphi)^* = -i\,\varphi^*$, on a de façon évidente

$$\varphi \text{ hermitienne} \iff i\varphi \text{ anti-hermitienne}, \qquad \varphi \text{ anti-hermitienne} \iff i\varphi \text{ hermitienne},$$

donc φ est anti-hermitienne si et seulement si $\varphi(x,x) \in i\mathbb{R}$ (imaginaire pur) pour tout $x \in E$. En particulier, pour φ hermitienne, les coefficients diagonaux vérifient $c_{jj} \in \mathbb{R}$, tandis que pour φ anti-hermitienne on a $c_{jj} \in i\mathbb{R}$.

Toute forme sesquilinéaire φ se décompose de manière unique en une somme

$$\varphi = \sigma + \alpha, \qquad \text{avec } \sigma \text{ hermitienne et } \alpha \text{ anti-hermitienne},$$

ces composantes étant déterminées par les formules

$$\sigma = \frac{1}{2}(\varphi + \varphi^*), \qquad \alpha = \frac{1}{2}(\varphi - \varphi^*).$$

On en déduit la proposition :

Proposition 4.7. *On a la décomposition en somme directe de sous-espaces vectoriels* réels

$$\mathrm{Sesq}(E) = \mathrm{Herm}(E) \oplus \mathrm{Antiherm}(E)$$

avec

$$\dim_{\mathbb{R}} \mathrm{Herm}(E) = \dim_{\mathbb{R}} \mathrm{Antiherm}(E) = \frac{1}{2}\dim_{\mathbb{R}} \mathrm{Sesq}(E) = n^2.$$

Démonstration. On a une bijection $\mathbb{R}$-linéaire

$$(*) \qquad \mathrm{Herm}(E) \xrightarrow{\simeq} \mathrm{Antiherm}(E), \qquad \varphi \mapsto i\,\varphi.$$

En particulier, $\mathrm{Herm}(E)$ et $\mathrm{Antiherm}(E)$ ne sont pas stables par multiplication par le scalaire $\lambda = i$, ce ne sont donc pas des sous-espaces vectoriels complexes ; il est clair cependant que ce sont des sous-espaces vectoriels réels. On a $\dim_{\mathbb{R}} \mathrm{Sesq}(E) = 2\dim_{\mathbb{C}} \mathrm{Sesq}(E) = 2n^2$, et l'isomorphisme $(*)$ donne bien les dimensions annoncées. □

Formule de polarisation hermitienne. Soit φ une forme sesquilinéaire hermitienne. La **forme quadratique hermitienne** associée est par définition

$$q(x) = \varphi(x,x), \qquad x \in E.$$

On sait que $q(x) \in \mathbb{R}$, par conséquent q est également associée à la forme $\mathbb{R}$-bilinéaire $(x,y) \mapsto \psi(x,y) = \mathrm{Re}\,\varphi(x,y)$; on notera que celle-ci est symétrique, car

$$\mathrm{Re}\,\varphi(y,x) = \mathrm{Re}\,\overline{\varphi(x,y)} = \mathrm{Re}\,\varphi(x,y).$$

En particulier, q est bien une *forme quadratique réelle*, lorsqu'on considère E comme un espace vectoriel sur le corps $\mathbb{K} = \mathbb{R}$. D'autre part, q vérifie la propriété supplémentaire

$$(**) \qquad q(\lambda x) = |\lambda|^2 q(x), \qquad \forall \lambda \in \mathbb{C},\ \forall x \in E,$$

propriété qui est quant à elle spécifique des formes quadratiques hermitiennes (voir la Proposition 4.8 ci-dessous). On notera qu'il s'agit d'un concept différent du concept de *forme quadratique complexe* associée à une forme $\mathbb{C}$-bilinéaire sur E, car dans ce cas on doit avoir $q(\lambda x) = \lambda^2 q(x)$ pour tout scalaire $\lambda \in \mathbb{C}$ (ce qui interdit entre autres que q soit réelle, si du moins $q \neq 0$).

La formule de polarisation a ici encore pour but de calculer φ en fonction de q. On part des identités

$$q(x+y) = \varphi(x+y, x+y) = \varphi(x,x) + \varphi(x,y) + \varphi(y,x) + \varphi(y,y),$$
$$q(x-y) = \varphi(x-y, x-y) = \varphi(x,x) - \varphi(x,y) - \varphi(y,x) + \varphi(y,y),$$
$$\frac{1}{4}\big(q(x+y) - q(x-y)\big) = \frac{1}{2}\big(\varphi(x,y) + \varphi(y,x)\big) = \operatorname{Re}\varphi(x,y),$$

la dernière égalité provenant de ce que $\varphi(y,x) = \overline{\varphi(x,y)}$. [Notons que la dernière ligne n'est en fait rien d'autre que la formule de polarisation appliquée à la forme bilinéaire symétrique $(x,y) \mapsto \operatorname{Re}\varphi(x,y)$]. En remplaçant y par iy on obtient

$$\frac{1}{4}\big(q(x+iy) - q(x-iy)\big) = \operatorname{Re}\varphi(x,iy) = \operatorname{Re}\big(i\,\varphi(x,y)\big) = -\operatorname{Im}\varphi(x,y),$$

car pour tout nombre complexe $a = u+iv$, on a $\operatorname{Re}(ia) = \operatorname{Re}(iu-v) = -v = -\operatorname{Im} a$. Comme $\varphi(x,y) = \operatorname{Re}\varphi(x,y) + i\operatorname{Im}\varphi(x,y)$, on obtient la **formule de polarisation hermitienne**

$$\text{(PH)} \qquad \varphi(x,y) = \frac{1}{4}\big(q(x+y) - q(x-y) - i\,q(x+iy) + i\,q(x-iy)\big).$$

Proposition 4.8. *Si q est une forme quadratique réelle vérifiant la propriété additionnelle $q(ix) = q(x)$ pour tout $x \in E$, alors la formule* (PH) *définit une forme sesquilinéaire hermitienne φ. Par conséquent q est alors une forme quadratique hermitienne.*

Démonstration. L'hypothèse que q soit une forme quadratique réelle implique que

$$\psi(x,y) = \frac{1}{4}\big(q(x+y) - q(x-y)\big)$$

est une forme bilinéaire symétrique réelle. On en déduit que

$$\varphi(x,y) = \frac{1}{4}\big(q(x+y) - q(x-y) - i\,q(x+iy) + i\,q(x-iy)\big) = \psi(x,y) - i\,\psi(x,iy)$$

est une forme $\mathbb{R}$-bilinéaire. On trouve par ailleurs

$$\varphi(x,iy) = \frac{1}{4}\big(q(x+iy) - q(x-iy) - i\,q(x-y) + i\,q(x+y)\big) = i\,\varphi(x,y),$$

tandis que l'hypothèse $q(ix) = q(x)$ implique $q(x) = q(-x) = q(-ix)$, d'où

$$\varphi(ix,y) = \frac{1}{4}\big(q(x-iy) - q(x+iy) - i\,q(x+y) + i\,q(x-y)\big) = -i\,\varphi(x,y).$$

Il en résulte facilement que φ est hermitienne. □

Méthode de Gauss. On va expliquer comment fonctionne ici la méthode de Gauss dans le cas hermitien. Les calculs sont un peu plus compliqués que dans le cas réel à cause de la présence de complexes conjugués dont il faut impérativement tenir compte.

Exemple 4.9. Considérons dans $E = \mathbb{C}^2$ la forme quadratique hermitienne

$$q(z,w) = 3z\overline{z} - 2iz\overline{w} + 2iw\overline{z} - 5w\overline{w}.$$

(on notera que les coefficients de $z\overline{w}$ et $w\overline{z}$ doivent nécessairement être conjugués, sinon q n'est pas réelle et il ne s'agit pas d'une forme quadratique hermitienne...) La forme sesquilinéaire hermitienne associée φ est donnée par

$$\varphi((z,w),(z',w')) = 3\overline{z}\,z' - 2i\overline{w}\,z' + 2i\overline{z}\,w' - 5\overline{w}\,w'$$

(seules les conjugués $\overline{z}$ et $\overline{w}$ doivent apparaître pour qu'il s'agisse d'une forme sesquilinéaire, et les coefficients doivent vérifier $\overline{c_{kj}} = c_{jk}$). La matrice de φ dans la base canonique $\mathcal{B}$ de $\mathbb{C}^2$ est

$$\mathrm{Mat}_{\mathcal{B}}(\varphi) = \begin{pmatrix} 3 & 2i \\ -2i & -5 \end{pmatrix}.$$

On cherche maintenant à décomposer ici $q(z,w)$ en somme de carrés $|\ell_j(z,w)|^2$ de formes linéaires. Lorsqu'on a un terme carré du type $\overline{z}\,z$, on regroupe tout les termes contenant la variable z, soit ici (en remettant les conjugués en premier)

$$3\overline{z}\,z - 2i\overline{w}\,z + 2i\overline{z}\,w = 3\,\overline{\Big(z + \frac{2i}{3}w\Big)}\,\Big(z + \frac{2i}{3}w\Big) - \frac{4}{3}\overline{w}\,w.$$

On obtient donc

$$q(z,w) = 3\,\overline{\Big(z + \frac{2i}{3}w\Big)}\,\Big(z + \frac{2i}{3}w\Big) - \frac{19}{3}\overline{w}\,w = 3\,|\ell_1(z,w)|^2 - \frac{19}{3}\,|\ell_2(z,w)|^2$$

avec

$$\ell_1(z,w) = z + \frac{2i}{3}w, \qquad \ell_2(z,w) = w.$$

Pour trouver une base orthogonale, on effectue le changement de coordonnées

$$\begin{cases} \widetilde{z} = z + \frac{2i}{3}w \\ \widetilde{w} = w \end{cases} \iff \begin{cases} z = \widetilde{z} - \frac{2i}{3}\widetilde{w} \\ w = \widetilde{w} \end{cases}$$

soit

$$\widetilde{X} = LX \qquad \Longleftrightarrow \qquad X = P\widetilde{X}$$

où L est la matrice de la famille de formes linéaires (ℓ_1, ℓ_2) et $P = L^{-1}$. On trouve donc ici une base $(\widetilde{e}_1, \widetilde{e}_1)$ donnée par la matrice de passage

$$P = \begin{pmatrix} 1 & -\frac{2i}{3} \\ 0 & 1 \end{pmatrix}, \quad \text{soit} \quad \widetilde{e}_1 = (1,0) \in \mathbb{C}^2, \ \ \widetilde{e}_2 = \Big(-\frac{2i}{3}, 1\Big) \in \mathbb{C}^2,$$

et les formes q et φ s'expriment dans les nouvelles coordonnées par

$$\begin{aligned} q(z,w) &= 3|\widetilde{z}|^2 - \frac{19}{3}|\widetilde{w}|^2 \\ \varphi((z,w),(z',w')) &= 3\,\overline{\widetilde{z}}\,\widetilde{z}' - \frac{19}{3}\,\overline{\widetilde{w}}\,\widetilde{w}', \\ \mathrm{Mat}_{(\widetilde{e}_1,\widetilde{e}_2)}(\varphi) &= \begin{pmatrix} 3 & 0 \\ 0 & -\frac{19}{3} \end{pmatrix}. \end{aligned}$$

On obtient ainsi une base orthogonale $(\widetilde{e}_1, \widetilde{e}_2)$, et on voit que q est de rang 2, de signature $(1,1)$.

Cas général. Reprenons le cas général d'une forme quadratique hermitienne sur $\mathbb{C}^n$

$$q(x_1, \ldots, x_n) = \sum_{1 \le j,k \le n} c_{jk}\,\overline{x}_j x_k.$$

On va montrer par récurrence sur n que q se décompose en une somme de carrés de modules de formes linéaires complexes indépendantes

$$q = \sum_{j=1}^{r} a_j |\ell_j|^2 \qquad \text{avec } \ell_j \in (\mathbb{C}^n)^*,\ a_j \in \mathbb{R}^* \text{ et } 0 \le r \le n.$$

On suppose d'abord que q comporte un "terme carré" non nul, par exemple $c_{11}\,\overline{x}_1 x_1$. On regroupe alors tous les termes contenant x_1 en factorisant le coefficient $c_{11} \neq 0$ (qui est réel). On obtient

$$c_{11}\Big(\overline{x}_1 x_1 + \frac{c_{12}}{c_{11}}\overline{x}_1 x_2 + \ldots + \frac{c_{1n}}{c_{11}}\overline{x}_1 x_n + \frac{c_{21}}{c_{11}}\overline{x}_2 x_1 + \ldots + \frac{c_{n1}}{c_{11}}\overline{x}_n x_1\Big).$$

Compte tenu des relations de conjugaison $\overline{c_{kj}} = c_{jk}$, ceci se factorise sous la forme

$$c_{11}\overline{\Big(x_1 + \frac{c_{12}}{c_{11}}\,x_2 + \ldots + \frac{c_{1n}}{c_{11}}\,x_n\Big)}\Big(x_1 + \frac{c_{12}}{c_{11}}\,x_2 + \ldots + \frac{c_{1n}}{c_{11}}\,x_n\Big) - q\,'(x_2, \ldots, x_n).$$

Si on introduit la forme linéaire

$$\ell_1(x_1, \ldots, x_n) = x_1 + \frac{c_{12}}{c_{11}}\,x_2 + \ldots + \frac{c_{1n}}{c_{11}}\,x_n,$$

il vient

$$q(x_1, \ldots, x_n) = c_{11}\big|\ell_1(x_1, \ldots, x_n)\big|^2 + \widetilde{q}\,(x_2, \ldots, x_n)$$

où $\widetilde{q}\,(x_2, \ldots, x_n)$ est une forme quadratique hermitienne ne portant plus que sur $n-1$ variables. Par hypothèse de récurrence $\widetilde{q} = \sum_{j=2}^{r} a_j|\ell_j|^2$ avec $a_j \in \mathbb{R}^*$ et $(\ell_2, \ldots, \ell_r)$ indépendantes en les variables $(x_2, \ldots, x_n)$. Ceci implique aussitôt que $\ell_1, \ldots, \ell_r \in (\mathbb{C}^n)^*$ sont indépendantes.

Dans le cas où il n'y a pas de termes carrés non nuls, mais seulement des termes rectangles non nuls, par exemple $c_{12}\,\overline{x}_1 x_2 + c_{21}\,\overline{x}_2 x_1$, on regroupe tous les termes contenant x_1 ou x_2, et on essaie de les factoriser en un produit $c_{12}\overline{(x_1 + \ldots)}(x_2 + \ldots) + \text{conjugué}$. Pour éviter d'écrire 2 fois chaque terme et son conjugué, on se bornera par exemple à n'écrire que les termes $\overline{x}_j x_k$ comportant $\overline{x}_1$ (conjugué) ou x_2 (non conjugué). Ceci donne

$$\begin{aligned}
&q(x_1, \ldots, x_n)\\
&\quad = c_{12}\Big(\overline{x}_1 x_2 + \frac{c_{13}}{c_{12}}\overline{x}_1 x_3 + \ldots + \frac{c_{1n}}{c_{12}}\overline{x}_1 x_n + \frac{c_{32}}{c_{12}}\overline{x}_3 x_2 + \ldots + \frac{c_{n2}}{c_{12}}\overline{x}_n x_2\Big) + \text{conjugué} + [x_3 \ldots x_n]\\
&\quad = c_{12}\overline{\Big(x_1 + \frac{c_{23}}{c_{21}}x_3 + \ldots + \frac{c_{2n}}{c_{21}}x_n\Big)}\Big(x_2 + \frac{c_{13}}{c_{12}}x_3 + \ldots + \frac{c_{1n}}{c_{12}}x_n\Big) + \text{conjugué} + [x_3 \ldots x_n]
\end{aligned}$$

où $[x_3 \ldots x_n]$ désigne des termes qui ne contiennent plus que les variables $x_3, \ldots, x_n$ ou leurs conjuguées. On trouve ainsi

$$q(x_1, \ldots, x_n) = \overline{A}\,B + \overline{B}\,A + \widetilde{q}\,(x_3, \ldots, x_n)$$

avec

$$A = x_1 + \frac{c_{23}}{c_{21}}x_3 + \ldots + \frac{c_{2n}}{c_{21}}x_n, \qquad B = c_{12}\Big(x_2 + \frac{c_{13}}{c_{12}}x_3 + \ldots + \frac{c_{1n}}{c_{12}}x_n\Big).$$

Pour mettre q sous forme de somme ou différence de carrés, on écrit simplement

$$\overline{A}\,B + \overline{B}\,A = \frac{1}{2}\big(\overline{(A+B)}(A+B) - \overline{(A-B)}(A-B)\big) = \frac{1}{2}\big(|A+B|^2 - |A-B|^2\big).$$

Ceci fait apparaître les deux formes linéaires indépendantes

$$\begin{aligned}
\ell_1(x_1, \ldots, x_n) &= A + B = x_1 + \frac{c_{23}}{c_{21}}x_3 + \ldots + \frac{c_{2n}}{c_{21}}x_n + c_{12}\Big(x_2 + \frac{c_{13}}{c_{12}}x_3 + \ldots + \frac{c_{1n}}{c_{12}}x_n\Big),\\
\ell_2(x_1, \ldots, x_n) &= A - B = x_1 + \frac{c_{23}}{c_{21}}x_3 + \ldots + \frac{c_{2n}}{c_{21}}x_n - c_{12}\Big(x_2 + \frac{c_{13}}{c_{12}}x_3 + \ldots + \frac{c_{1n}}{c_{12}}x_n\Big)
\end{aligned}$$

(A, B sont indépendantes du fait que $c_{12} \neq 0$, donc ℓ_1, ℓ_2 le sont aussi). On obtient alors

$$q(x_1, \ldots, x_n) = \frac{1}{2}\big|\ell_1(x_1, \ldots, x_n)\big|^2 - \frac{1}{2}\big|\ell_2(x_1, \ldots, x_n)\big|^2 + \widetilde{q}\,(x_3, \ldots, x_n),$$

et on poursuit le calcul avec $\widetilde{q}$ qui comporte 2 variables de moins et qui, par hypothèse de récurrence, se décompose en somme ou différence de carrés $\widetilde{q} = \sum_{j=3}^{r} a_j|\ell_j|^2$ de formes

linéaires indépendantes $\ell_3, \dots, \ell_r$ en les variables $x_3, \dots, x_n$. Nous pouvons résumer les résultats obtenus comme suit.

Théorème 4.10. *Soit E un espace vectoriel complexe de dimension finie munie d'une base $(e_1, \dots, e_n)$. Toute forme quadratique hermitienne q sur E admet une décomposition en somme de carrés*

$$q = \sum_{j=1}^{r} a_j |\ell_j|^2 \qquad \text{avec } \ell_j \in E^*,\ a_j \in \mathbb{R}^* \text{ et } 0 \leq r \leq n$$

où $(\ell_1, \dots, \ell_r)$ sont indépendantes. La forme polaire φ associée est la forme sesquilinéaire hermitienne

$$\varphi(x, y) = \sum_{1 \leq j \leq r} a_j \, \overline{\ell_j(x)} \, \ell_j(y).$$

Si l'on complète les ℓ_j en une base $(\ell_1, \dots, \ell_n)$ de E^* (par exemple en introduisant les coordonnées $\ell_j(x) = x_{k_j}$ qui n'ont pas été utilisées dans la méthode de Gauss), on obtient de nouvelles coordonnées $\widetilde{x}_j = \ell_j(x)$, c'est-à-dire en écriture matricielle $\widetilde{X} = LX \Leftrightarrow X = P\widetilde{X}$ avec $P = L^{-1}$. La matrice de passage P donne une base orthogonale $(\widetilde{e}_1, \dots, \widetilde{e}_n)$ dans laquelle

$$\mathrm{Mat}_{(\widetilde{e}_1, \dots, \widetilde{e}_n)}(q) = \begin{pmatrix} a_1 & \dots & 0 \\ \vdots & a_2 & \vdots \\ 0 & \dots & a_n \end{pmatrix} \quad \text{avec } a_1, \dots, a_r \neq 0,\ \ a_{r+1} = \dots = a_n = 0,\ \ r = \mathrm{rang}(q).$$

Changement de base. Supposons que la forme sesquilinéaire hermitienne φ s'écrive

$$\varphi(x, y) = X^* C Y \quad \text{dans une base } \mathcal{B} = (E_1, \dots, e_n).$$

Soit $\widetilde{\mathcal{B}} = (\widetilde{e}_1, \dots, \widetilde{e}_n)$ une nouvelle base, donnée par une matrice de passage P. On a $X = P\widetilde{X}$, $Y = P\widetilde{Y}$, d'où $X^* = (\widetilde{X})^* P^*$ et

$$\varphi(x, y) = (\widetilde{X})^* P^* C P \, \widetilde{Y}.$$

Ceci donne la formule

$$\widetilde{C} = \mathrm{Mat}_{\widetilde{\mathcal{B}}}(\varphi) = P^* C \, P.$$

Le Théorème 4.10 peut alors se reformuler ainsi : pour toute matrice hermitienne C, il existe une matrice de passage P telle que $\widetilde{C} = P^* C \, P$ soit diagonale de coefficients $a_1 \dots, a_n \in \mathbb{R}$. Le raisonnement conduisant à la preuve du théorème d'inertie de Sylvester est inchangée par rapport au cas réel (si ce n'est qu'on considère les dimensions sur $\mathbb{C}$).

Théorème 4.11 (Théorème d'inertie de Sylvester, cas hermitien)**.** *Soit q une forme quadratique hermitienne de rang r sur un espace vectoriel complexe E de dimension n. Il existe une décomposition en somme directe*

$$E = F_+ \oplus F_- \oplus \mathrm{Ker}(q)$$

associée à toute base orthogonale $(b_1, \dots b_n)$, où F_+ (resp. F_-) est engendré par les vecteurs b_j tels que $a_j = q(b_j) > 0$ (resp. $a_j = q(b_j) < 0$), et où $\mathrm{Ker}(q)$ est engendré par les vecteurs b_j tels que $a_j = q(b_j) = 0$. Les dimensions

$$r_+ = \dim_{\mathbb{C}} F_+, \qquad r_- = \dim_{\mathbb{C}} F_-$$

sont indépendantes de la base orthogonale $(b_1, \dots, b_n)$ choisie, et on a

$$r_+ + r_- = r = n - \dim_{\mathbb{C}} \mathrm{Ker}(q).$$

On dit que le couple (r_+, r_-) est la signature de q.

Définition 4.12. Si φ est une forme sesquilinéaire hermitienne sur E et x, y des vecteurs de E, on écrit $x \perp_\varphi y$ (ou simplement $x \perp y$) si $\varphi(x,y) = 0$.

On a bien $x \perp y \Leftrightarrow y \perp x$, du fait de la relation $\varphi(x,y) = \overline{\varphi(y,x)}$.

Théorème 4.13. *Pour tout sous-espace complexe F de E, l'orthogonal*

$$F^\perp = \big\{ y \in E \,;\ \forall x \in F,\ \varphi(x,y) = 0 \big\}$$

de F par rapport à une forme sesquilinéaire hermitienne φ est un sous-espace vectoriel complexe de E. On a toujours $F \subset (F^\perp)^\perp$.

On définit ici encore

$$\mathrm{Ker}(\varphi) = E^\perp = \{ y \,;\ \forall x \in E,\ \varphi(x,y) = 0 \},$$

$\mathrm{Ker}(\varphi)$ se calcule matriciellement en cherchant le noyau $\mathrm{Ker}(C) = \{ Y \,;\ CY = 0 \}$ de la matrice C. L'ensemble des vecteurs isotropes de $q(x) = \varphi(x,x)$ est

$$\mathrm{Isotrope}(q) = \{ x \in E \,;\ q(x) = 0 \},$$

c'est ici un cône complexe, c'est-à-dire que $x \in \mathrm{Isotrope}(q)$ implique $\lambda x \in \mathrm{Isotrope}(q)$ pour tout scalaire complexe λ. On a en général $\mathrm{Ker}(\varphi) \subset \mathrm{Isotrope}(q)$, l'inclusion pouvant être stricte.

Exemple 4.14. Sur $E = \mathbb{C}^2$, la forme quadratique hermitienne $q(z_1, z_2) = |z_1|^2 - |z_2|^2$ est non dégénérée, sa forme polaire φ admet pour matrice dans la base canonique $\mathcal{B}$

$$C = \mathrm{Mat}_{\mathcal{B}}(\varphi) = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

qui est une matrice inversible. On a donc

$$\mathrm{Ker}(\varphi) = \{0\} \neq \mathrm{Isotrope}(q) = \big\{ (z_1, z_2) \in \mathbb{C}^2 \,;\ |z_1| = |z_2| \big\}$$

Théorème 4.15. *Si φ est une forme sesquilinéaire* **non dégénérée** *(c'est-à-dire de noyau $\mathrm{Ker}(\varphi) = \{0\}$) sur un espace vectoriel complexe E de dimension finie, alors*

$$\dim_{\mathbb{C}} F^\perp = \dim_{\mathbb{C}} E - \dim_{\mathbb{C}} F \quad \textit{et} \quad (F^\perp)^\perp = F$$

pour tout sous-espace complexe F de E. On a en particulier

$$E = F \oplus F^\perp$$

dès que $F \cap F^\perp = \{0\}$, ce qui est toujours le cas si φ est définie positive.

Théorème 4.16 (Inégalité de Cauchy-Schwarz)**.** *Soit E un espace vectoriel complexe muni d'une forme hermitienne φ de forme quadratique associée q semi-positive. Alors pour tous $x, y \in E$ on a*

$$|\varphi(x,y)| \leq \|x\|\,\|y\| = \sqrt{q(x)}\,\sqrt{q(y)}.$$

De plus, si q est définie positive, l'égalité a lieu si et seulement si les vecteurs x, y sont $\mathbb{C}$-linéairement dépendants.

Démonstration. On a déjà observé que la forme quadratique hermitienne q est associée à la forme $\mathbb{R}$-bilinéaire symétrique $\psi(x,y) = \mathrm{Re}\,\varphi(x,y)$. Pour $x, y \in E$ (vu comme espace vectoriel sur $\mathbb{R}$), l'inégalité de Cauchy-Schwarz réelle implique donc

$$|\,\mathrm{Re}\,\varphi(x,y)| \leq \sqrt{q(x)}\,\sqrt{q(y)}.$$

Utilisons une écriture du nombre complexe $z = \varphi(x,y)$ en coordonnées polaires :

$$\varphi(x,y) = r\,e^{i\theta}, \qquad r = |\varphi(x,y)|.$$

Nous avons

$$r = e^{-i\theta}\varphi(x,y) = \varphi(e^{i\theta}x, y) = \mathrm{Re}\,\big(\varphi(e^{i\theta}x, y)\big)$$

puisque $e^{-i\theta} = \overline{e^{i\theta}}$ et $r \in \mathbb{R}_+$. On en déduit

$$|\varphi(x,y)| = \big|\,\mathrm{Re}\,\big(\varphi(e^{i\theta}x, y)\big)\big| \leq \sqrt{q(e^{i\theta}x)}\,\sqrt{q(y)} = \sqrt{q(x)}\,\sqrt{q(y)}.$$

Si q est définie positive, on sait d'après le cas réel que l'égalité a lieu si et seulement si $e^{i\theta}x$ et y sont $\mathbb{R}$-linéairement dépendants. Ceci entraîne bien que x, y sont $\mathbb{C}$-linéairement dépendants. Réciproquement, si x, y sont $\mathbb{C}$-linéairement dépendants, on a par exemple $y = \lambda x$ avec $\lambda \in \mathbb{C}$, et il vient

$$\varphi(x,y) = \lambda\,\varphi(x,x) = \lambda\,q(x), \qquad \sqrt{q(x)}\,\sqrt{q(y)} = |\lambda|\sqrt{q(x)}\,\sqrt{q(x)} = |\lambda|\,q(x).$$

Par conséquent on a bien $|\varphi(x,y)| = \sqrt{q(x)}\,\sqrt{q(y)}$ dans ce cas. □

Corollaire 4.17. *Si q est une forme quadratique hermitienne semi-positive, on a*

$$\mathrm{Ker}(q) = \mathrm{Isotrope}(q),$$

et q est définie positive si et seulement si $\mathrm{Ker}(q) = 0$.

Exemple 4.18. Soit $E = C^0([a,b], \mathbb{C})$ l'espace des fonctions continues $[a,b] \to \mathbb{C}$. Si $w \geq 0$ est une fonction continue sur $[a,b]$, la forme

$$\varphi_w(f,g) = \int_a^b w(t)\,\overline{f(t)}\,g(t)\,dt$$

est une forme sesquilinéaire hermitienne, de forme quadratique associée semi-positive

$$q_w(f) = \int_a^b w(t)\,|f(t)|^2\,dt \geq 0.$$

On en déduit l'inégalité de Cauchy-Schwarz

$$(*) \qquad \left|\int_a^b w(t)\,\overline{f(t)}\,g(t)\,dt\right| \leq \sqrt{\int_a^b w(t)\,|f(t)|^2\,dt}\ \ \sqrt{\int_a^b w(t)\,|g(t)|^2\,dt}.$$

Observons que q_w est définie positive dès que $w > 0$, ou même dès que $w \geq 0$ ne s'annule qu'en des points isolés : l'intégrale d'une fonction continue ≥ 0 non nulle est strictement postive, donc $q_w(f) = 0$ entraîne $w|f|^2 = 0$, et donc $f(t) = 0$ en tout point $t \in [a,b]$ où $w(t) \neq 0$; mais si w ne s'annule qu'en des points isolés, ceci entraîne $f(t) = 0$ partout. Dans ce cas, l'égalité a lieu dans $(*)$ si et seulement si les fonctions f, g sont proportionnelles, i.e. s'il existe $\lambda \in \mathbb{C}$ tel que $g(t) = \lambda f(t)$ pour tout $t \in [a,b]$, ou $f(t) = \lambda g(t)$ pour tout $t \in [a,b]$.

Définition 4.19. *On appelle* espace hermitien *un espace vectoriel complexe muni d'une forme quadratique hermitienne q définie positive, et de sa forme sesquilinéaire hermitienne associée φ. On la note en général $(x,y) \mapsto \langle x,y\rangle = \varphi(x,y)$ et on appelle $\langle x,y\rangle$ le* produit scalaire complexe.

Procédé d'orthonormalisation de Gram-Schmidt et projections orthogonales. Si F est un sous-espace complexe de dimension finie d'un espace hermitien (E,φ), le procédé d'orthonormalisation de Gram-Schmidt permet de déterminer une base orthonormée $(b_1,\dots,b_p)$ de F à partir d'une base quelconque $(a_1,\dots,a_p)$ de ce sous-espace, en sorte que

$$F_j = \mathrm{Vect}(a_1,\dots,a_j) = \mathrm{Vect}(b_1,\dots,b_j) \quad \text{pour tout } j = 1,\dots,p.$$

On utilise les mêmes formules que dans le cas réel : on calcule $b_1 = \frac{1}{\|a_1\|}a_1$, puis de proche en proche par récurrence sur j

$$(*) \qquad \widetilde{a}_j = a_j - \sum_{k=1}^{j-1}\langle b_k, a_j\rangle b_k = a_j - \pi_{F_{j-1}}(a_j), \qquad\qquad (**) \qquad b_j = \frac{1}{\|\widetilde{a}_j\|}\widetilde{a}_j.$$

De façon équivalente, on peut calculer directement les $\widetilde{a}_j$ en prenant $\widetilde{a}_1 = a_1$ et

$$(***) \qquad \widetilde{a}_j = a_j - \sum_{k=1}^{j-1}\frac{\langle \widetilde{a}_k, a_j\rangle}{\langle \widetilde{a}_k, \widetilde{a}_k\rangle}\,\widetilde{a}_k.$$

On a ici encore une décomposition en somme directe orthogonale

$$E = F \oplus F^\perp,$$

et les projections orthogonales $\pi_F : E \to F$, $\pi_{F^\perp} : E \to F^\perp$ (resp. les symétries orthogonales σ_F, $\sigma_{F^\perp}$ de E par rapport à F et $F^\perp$) se calculent par les formules

$$\pi_F(x) = \sum_{j=1}^{p}\langle b_j, x\rangle\, b_j \in F, \qquad \pi_{F^\perp}(x) = x - \sum_{j=1}^{p}\langle b_j, x\rangle\, b_j \in F^\perp,$$
$$\sigma_F(x) = 2\pi_F(x) - x = 2\Big(\sum_{j=1}^{p}\langle b_j, x\rangle\, b_j\Big) - x,$$
$$\sigma_{F^\perp}(x) = x - 2\pi_F(x) = x - 2\Big(\sum_{j=1}^{p}\langle b_j, x\rangle\, b_j\Big) = -\sigma_F(x).$$

Il faut juste prendre garde au fait de placer le vecteur x du bon côté des produits scalaires complexes, pour tenir compte de la $\mathbb{C}$-linéarité.

5. Normes et distances, méthode des moindres carrés

Soit $(E, \langle\ ,\ \rangle)$ un espace euclidien ou hermitien, et $\|x\| = \sqrt{\langle x, x\rangle}$ la norme associée. Pour tous $x, y \in E$, nous avons

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

et comme $\langle y, x\rangle = \overline{\langle x, y\rangle}$ on obtient la formule générale

$$(*) \qquad \|x+y\|^2 = \|x\|^2 + 2\,\mathrm{Re}\langle x, y\rangle + \|y\|^2.$$

(la partie réelle étant bien sûr sans effet dans le cas euclidien).

Théorème 5.1 (Inégalité triangulaire)**.** *Pour tous $x, y \in E$ on a*

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

et l'égalité a lieu si et seulement si x, y sont $\mathbb{R}$-colinéaires et de même sens, c'est-à-dire s'il existe $\lambda \in \mathbb{R}_+$ tel que $y = \lambda x$ ou $x = \lambda y$.

Démonstration. On peut toujours se ramener au cas euclidien en remplaçant au besoin le produit scalaire hermitien complexe par le produit scalaire euclidien réel $\langle x, y\rangle_{\mathbb{R}} = \mathrm{Re}\langle x, y\rangle$. Il suffit donc de raisonner dans le cas euclidien. L'inégalité de Cauchy-Schwarz implique

$$\langle x, y\rangle \le |\langle x, y\rangle| \le \|x\|\|y\|$$

et les égalités ont lieu si et seulement si x, y sont $\mathbb{R}$-colinéaires et de même sens. D'après la formule $(*)$, on en déduit

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

c'est-à-dire

$$\|x+y\|^2 \leq (\|x\|+\|y\|)^2.$$

En prenant la racine carrée, on obtient le résultat voulu, et l'égalité se produit si et seulement si x, y sont $\mathbb{R}$-colinéaires et de même sens. □

Il est souvent utile également de savoir minorer la norme d'une somme de vecteurs :

Proposition 5.2. *Pour tous $x, y \in E$, on a*

$$\|x+y\| \geq \big|\|x\|-\|y\|\big|.$$

Démonstration. L'inégalité triangulaire appliquée à $x' = x+y$ et $y' = -y$ donne

$$\|x\| = \|(x+y)+(-y)\| \leq \|x+y\|+\|y\|,$$

donc $\|x+y\| \geq \|x\|-\|y\|$. En échangeant les rôles de x et y on obtient de même l'inégalité $\|x+y\| \geq \|y\|-\|x\|$, ce qui démontre la proposition. On pourra vérifier ici que l'égalité se produit si et seulement si x, y sont $\mathbb{R}$-colinéaires et de sens contraires. □

Une autre conséquence immédiate de la formule $(*)$ est le "théorème de Pythagore généralisé" :

Théorème 5.3. *Soit $(E, \langle\ ,\ \rangle)$ un espace euclidien ou hermitien. Alors pour tous $x, y \in E$, on a*

$$\|x+y\|^2 = \|x\|^2+\|y\|^2 \iff \mathrm{Re}(\langle x, y\rangle) = 0.$$

En particulier, si E est un espace **euclidien** *on a*

$$\|x+y\|^2 = \|x\|^2+\|y\|^2 \iff x \perp y.$$

Les deux propositions suivantes sont souvent très utiles.

Proposition 5.4. *Soit $(E, \langle\ ,\ \rangle)$ un espace euclidien ou hermitien, et soient $x_1, \ldots, x_k \in E$ une famille de vecteurs deux à deux orthogonaux. Alors on a*

$$\|x_1+\ldots+x_k\|^2 = \|x_1\|^2+\ldots+\|x_k\|^2.$$

Démonstration. Supposons $x_1, \ldots, x_k \in E$ deux à deux orthogonaux. On a donc

$$\langle x_i, x_j\rangle = 0 \text{ pour tout } i \neq j.$$

Mais alors, on a

$$\|x_1+\ldots+x_k\|^2 = \langle x_1+\ldots+x_k, x_1+\ldots+x_k\rangle = \sum_{i,j}\langle x_i, x_j\rangle.$$

Mais puisque $\langle x_i, x_j\rangle = 0$ pour tout $i \neq j$, on obtient

$$\|x_1+\ldots+x_k\|^2 = \sum_{i=1}^{k}\langle x_i, x_i\rangle = \sum_{i=1}^{k}\|x_i\|^2,$$

ce que l'on voulait démontrer. □

Proposition 5.5. *Soit $(E, \langle\ ,\ \rangle)$ un espace euclidien ou hermitien, et soient $a_1, \ldots, a_k \in E$ une famille de vecteurs* **non nuls** *deux à deux orthogonaux. Alors $(a_1, \ldots, a_k)$ est une famille libre.*

Démonstration. Soient $\lambda_1, \ldots, \lambda_k \in \mathbb{K}$ tels que

$$\lambda_1 a_1 + \ldots + \lambda_k a_k = 0_E.$$

Soit $1 \leq j \leq k$. On a

$$\langle a_j, \lambda_1 a_1 + \ldots + \lambda_k a_k \rangle = \langle a_j, 0_E \rangle = 0,$$

et donc

$$\sum_{i=1}^{k} \lambda_i \langle a_j, a_i \rangle = 0.$$

Puisque les a_i sont deux à deux orthogonaux, cela s'écrit

$$\lambda_j \langle a_j, a_j \rangle = 0.$$

Puisque par hypothèse $a_j \neq 0$, on a $\langle a_j, a_j \rangle > 0$, et donc $\lambda_j = 0$. Ceci achève la démonstration. □

Le théorème suivant donne une caractérisation de la projection orthogonale comme solution d'un problème de minimisation, qui généralise une propriété bien connue de la projection orthogonale d'un vecteur de $\mathbb{R}^2$ ou $\mathbb{R}^3$.

Proposition 5.6. *Soit $(E, \langle\ ,\ \rangle)$ un espace euclidien ou hermitien, et soit F un sous-espace de E de dimension finie. Soit $x \in E$. Alors la projection orthogonale $v = \pi_F(x)$ est* **l'unique** *élément $v \in F$ vérifiant*

$$\|x - v\| = \min_{w \in F} \|x - w\|.$$

Démonstration. Par définition de la projection orthogonale $v = \pi_F(x)$, nous avons

$$x = v + (x - v), \qquad v \in F,\ x - v \in F^{\perp}.$$

Soit $w \in F$. On peut écrire

$$x - w = (v - w) + (x - v), \qquad v - w \in F,\ x - v \in F^{\perp}$$

et le théorème de Pythagore donne

$$\|x - w\|^2 = \|v - w\|^2 + \|x - v\|^2.$$

Ceci implique bien $\|x - w\| \geq \|x - v\|$, avec égalité si et seulement $\|v - w\| = 0$, c'est-à-dire si $w = v$. □

Ce résultat permet souvent de résoudre des problèmes de minimisation – précisément ceux qui mettent en jeu des formes quadratiques – c'est la méthode dite des "moindres carrés".

Exemple 5.7. Considérons le problème suivant. On veut mesurer une donnée y (pH d'une solution, température) en fonction d'un paramètre x (concentration d'un ion, temps). Considérons les n points $P_1 := (x_1, y_1), \ldots, P_n := (x_n, y_n)$ de $\mathbb{R}^2$ représentant par exemple le résultat de n expérimentations. Supposons que la théorie nous dise que y varie linéairement en fonction de x. A cause des erreurs de manipulations ou de mesure, les n points $P_1, \ldots, P_n$ ne sont pas alignés.

Comment trouver la droite de meilleure approximation, c'est-à-dire la droite d'équation $y = \alpha x + \beta$ telles que les points théoriques $Q_1 := (x_1, \alpha x_1 + \beta), \ldots, Q_n := (x_n, \alpha x_n + \beta)$ soient les plus proches possibles des points expérimentaux $P_1, \ldots, P_n$?

Plus précisément, comment choisir la droite $y = \alpha x + \beta$ telle que

$$d := P_1 Q_1^2 + \ldots + P_n Q_n^2$$

soit minimale ?

On veut donc trouver $\alpha, \beta \in \mathbb{R}^2$ tels que

$$d := (y_1 - (\alpha x_1 + \beta))^2 + \ldots + (y_n - (\alpha x_n + \beta))^2$$

soit minimale. Posons

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \text{et} \quad U = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

On voit facilement que

$$Y - (\alpha X + \beta U) = \begin{pmatrix} y_1 - (\alpha x_1 + \beta) \\ \vdots \\ y_n - (\alpha x_n + \beta) \end{pmatrix},$$

et donc

$$d = \|Y - (\alpha X + \beta U)\|^2.$$

Posons $F = \mathrm{Vect}(X, U)$. On veut donc minimiser $\|Y - V\|$, lorsque V décrit F. D'après les propriétés vues ci-dessus, le minimum est obtenu pour la projection orthogonale $V = \pi_F(Y)$. Les coefficients α et β seront alors donnés par la relation

$$\pi_F(Y) = \alpha X + \beta U.$$

Posons

$$\overline{x} = \frac{x_1 + \ldots + x_n}{n}, \qquad \overline{y} = \frac{y_1 + \ldots + y_n}{n}.$$

Appliquons l'algorithme de Gram-Schmidt à la base $(a_1, a_2) = (U, X)$ de F. On a $b_1 = \frac{1}{\|U\|}U$ et

$$\widetilde{a}_2 = a_2 - \langle b_1, a_2\rangle b_1 = X - \frac{\langle U, X\rangle}{\langle U, U\rangle}U = X - \overline{x}U, \qquad b_2 = \frac{1}{\|\widetilde{a}_2\|}\widetilde{a}_2.$$

Or $\widetilde{a}_2 = X - \overline{x}U$ est le vecteur ayant pour composantes $x_i - \overline{x}$, et on a donc

$$\begin{aligned} \pi_F(Y) &= \langle b_1, Y\rangle b_1 + \langle b_2, Y\rangle b_2 = \frac{\langle U, Y\rangle}{\langle U, U\rangle}U + \frac{\langle \widetilde{a}_2, Y\rangle}{\langle \widetilde{a}_2, \widetilde{a}_2\rangle}\widetilde{a}_2 \\ &= \overline{y}U + \frac{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}(X - \overline{x}U). \end{aligned}$$

On remarque que

$$\sum_{i=1}^{n}(x_i - \overline{x})y_i = \Big(\sum_{i=1}^{n} x_i y_i\Big) - n\overline{x}\,\overline{y} = \sum_{i=1}^{n}(x_i y_i - \overline{x}\,\overline{y}),$$

ce qui donne

$$\pi_F(Y) = \frac{\displaystyle\sum_{i=1}^{n}(x_i y_i - \overline{x}\,\overline{y})}{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}X + \left(\overline{y} - \overline{x}\frac{\displaystyle\sum_{i=1}^{n}(x_i y_i - \overline{x}\,\overline{y})}{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)U = \alpha X + \beta U.$$

Ainsi, la droite de meilleure approximation est donnée par $y = \alpha x + \beta$, soit encore

$$y = \frac{\sum_{i=1}^{n}(x_i y_i - \overline{x}\,\overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}(x - \overline{x}) + \overline{y}.$$

Exemple 5.8. On cherche maintenant à approximer une fonction $f : [a, b] \to \mathbb{R}$ par une droite $y = \alpha x + \beta$. Dans ce cas, la méthode précédente ne fonctionne plus telle quelle, puisque l'on doit a priori considérer une infinité de points.

L'idée est de considérer un grand nombre de points sur le graphe de f, dont les abcisses sont régulièrement espacés $P_1 = (x_1, f(x_1)), \ldots, P_n = (x_n, f(x_n))$, avec $x_i = a + \dfrac{(b-a)i}{n}$ et de considérer la droite de meilleure approximation pour ces points. Bien sûr, plus n est grand, meilleure sera l'approximation. L'entier n étant fixé, on doit donc minimiser

$$d := (f(x_1) - (\alpha x_1 + \beta))^2 + \ldots + (f(x_n) - (\alpha x_n + \beta))^2.$$

Ceci revient aussi à minimiser

$$S_n := \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - (\alpha x_i + \beta))^2, \text{ avec } x_i = a + \frac{(b-a)i}{n}.$$

D'après la théorie de l'intégrale de Riemann, la somme S_n converge vers

$$\int_a^b (f(x) - (\alpha x + \beta))^2 dx.$$

En particulier, S_n est très proche de cette intégrale lorsque n est suffisamment grand. Il est alors naturel de définir la droite de meilleure approximation $y = \alpha x + \beta$ comme celle qui minimise l'intégrale ci-dessus. Ce genre d'intégrale s'interprète souvent comme l'énergie d'un système. Ainsi, le problème de minimisation précédent revient à demander à minimiser cette énergie.

Considérons par exemple le problème de minimisation suivant : trouver $\alpha, \beta \in \mathbb{R}$ tel que l'intégrale

$$\int_0^{\frac{\pi}{2}} \big(\cos x - \alpha - \beta x\big)^2 dx$$

soit minimale. Pour cela, on introduit l'espace vectoriel E des fonctions continues $f : [0, \frac{\pi}{2}] \to \mathbb{R}$, et la forme

$$\langle\ ,\ \rangle : E \times E \to \mathbb{R}, \qquad (f, g) \mapsto \int_0^{\frac{\pi}{2}} f(x)g(x)\,dx.$$

On sait que $\langle\ ,\ \rangle$ est un produit scalaire euclidien sur E. Considérons le sous-espace F de E engendré par les fonctions $x \mapsto 1$ et $x \mapsto x$, c'est-à-dire

$$F = \{x \mapsto \alpha + \beta x\,;\ \alpha, \beta \in \mathbb{R}\}.$$

Le problème de minimisation se reformule alors ainsi : *trouver $v \in F$ tel que $\|\cos - v\|$ soit minimal*. D'après la Proposition 5.6, cela revient à dire que $v = \pi_F(x \mapsto \cos x)$. On cherche donc à calculer la projection orthogonale de $x \mapsto \cos x$ sur F.

Appliquons le procédé de Gram-Schmidt à la base $a_1 = (x \mapsto 1)$, $a_2 = (x \mapsto x)$ de F. On a

$$\|a_1\|^2 = \frac{\pi}{2}, \qquad b_1 = \sqrt{\frac{2}{\pi}}a_1 = \Big(x \mapsto \sqrt{\frac{2}{\pi}}\Big), \qquad \langle b_1, a_2\rangle b_1 = \frac{2}{\pi}\langle a_1, a_2\rangle a_1 = \frac{2}{\pi}\frac{\pi^2}{8}a_1 = \frac{\pi}{4}a_1$$

et

$$\widetilde{a}_2 = a_2 - \langle b_1, a_2\rangle b_1 = a_2 - \frac{\pi}{4}a_1 : x \mapsto x - \frac{\pi}{4}, \qquad \|\widetilde{a}_2\|^2 = \frac{2}{3}\Big(\frac{\pi}{4}\Big)^3, \qquad b_2 = \frac{1}{\|\widetilde{a}_2\|}\widetilde{a}_2.$$

On obtient en définitive

$$v = \pi_F(\cos) = \langle b_1, \cos\rangle b_1 + \langle b_2, \cos\rangle b_2 = \frac{\langle a_1, \cos\rangle}{\|a_1\|^2}a_1 + \frac{\langle \widetilde{a}_2, \cos\rangle}{\|\widetilde{a}_2\|^2}\widetilde{a}_2.$$

Un calcul explicite donne

$$\langle a_1, \cos\rangle = 1, \qquad \langle \widetilde{a}_2, \cos\rangle = \int_0^{\pi/2} (x - \pi/4)\cos x\, dx = \frac{\pi}{4} - 1,$$

d'où

$$v(x) = \frac{2}{\pi} + \frac{96}{\pi^3}\Big(\frac{\pi}{4} - 1\Big)\Big(x - \frac{\pi}{4}\Big) \simeq 1.158469 - 0.664439\, x.$$

Exemple 5.9. Dans une situation plus générale, on cherche à approximer une fonction $f : [a, b] \to \mathbb{R}$ par une fonction v appartenant à un sous-espace vectoriel $F = \mathrm{Vect}(a_1, \ldots, a_p)$ de $C^0([a, b], \mathbb{R})$, de façon à ce que l'intégrale

$$\int_a^b (f(x) - v(x))^2 dx$$

soit minimale lorsque v décrit F. Comme précédemment, le calcul consiste à utiliser la méthode de Gram-Schmidt pour trouver une base orthogonale $(\widetilde{a}_1, \ldots, \widetilde{a}_p)$ (resp. une base orthonormée $(b_1, \ldots, b_p)$) de F, et on prend

$$v = \pi_F(f) = \sum_{i=1}^{p} \langle b_i, f\rangle b_i = \sum_{i=1}^{p} \frac{\langle \widetilde{a}_i, f\rangle}{\langle \widetilde{a}_i, \widetilde{a}_i\rangle}\widetilde{a}_i.$$

6. Endomorphismes symétriques, anti-symétriques, orthogonaux et unitaires

On suppose ici que E est un espace vectoriel euclidien ($\mathbb{K} = \mathbb{R}$) ou hermitien ($\mathbb{K} = \mathbb{C}$) de dimension finie $n = \dim_{\mathbb{K}} E$, et on note $(x, y) \mapsto \langle x, y\rangle$ le produit scalaire de E. Fixons une base orthonormée $(e_1, \ldots, e_n)$. Si $X = (x_j)$ et $Y = (y_j)$ sont les matrices colonnes des vecteurs $x, y \in E$ dans la base (e_j), on peut écrire

$$\langle x, y\rangle = \sum_{j=1}^{n} \overline{x}_j y_j = X^* Y$$

(la conjugaison étant sans effet si $\mathbb{K} = \mathbb{R}$, dans ce cas $M^* = M^t$ pour toute matrice réelle).

Théorème 6.1. *Soit $u \in \mathcal{L}_{\mathbb{K}}(E, E) = \mathrm{End}_{\mathbb{K}}(E)$ un endomorphisme de matrice A dans la base orthonormée $(e_1, \ldots, e_n)$. Alors, il existe un endomorphisme $u^* \in \mathcal{L}_{\mathbb{K}}(E, E)$ unique tel que*

$$\forall x, y \in E, \quad \langle u^*(x), y\rangle = \langle x, u(y)\rangle.$$

De plus, on a

$$\mathrm{Mat}_{(e_j)}(u^*) = A^* = \big(\mathrm{Mat}_{(e_j)}(u)\big)^*.$$

On appelle u^* l'endomorphisme adjoint *de u.*

Démonstration. Le vecteur $u(y)$ admet pour matrice AY, on a donc

$$\langle x, u(y)\rangle = X^*(AY) = (X^*A)Y = (A^*X)^*Y.$$

Si l'on définit u^* comme étant l'endomorphisme de matrice A^*, on a donc bien la relation voulue $\langle u^*(x), y\rangle = \langle x, u(y)\rangle$. Il n'y a pas d'autre choix possible, car si B est la matrice

de u^*, la relation $\langle u^*(x), y\rangle = \langle x, u(y)\rangle$ se traduit par $(BX)^*Y = X^*(AY)$, c'est-à-dire $X^*(B^* - A)Y = 0$ pour toutes matrices colonnes X, Y. Ceci entraîne $B^* - A = 0$ lorsqu'on considère tous les couples $(x, y) = (e_j, e_k)$ de vecteurs de base, donc $B = A^*$. □

La définition de l'adjoint implique aussitôt que $(u^*)^* = u$. On prendra garde au fait que l'adjonction est *anti-linéaire* : $(\lambda_1 u_1 + \lambda_2 u_2)^* = \overline{\lambda}_1\, u_1^* + \overline{\lambda}_2\, u_2^*$ lorsque $\lambda_1, \lambda_2 \in \mathbb{C}$.

Définition 6.2. *On dit que*

- *u est un endomorphisme symétrique ($\mathbb{K} = \mathbb{R}$), resp. hermitien ($\mathbb{K} = \mathbb{C}$), si $u^* = u$.*
- *u est un endomorphisme anti-symétrique, resp. anti-hermitien, si $u^* = -u$.*

On utilise aussi indifféremment la terminologie "endomorphisme symétrique" ou "endomorphisme anti-symétrique" dans le cas complexe.

Exemple 6.3. Soit F un sous-espace vectoriel de E, et $\pi_F : E \to F$ la projection orthogonale sur F. Si on utilise la décomposition en somme directe

$$E = F \oplus F^\perp, \qquad x = x' + x'', \quad y = y' + y'', \quad x', y' \in F, \quad x'', y'' \in F^\perp$$

nous obtenons

$$\langle x, y\rangle = \langle x', y'\rangle + \langle x'', y''\rangle.$$

Comme $\pi_F(x) = x'$ et $\pi_F(y) = y'$, ceci donne en particulier

$$\langle x, \pi_F(y)\rangle = \langle x' + x'', y'\rangle = \langle x', y'\rangle = \langle x', y' + y''\rangle = \langle \pi_F(x), y\rangle$$

et on voit donc que $\pi_F = \pi_F^*$ est un endomorphisme symétrique. La symétrie orthogonale σ_F est donnée par $\sigma_F = 2\pi_F - \mathrm{Id}_E$, et comme $(\mathrm{Id}_E)^* = \mathrm{Id}_E$, on voit que $\sigma_F = \sigma_F^*$ est aussi un endomorphisme symétrique.

Le théorème suivant relie les caractéristiques géométriques d'un endomorphisme u et celles de son adjoint u^*.

Théorème 6.4. *Soit $u \in \mathcal{L}_{\mathbb{K}}(E, E)$ un endomorphisme d'un espace vectoriel E euclidien ou hermitien de dimension finie. Alors*

- $\mathrm{Ker}(u^*) = \big(\mathrm{Im}(u)\big)^\perp$;
- $\mathrm{Im}(u^*) = \big(\mathrm{Ker}(u)\big)^\perp$;
- *Si S est un sous-espace de E stable par u, i.e. $u(S) \subset S$, alors son orthogonal $S^\perp$ est stable par u^*, i.e. $u^*(S^\perp) \subset S^\perp$.*

Démonstration. Soit $x \in$. On a $x \in \mathrm{Ker}(u^*)$ si et seulement si $u^*(x) = 0$, ce qui équivaut à $\langle u^*(x), y\rangle = 0$ pour tout $y \in E$ du fait que le produit scalaire est supposé non dégénéré. Comme $\langle u^*(x), y\rangle = \langle x, u(y)\rangle$, ce terme s'annule pour tout y si et seulement x est orthogonal à $\mathrm{Im}(u)$ soit $x \in (\mathrm{Im}(u))^\perp$. Ceci démontre la première égalité. En remplaçant u par u^* dans celle-ci, on obtient $\mathrm{Ker}(u) = \mathrm{Ker}(u^{**}) = (\mathrm{Im}\, u^*)^\perp$, donc en passant aux orthogonaux

$$\big(\mathrm{Ker}(u)\big)^\perp = \big((\mathrm{Im}(u^*))^\perp\big)^\perp = \mathrm{Im}(u^*)$$

et la deuxième égalité s'ensuit. Supposons maintenant que $u(S) \subset S$ et soit $x \in S^\perp$. Alors pour tout $y \in S$ on a $u(y) \in S$, donc

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

puisque $x \in S^\perp$. Ceci montre bien que $u^*(x) \in S^\perp$, donc $u^*(S^\perp) \subset S^\perp$. □

Comme exemple, on va considérer le cas des projections et symétries *obliques*, relatives à une décomposition en somme directe $E = F' \oplus F''$, avec F' et F'' non (nécessairement) orthogonaux. Pour $x \in E$, on note

$$x = x' + x'', \quad x' \in F', \quad x'' \in F'', \qquad \pi_{F',F''}(x) = x', \quad \sigma_{F',F''} = x' - x'',$$

où $\pi_{F',F''}$ est la projection sur F' parallèlement à F'', et $\sigma_{F',F''} = 2\pi_{F',F''} - \mathrm{Id}_E$ la symétrie par rapport à F' parallèlement à F''. Rappelons la caractérisation des projections et symétries.

Théorème 6.5. *Soient $p, s \in \mathcal{L}_{\mathbb{K}}(E,E)$ des endomorphismes d'un espace vectoriel E de dimension finie. Alors*

• *p est une projection si et seulement si $p \circ p = p$. Dans ce cas, p est la projection sur $F' = \mathrm{Im}(p)$ parallèlement à $F'' = \mathrm{Ker}(p)$. De plus*

$$F' = \mathrm{Im}(p) = \mathrm{Ker}(p - \mathrm{Id}_E) = \{x \in E\,;\ p(x) = x\}.$$

• *s est une symétrie si et seulement si $s \circ s = \mathrm{Id}_E$ (i.e. s est une "involution"). Dans ce cas, s est la symétrie par rapport à $F' = \mathrm{Ker}(s - \mathrm{Id}_E) = \{x \in E\,;\ s(x) = x\}$ (vecteurs invariants), parallèlement à $F'' = \mathrm{Ker}(s + \mathrm{Id}_E) = \{x \in E\,;\ s(x) = -x\}$.*

Démonstration. La projection $p = \pi_{F',F''}$ et la symétrie $s = \sigma_{F',F''}$ vérifient bien les propriétés énoncées. Réciproquement, supposons $p \circ p = p$ et soient $F' = \mathrm{Im}(p)$, $F'' = \mathrm{Ker}(p)$. Alors pour tout $x \in E$ on peut écrire

$$x = x' + x'', \qquad x' = p(x), \quad x'' = x - p(x).$$

Nous avons $p(x') = p \circ p(x) = p(x) = x'$, tandis que $p(x'') = p(x) - p \circ p(x) = 0$. Ceci montre bien que $x' \in F' = \mathrm{Im}(p)$ et $x'' \in F'' = \mathrm{Ker}(p)$. Vérifions maintenant que $F' \cap F'' = \{0\}$: soit $x = p(v) \in F' = \mathrm{Im}(p)$. Si $x \in F'' = \mathrm{Ker}(p)$, alors $x = p(v) = p \circ p(v) = p(x) = 0$, donc $x = 0$. Enfin, on a bien $\mathrm{Im}(p) \subset \mathrm{Ker}(p - \mathrm{Id}_E)$ du fait que $(p - \mathrm{Id}_E) \circ p = p \circ p - p = 0$, et inversement il est évident que

$$\mathrm{Ker}(p - \mathrm{Id}_E) = \{x \in E\,;\ p(x) = x\} \subset \mathrm{Im}(p).$$

Les propriétés relatives à s se démontrent en posant $s = 2p - \mathrm{Id}_E$, ce qui équivaut à prendre $p = \frac{1}{2}(s + \mathrm{Id}_E)$. La propriété d'involution $s \circ s = \mathrm{Id}_E$ implique

$$p \circ p = \frac{1}{4}\big(s \circ s + 2s + \mathrm{Id}_E\big) = \frac{1}{2}(s + \mathrm{Id}_E) = p$$

et on voit que $F'' = \mathrm{Ker}(p) = \mathrm{Ker}(s + \mathrm{Id}_E)$, $F' = \mathrm{Ker}(p - \mathrm{Id}_E) = \mathrm{Ker}(s - \mathrm{Id}_E)$. □

Exemple 6.6. Considérons une projection oblique $p = \pi_{F',F''}$. Alors $p \circ p = p$ implique $p^* \circ p^* = p^*$, ce qui montre que p^* est encore une projection. Or $\mathrm{Ker}(p^*) = (\mathrm{Im}(p))^\perp = (F')^\perp$ et $\mathrm{Im}(p^*) = (\ker(p))^\perp = (F'')^\perp$. Ceci implique $p^* = \pi_{(F'')^\perp,(F')^\perp}$, autremennt dit on a la formule

$$(\pi_{F',F''})^* = \pi_{(F'')^\perp,(F')^\perp}.$$

La relation $s = 2p - \mathrm{Id}_E$ donne de même la relation entre les symétries obliques :

$$(\sigma_{F',F''})^* = \sigma_{(F'')^\perp,(F')^\perp}.$$

Corollaire 6.7. *Un endomorphisme $p \in \mathcal{L}_{\mathbb{K}}(E,E)$ est une projection orthogonale si et seulement si $p \circ p = p$ et $p^* = p$. Un endomorphisme $s \in \mathcal{L}_{\mathbb{K}}(E,E)$ est une symétrie orthogonale si et seulement si $s \circ s = \mathrm{Id}_E$ et $s^* = s$, autrement dit le caractère orthogonal de ces endomorphismes est caractérisé par la condition de symétrie $u^* = u$.*

Démonstration. On sait que la condition $p \circ p = p$ implique l'existence d'une somme directe $E = F' \oplus F''$ telle que $p = \pi_{F',F''}$. Comme $p^* = \pi_{(F'')^\perp,(F')^\perp}$, la condition $p^* = p$ équivaut à prendre $F'' = (F')^\perp$. Même raisonnement pour s. □

Endomorphismes orthogonaux et unitaires.

Théorème 6.8. *Soit $u \in \mathcal{L}_{\mathbb{K}}(E,E)$ un endomorphisme d'un espace euclidien ou hermitien E de dimension finie sur $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Les propriétés suivantes sont équivalentes* :

(1) *u préserve la norme* : $\forall x \in E$, $\|u(x)\| = \|x\|$;

(2) *u préserve le produit scalaire* : $\forall x,\, y \in E$, $\langle u(x), u(y)\rangle = \langle x, y\rangle$;

(3) *u est inversible et $u^* = u^{-1} \iff u \circ u^* = u^* \circ u = \mathrm{Id}_E$* ;

(4) *il existe une base orthonormée $(b_j)_{1\leq j\leq n}$ telle que l'image $(u(b_j))$ soit orthonormée* ;

(5) *pour toute base orthonormée $(b_j)_{1\leq j\leq n}$, l'image $(u(b_j))$ est une base orthonormée.*

On dit alors que u est un endomorphisme orthogonal ($\mathbb{K} = \mathbb{R}$)*, resp. un endomorphisme* unitaire ($\mathbb{K} = \mathbb{C}$) *; dans les deux cas, on dit aussi que u est une* isométrie.

Démonstration. Il est clair que (2) $\Rightarrow$ (1) en prenant $x = y$. L'implication (1) $\Rightarrow$ (2) résulte de la formule de polarisation. Par exemple dans le cas réel, l'invariance de la norme implique

$$\begin{aligned}\langle u(x), u(y)\rangle &= \frac{1}{4}\big(\|u(x)+u(y)\|^2 - \|u(x)-u(y)\|^2\big) = \frac{1}{4}\big(\|u(x+y)\|^2 - \|u(x-y)\|^2\big)\\ &= \frac{1}{4}\big(\|x+y\|^2 - \|x-y\|^2\big) = \langle x, y\rangle.\end{aligned}$$

Remarquons maintenant que si l'on remplace x par $u(x)$ dans la formule d'adjonction $\langle u^*(x), y\rangle = \langle x, u(y)\rangle$ il vient

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

On en déduit

$$\langle u(x), u(y)\rangle - \langle x, y\rangle = \langle u^* \circ u(x) - x, y\rangle = \langle (u^* \circ u - \mathrm{Id}_E)(x), y\rangle.$$

Cette expression est nulle pour tous $x, y \in E$ si et seulement si $u^* \circ u - \mathrm{Id}_E = 0$, c'est-à-dire $u^* \circ u = \mathrm{Id}_E$. Mais comme E est de dimension finie, ceci équivaut à dire que u est inversible ($\det(u) \neq 0$) et $u^{-1} = u^*$. On voit ainsi que (2) et (3) sont équivalents.

Montrons maintenant que (2) $\Rightarrow$ (5) : on observe simplement que l'invariance du produit scalaire implique

$$\langle u(b_j), u(b_k)\rangle = \langle b_j, b_k\rangle = \delta_{jk}\ (=1 \ \text{ si } j=k,\ =0 \text{ si } j\neq k).$$

L'implication (5) $\Rightarrow$ (4) est évidente (on utilise tout de même l'existence de bases orthonormées pour tout produit scalaire positif non dégénéré !) Pour boucler la boucle, il reste à voir par exemple que (4) $\Rightarrow$ (1). Or si $x = \sum_{1\leq j\leq n} x_j b_j$, nous avons $\|x\|^2 = \sum_{1\leq j\leq n} |x_j|^2$ puisque la base (b_j) est orthonormée, et l'égalité $u(x) = \sum_{1\leq j\leq n} x_j u(b_j)$ implique de même $\|u(x)\|^2 = \sum_{1\leq j\leq n} |x_j|^2$ si la base $(u(b_j))$ est orthonormée. Si (4) est vérifié, on voit donc que u préserve la norme (propriété (1)). □

Corollaire 6.9. *Soit $(e_1, \ldots, e_n)$ une base orthonormée de E et $A = \mathrm{Mat}_{(e_j)}(u)$ la matrice d'un endomorphisme $u \in \mathcal{L}_{\mathbb{K}}(E,E)$. Alors u est orthogonal, resp. unitaire, si et seulement si la matrice $A = (a_{jk})$ vérifie l'une des conditions équivalentes suivantes* :

(1) *A est inversible et $A^* = A^{-1} \iff AA^* = A^*A = I_n$ (matrice unité $n \times n$)* ;

(2) *les vecteurs colonnes $C_k = (a_{jk})_{1\leq j\leq n}$ forment une base orthonormée de $\mathbb{K}^n$ pour le produit scalaire usuel.*

(3) *les vecteurs lignes $L_j = (a_{jk})_{1\leq k\leq n}$ forment une base orthonormée de $\mathbb{K}^n$ pour le produit scalaire usuel.*

Une telle matrice A est appelée matrice orthogonale (cas réel), resp. unitaire (cas complexe).

Démonstration. On raisonne dans le cas des matrices complexes, qui contient le cas réel. L'équivalence de (1) et (2) résulte de l'équivalence de (3), (4) et (5) du théorème précédent. Maintenant, si A est unitaire, alors $B = A^t$ l'est aussi (et réciproquement) :

$$B^{-1} = (A^{-1})^t = (A^*)^t = (\overline{A^t})^t = (A^t)^* = B^*.$$

Comme les lignes de A sont les colonnes de B, on voit que (2) et (3) sont équivalents. On peut voir aussi que A est unitaire si et seulement si $\overline{A}$ est unitaire. □

Théorèmes spectraux. En dimension finie, le spectre d'un endomorphisme est par définition l'ensemble de ses valeurs propres. La théorie spectrale est l'étude des valeurs propres et vecteurs propres des endomorphismes (éventuellement généralisée à la dimension infinie, afin de prendre en compte des opérateurs tels que l'opérateur de Schrödinger de la mécanique quantique...). Mais il faut bien commencer par le début, et nous allons ici nous contenter de la théorie spectrale des endomorphismes symétriques, anti-symétriques, unitaires et orthogonaux en dimension finie.

Théorème 6.10 (théorème spectral, cas complexe)**.** *Soit E un espace hermitien de dimension finie $n = \dim_{\mathbb{C}} E$, et soit $u \in \mathcal{L}_{\mathbb{C}}(E, E)$.*

(1) *L'endomorphisme u est hermitien (u^*=u) si et seulement u possède une base* orthonormée *$(b_j)_{1\leq j\leq n}$ de vecteurs propres associés à des valeurs propres* réelles *λ_j, i.e. $u(b_j) = \lambda_j b_j$, $\lambda_j \in \mathbb{R}$. On a alors*

$$\mathrm{Mat}_{(b_j)}(u) = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \lambda_j & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix}, \qquad \lambda_j \in \mathbb{R}.$$

(2) *L'endomorphisme u est unitaire ($u^* = u^{-1}$) si et seulement u possède une base* orthonormée *$(b_j)_{1\leq j\leq n}$ de vecteurs propres associés à des valeurs propres λ_j de* module 1*, i.e. $u(b_j) = \lambda_j b_j$, $\lambda_j \in \mathbb{C}$, $|\lambda_j| = 1$. On a alors*

$$\mathrm{Mat}_{(b_j)}(u) = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \lambda_j & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix}, \qquad \lambda_j \in \mathbb{C}, \quad |\lambda_j| = 1.$$

(3) *Dans les deux cas* (1) *et* (2)*, les sous-espaces propres S, S' associés à des valeurs propres λ, λ' distinctes sont orthogonaux entre eux.*

Démonstration. Pour (1) et (2), on raisonne par récurrence sur $n = \dim_{\mathbb{C}} E$. Si $n = 1$ la matrice A de u est déjà diagonale, $A = (\lambda)$ (et tout vecteur non nul est vecteur propre de valeur propre λ !). Nous avons $A^* = (\overline{\lambda})$, donc $A^* = A$ équivaut à $\lambda \in \mathbb{R}$, et la condition $A^*A = (1)$ équivaut à $\overline{\lambda}\lambda = |\lambda|^2 = 1$. Les résultats annoncés sont évidents si $n = 1$.

Supposons maintenant le théorème démontré en dimension $n - 1$. Comme on est sur $\mathbb{C}$, il existe une valeur propre $\lambda \in \mathbb{C}$ (n'importe quelle racine du polynôme caractéristique) et un vecteur propre $v \in E$ associé, soit $v \neq 0$ tel que $u(v) = \lambda v$. Si $u^* = u$, nous avons $\langle u(v), v\rangle = \langle v, u(v)\rangle$, ce qui implique

$$\langle \lambda v, v\rangle = \langle v, \lambda v\rangle \implies \overline{\lambda}\,\|v\|^2 = \lambda\,\|v\|^2 \implies \overline{\lambda} = \lambda \implies \lambda \in \mathbb{R}.$$

Si u est unitaire, nous avons $\|u(v)\| = \|v\|$, donc

$$\|\lambda v\| = \|v\| \implies |\lambda|\,\|v\| = \|v\| \implies |\lambda| = 1.$$

On voit donc que les valeurs propres sont réelles dans le cas u hermitien, et qu'elles sont de module 1 dans le cas u unitaire. Choisissons une telle valeur propre $\lambda = \lambda_1$, un vecteur

propre associé $b_1 = \frac{1}{\|v\|} v$ de norme 1, et soit $D = \mathbb{C}b_1$ la droite engendrée par ce vecteur propre.

Cette droite est stable : $u(D) \subset D$. On sait alors que l'hyperplan orthogonal $H = D^\perp$ est stable par u^*. Dans le cas u hermitien, nous obtenons $u(H) \subset H$, tandis que dans le cas u unitaire nous trouvons $u^{-1}(H) \subset H$. Cette dernière inclusion est une égalité car u^{-1} est bijective et donc $\dim u^{-1}(H) = \dim H = n-1$; par conséquent $u(H) = u(u^{-1}(H)) = H$, de sorte que H est stable dans tous les cas. La restriction $u_{|H}$ est alors un endomorphisme de H. Si u est hermitien (resp. unitaire), il est évident de vérifier que $u_{|H}$ est encore hermitien (resp. unitaire). D'après l'hypothèse de récurrence, l'hyperplan H, qui est de dimension $n-1$, admet une base orthonormée $(b_2, \dots, b_n)$ de vecteurs propres pour $u_{|H}$, avec $u(b_j) = u_{|H}(b_j) = \lambda_j b_j$ pour $j = 2, \dots, n$ et $\lambda_j \in \mathbb{R}$ (resp. $\lambda_j \in \mathbb{C}$, $|\lambda_j| = 1$). Comme $E = \mathbb{C}b_1 \oplus H$, on obtient ainsi une base orthonormée $(b_1, \dots, b_n)$ de E formée de vecteurs propres de u, et le théorème est démontré par récurrence sur n, puisque les conditions énoncées en (1) et (2) sont trivialement suffisantes.

(3) Supposons $u(x) = \lambda x$ et $u(y) = \lambda' y$ avec $x, y \neq 0$ et avec des valeurs propres $\lambda \neq \lambda'$. Alors si $u^* = u$, la relation $\langle u(x), y\rangle = \langle x, u(y)\rangle$ implique

$$\langle \lambda x, y\rangle = \langle x, \lambda' y\rangle \Longrightarrow (\overline{\lambda} - \lambda')\langle x, y\rangle = 0 \Longrightarrow (\lambda - \lambda')\langle x, y\rangle = 0 \Longrightarrow \langle x, y\rangle = 0$$

(on notera que λ est réel dans ce cas). Lorsque u est unitaire, la relation $\langle u(x), u(y)\rangle = \langle x, y\rangle$ implique

$$\langle \lambda x, \lambda' y\rangle = \langle x, y\rangle \Longrightarrow (\overline{\lambda}\lambda' - 1)\langle x, y\rangle = \overline{\lambda}(\lambda' - \lambda)\langle x, y\rangle = 0 \Longrightarrow \langle x, y\rangle = 0$$

(du fait que $|\lambda|^2 = 1$ ici). Dans les deux cas, on voit que les espaces propres S, S' associés à λ et λ' sont orthogonaux. □

Remarque 6.11. Un endomorphisme u est anti-hermitien si et seulement si iu est hermitien. On déduit alors du théorème précédent que u est anti-hermitien si et seulement si u admet une base orthonormée (b_j) de vecteurs propres correspondant à des valeurs propres $\lambda_j \in i\mathbb{R}$ purement imaginaires.

Nous allons maintenant étudier les analogues de ces résultats lorsque $\mathbb{K} = \mathbb{R}$. Le cas symétrique réel est sans changement par rapport au cas hermitien.

Théorème 6.12 (théorème spectral, cas symétrique réel)**.** *Un endomorphisme $u \in \mathcal{L}_{\mathbb{R}}(E, E)$ d'un espace euclidien E de dimension finie n est symétrique si et seulement si u admet une base orthonormée de vecteurs propres $(b_1, \dots, b_n)$ correspondant à des valeurs propres $\lambda_1, \dots, \lambda_n$ réelles. De plus, les espaces propres associés à des valeurs propres distinctes sont orthogonaux.*

Démonstration. Soit $(e_1, \dots, e_n)$ une base orthonormée de E et A la matrice symétrique réelle $n \times n$ qui représente u dans la base (e_j). Comme $\mathbb{R} \subset \mathbb{C}$, on peut aussi considérer A comme une matrice hermitienne $n \times n$. On sait d'après ce qui précède que toutes les valeurs propres sont réelles. Il existe donc des vecteurs propres réels, et on peut faire un raisonnement par récurrence sur la dimension, exactement comme dans le cas hermitien. □

Corollaire 6.13 (interprétation matricielle)**.** *Si A est une matrice symétrique $n \times n$ réelle, il existe une matrice de passage P orthogonale réelle telle que $P^{-1}AP = P^*AP = D$, où D est la matrice diagonale ayant les valeurs propres $\lambda_j \in \mathbb{R}$ de A comme coefficients diagonaux. Un résultat analogue est vrai pour une matrice A complexe hermitienne, avec cette fois une matrice de passage P unitaire.*

Remarque 6.14. Pour diagonaliser une forme quadratique réelle q, il suffit donc de diagonaliser la matrice symétrique A qui la représente en cherchant les valeurs propres et les vecteurs propres. On fera attention au fait que si des vecteurs propres correspondant à des valeurs propres différentes sont bien orthogonaux, en revanche des vecteurs propres correspondant à une valeur propre multiple ne sont pas automatiquement orthogonaux ; si cette dernière situation se produit, il faut rendre ces vecteurs orthogonaux par le procédé de Gram-Schmidt pour obtenir une base orthonormée...

Ce calcul est en général beaucoup plus compliqué que la méthode de Gauss (bien sûr, il peut toujours se produire des miracles, mais en pratique ceux-ci ne sont fréquents que dans les énoncés d'exercices et de problèmes de L2 !) On notera que la méthode de Gauss, quant à elle, ne fournit pas a priori une matrice de passage P orthonormée ; si $\mathbb{K} = \mathbb{Q}$, elle produit une matrice de passage P dans $\mathbb{Q}$, alors que la recherche des valeurs propres conduit en général (de nouveau, sauf miracle !) à des racines irrationnelles du polynôme caractéristique.

Exemple 6.15. Considérons la forme quadratique $q(x, y) = 3x^2 - 2xy + 5y^2$ sur $\mathbb{R}^2$. Sa matrice dans la base canonique de $\mathbb{R}^2$ est

$$A = \begin{pmatrix} 3 & -1 \\ -1 & 5 \end{pmatrix}$$

Le polynôme caractéristique est donné par

$$P(\lambda) = \det \begin{pmatrix} 3-\lambda & -1 \\ -1 & 5-\lambda \end{pmatrix} = (3-\lambda)(5-\lambda) - 1 = \lambda^2 - 8\lambda + 14$$

dont les racines sont $\lambda_1 = 4 + \sqrt{2}$, $\lambda_2 = 4 - \sqrt{2}$. Un calcul de $\mathrm{Ker}(A - \lambda I)$, $\lambda = \lambda_1$ ou λ_2, montre que les vecteurs propres correspondants sont

$$v_1 = \begin{pmatrix} 1 \\ -1-\sqrt{2} \end{pmatrix}, \qquad v_2 = \begin{pmatrix} 1 \\ -1+\sqrt{2} \end{pmatrix}.$$

Cette base (v_1, v_2) est nécessairement orthogonale (ceci résulte du fait que les valeurs propres sont distinctes, les sous-espaces propres étant alors orthogonaux), il suffit de diviser par les normes pour obtenir une base orthonormée de vecteurs propres :

$$b_1 = \frac{1}{\sqrt{4+2\sqrt{2}}} \begin{pmatrix} 1 \\ -1-\sqrt{2} \end{pmatrix}, \qquad b_2 = \frac{1}{\sqrt{4-2\sqrt{2}}} \begin{pmatrix} 1 \\ -1+\sqrt{2} \end{pmatrix}.$$

Dans cette base, la matrice de q devient

$$\mathrm{Mat}_{(b_1,b_2)}(q) = \begin{pmatrix} 4+\sqrt{2} & 0 \\ 0 & 4-\sqrt{2} \end{pmatrix}.$$

La méthode de Gauss donne quant à elle $q(x, y) = 3(x - \frac{1}{3}y)^2 + \frac{14}{3}y^2$, soit un changement de coordonnées $\widetilde{x} = x - \frac{1}{3}y$, $\widetilde{y} = y \Rightarrow x = \widetilde{x} + \frac{1}{3}\widetilde{y}$, et donc une matrice de passage (non orthogonale)

$$P' = \begin{pmatrix} 1 & \frac{1}{3} \\ 0 & 1 \end{pmatrix}$$

vers une base (e'_1, e'_2) dans laquelle

$$\mathrm{Mat}_{(e'_1,e'_2)}(q) = \begin{pmatrix} 3 & 0 \\ 0 & \frac{14}{3} \end{pmatrix}.$$

Théorème 6.16. *Si q, q' sont deux formes quadratiques sur un espace vectoriel E de dimension finie, avec q définie positive, alors il existe une base orthonormée $(b_1, \ldots, b_n)$ qui est orthonormée pour q et orthogonale pour q'.*

Démonstration. C'est juste une reformulation du théorème spectral dans le cas symétrique. On utilise q pour définir une structure d'espace euclidien sur E. Soit $(e_1, \dots, e_n)$ une base orthonormée pour q, et A la matrice de q' dans cette base. Il existe alors un matrice de passage P orthogonale (resp. unitaire) qui diagonalise la matrice A de q'. Ceci fournit une base orthonormée $(b_1, \dots, b_n)$ pour q qui est orthogonale pour q'. □

Théorème 6.17 (théorème spectral, cas orthogonal)**.** *Un endomorphisme $u \in \mathcal{L}_{\mathbb{R}}(E, E)$ d'un espace euclidien E de dimension finie n est orthogonal si et seulement si l'espace E admet une décomposition en somme directe orthogonale*

$$E = \Pi_1 \oplus \dots \oplus \Pi_s \oplus D_1 \oplus \dots \oplus D_t$$

formée de plans Π_j et de droites D_j stables par u, de sorte que $u_{|\Pi_j}$ soit une rotation d'angle θ_j et $u_{|D_j} = \pm \mathrm{Id}_{D_j}$. Choisissons une base orthonormée (a_j, b_j) de Π_j et un vecteur directeur unitaire v_j de D_j. Alors $\mathcal{B} = (a_1, b_1, \dots, a_s, b_s, v_1, \dots, v_t)$ est une base orthonormée de E, et dans cette base la matrice de u est donnée par

$$\widetilde{A} = \mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} \cos\theta_1 & -\sin\theta_1 & 0 & 0 & \dots & 0 & \dots & 0 \\ \sin\theta_1 & \cos\theta_1 & 0 & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & & & & \vdots \\ 0 & 0 & \dots & \cos\theta_s & -\sin\theta_s & 0 & \dots & 0 \\ 0 & 0 & \dots & \sin\theta_s & \cos\theta_s & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \lambda_1 & \dots & 0 \\ \vdots & & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & \lambda_t \end{pmatrix}, \qquad \lambda_j = \pm 1.$$

Autrement dit, si $(e_1, \dots, e_n)$ est une base orthonormée donnée d'avance et $A = \mathrm{Mat}_{(e_j)}(u)$ la matrice orthogonale représentant u, il existe une matrice de passage P (orthogonale) vers une nouvelle base orthonormée, de sorte que $\widetilde{A} = P^{-1}AP = P^tAP$ soit de la forme ci-dessus.

Démonstration. On raisonne par récurrence sur la dimension. Si $n = 1$, le résultat est évident, les seuls endomorphismes d'un espace E de dimension 1 sont les homothéties de rapport $\lambda \in \mathbb{R}$, et il s'agit d'un endomorphisme orthogonal si $\lambda = \pm 1$, d'où $u = \pm \mathrm{Id}_E$.

Si $n = 2$, soit (e_1, e_2) une base orthonormée de E et A la matrice de u dans cette base. Alors $(u(e_1), u(e_2))$ est une base orthonormée et on peut écrire $u(e_1) = \cos\theta e_1 + \sin\theta e_2$ pour un certain angle $\theta \in \mathbb{R}$ (qu'on peut choisir sans $[0, 2\pi[$ si on veut). Le vecteur $u(e_2)$ étant orthogonal à $u(e_1)$ et de norme 1, nous avons seulement deux possibilités, à savoir

$$u(e_2) = -\sin\theta\, e_1 + \cos\theta\, e_2 \quad \text{ou} \quad u(e_2) = \sin\theta\, e_1 - \cos\theta\, e_2,$$

ce qui donne

$$A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \quad \text{ou} \quad A = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

Dans le premier cas, il s'agit d'une rotation d'angle θ (les valeurs propres de A sont les nombres complexes conjugués $\lambda = e^{i\theta}$, $\overline{\lambda} = e^{-i\theta}$), dans le deuxième on vérifie facilement que

$$A = P^{-1}A'P, \qquad P = \begin{pmatrix} \cos\theta/2 & -\sin\theta/2 \\ \sin\theta/2 & \cos\theta/2 \end{pmatrix}, \qquad A' = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

il s'agit d'une symétrie orthogonale par rapport à la droite vectorielle $D_{\theta/2}$ d'angle polaire $\theta/2$, dont les valeurs propres sont $+1$ et -1. Dans les deux cas on se ramène à une base dans laquelle u a une matrice $\widetilde{A} = A$ (resp. $\widetilde{A} = A'$) du type voulu.

Supposons maintenant que le résultat soit connu en dimension $\leq n-1$, avec $n \geq 3$. On sait d'après le Théorème 1.17 que u possède un sous-espace stable S de dimension 1 ou 2 (une droite ou un plan). On a une décomposition orthogonale $E = S \oplus S^\perp$ et on sait que $S^\perp$ est stable par u, de sorte que la restriction $u_{|S^\perp}$ est encore un endomorphisme orthogonal. On obtient alors l'existence d'une décomposition orthogonale de $S^\perp$ vérifiant les conclusions du théorème par l'hypothèse de récurrence, tandis que sur S on applique les observations déjà faites en dimension 1 et 2. □

Corollaire 6.18. *Si $u \in \mathcal{L}_{\mathbb{R}}(E,E)$ est un endomorphisme orthogonal, alors*

$$\det(u) = \lambda_1 \dots \lambda_t = \pm 1.$$

Définition 6.19. *Si $u \in \mathcal{L}_{\mathbb{R}}(E,E)$ est un endomorphisme orthogonal, on dit que u est une* rotation (*ou endomorphisme orthogonal positif*) *si* $\det(u) = +1$ *et un* anti-déplacement (*ou endomorphisme orthogonal négatif*) *si* $\det(u) = -1$. *On note $O(E)$ (resp. $O^+(E)$, $O^-(E)$) l'ensemble des endomorphismes orthogonaux (resp. orthogonaux positifs, négatifs).*

Remarque 6.20. Les matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \text{resp.} \quad \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix},$$

sont des rotation planes d'angle $\theta = 0$, resp. $\theta = \pi$, donc on peut toujours ne faire apparaître qu'au plus une fois les valeurs propres $\lambda_j = 1$ et $\lambda_j = -1$ (de sorte qu'il y a au plus deux droites dans la décomposition en somme directe orthogonale). Si $\dim_{\mathbb{R}} E = 3$, il existe toujours une décomposition orthogonale $E = \Pi \oplus D$ et une base orthonormée $\mathcal{B} = (a, b, v)$ où (a, b) une base orthonormée de Π et v un vecteur directeur de $D = \Pi^\perp$ dans laquelle

$$\mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{ou} \quad \mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

suivant que $u \in O^+(E)$ ou $u \in O^-(E)$. Dans le premier cas u est une rotation d'axe D et d'angle θ, dans le second cas u est la composée de cette rotation avec la symétrie orthogonale $(x, y, z) \mapsto (x, y, -z)$ par rapport au plan $z = 0$. On observera que l'angle θ n'est défini qu'au signe près (un changement d'orientation de la base (a, b) de Π change θ en $-\theta$). Chaque plan Π_j où u est une rotation d'angle θ_j correspond à des valeurs propres complexes conjuguées $\lambda_j = e^{i\theta_j}$, $\overline{\lambda_j} = e^{-i\theta_j}$ de la matrice de u.

Remarque 6.21. En Physique, le mouvement d'un solide est caractérisé par la variation dans le temps d'un repère orthonormé $(M_0(t); \mathcal{B}(t))$, en considérant le déplacement d'un point fixé M_0 du solide et la variation d'un système d'axes orthonormé associé au solide. On s'intéresse ici seulement à la variation de la base $\mathcal{B}(t)$ en fonction du temps.

Choisissons $(e_j) = \mathcal{B}(0)$ comme base de référence. Alors $\mathcal{B}(t)$ est donnée par une matrice de passage orthogonale $P(t)$ telle que $P(0) = I_n$ (matrice unité $n \times n$) ; le fait que la matrice soit orthogonale résulte de l'hypothèse qu'il s'agit d'un solide : les distance mutuelles de ses points ne changent pas, donc la transformation doit préserver la norme des vecteurs.

Maintenant, pour des raisons dues à la Physique, l'application matricielle $t \mapsto P(t)$ doit être continue (et même en général différentiable, la vitesse de déplacement ne peut être infinie...). Ceci entraîne que $t \mapsto \det(P(t))$ est continue. Comme $\det(P(t)) = \pm 1$ et qu'il ne peut y avoir de saut, et on a donc $\det(P(t)) = +1$ pour tout temps t, puisque $\det(P(0)) = \det(I_n) = 1$. Ceci entraîne qu'un mouvement de solides ne peut se faire que par rotations (et ne peut jamais conduire à un endomorphisme orthogonal négatif).

Réciproquement, toute matrice de rotation A peut s'obtenir par un mouvement continu $t \mapsto P(t)$ sur un intervalle de temps unité $[0,1]$, c'est-à-dire tel que $P(0) = I_n$ et $P(1) = A$ (et ceci en dimension n quelconque, à supposer que notre espace ne soit pas de dimension $3\ldots$). Pour le voir, on écrit $A = Q^{-1}(R_{\theta_1,\ldots,\theta_s;\,m})Q$ où Q est orthogonale et où $(R_{\theta_1\ldots\theta_s;\,m})$ est la matrice formée de s blocs de rotations planes d'angles $\theta_1, \ldots, \theta_s$ et de m valeurs propres $+1$ (les valeurs propres -1 sont en nombre pair, on peut les regrouper en rotations planes d'angle $\theta_j = \pi$). On pose alors

$$P(t) = Q^{-1}(R_{t\theta_1,\ldots,t\theta_s;\,m})Q,$$

ceci donne un mouvement continu (et même indéfiniment différentiable) pour $t \in [0,1]$, tel que $P(0) = Q^{-1}I_nQ = I_n$ et $P(1) = A$.

Théorème 6.22 (théorème spectral, cas anti-symétrique)**.** *Un endomorphisme $u \in \mathcal{L}_{\mathbb{R}}(E,E)$ d'un espace euclidien E de dimension finie n est anti-symétrique si et seulement si l'espace E admet une décomposition en somme directe orthogonale*

$$E = \Pi_1 \oplus \ldots \oplus \Pi_s \oplus K$$

formée de plans Π_j stables par u et du noyau $K = \mathrm{Ker}(u)$. Choisissons une base orthonormée (a_j, b_j) de Π_j et une base orthonormée $(v_1, ..., v_t)$ de K. Alors $\mathcal{B} = (a_1, b_1, \ldots, a_s, b_s, v_1, \ldots, v_t)$ est une base orthonormée de E, et quitte à changer b_j en $-b_j$, la matrice de u se met sous la forme

$$\widetilde{A} = \mathrm{Mat}_{\mathcal{B}}(u) = \begin{pmatrix} 0 & -\alpha_1 & 0 & 0 & \ldots & 0 & \ldots & 0 \\ \alpha_1 & 0 & 0 & 0 & \ldots & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & & & & \vdots \\ 0 & 0 & \ldots & 0 & -\alpha_s & 0 & \ldots & 0 \\ 0 & 0 & \ldots & \alpha_s & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\ \vdots & & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \end{pmatrix}, \quad \alpha_j > 0, \quad r = \mathrm{rang}(u) = 2s.$$

Autrement dit, si $(e_1, \ldots, e_n)$ est une base orthonormée donnée d'avance et $A = \mathrm{Mat}_{(e_j)}(u)$ la matrice anti-symétrique représentant u, il existe une matrice de passage P (orthogonale) vers une nouvelle base orthonormée, de sorte que $\widetilde{A} = P^{-1}AP = P^tAP$ soit de la forme ci-dessus.

Démonstration. Le sous-espace $K = \mathrm{Ker}\, u$ étant stable, son orthogonal $K^\perp$ est stable par $u^* = -u$, donc également par u. La restriction $u_{|K^\perp}$ est encore anti-symétrique, on sait que sa matrice n'admet que des valeurs propres purement imaginaires $\lambda_j = i\alpha_j$, $\alpha_j \in \mathbb{R}$, $\alpha_j \neq 0$ (on ne peut avoir $\lambda_j = 0$, sinon on obtiendrait un vecteur non nul du noyau dans $K^\perp$, ce qui est impossible puisque $K \cap K^\perp = \{0\}$). Si $K = E$, on a $u = 0$ et il n'y a rien à faire. Sinon, prenons par exemple la valeur propre $\lambda_1 = i\alpha_1$, on obtient un plan stable Π_1 dans lequel la matrice associée est antisymétrique. Ayant choisi une base orthonormée (a_1, b_1) de Π_1, la seule possibilité est une matrice de la forme

$$\mathrm{Mat}_{(a_1,b_1)}(u_{|\Pi_1}) = \begin{pmatrix} 0 & -\alpha_1 \\ \alpha_1 & 0 \end{pmatrix}$$

dont les valeurs propres sont précisément $\lambda_1 = i\alpha_1$, $\overline{\lambda}_1 = -i\alpha_1$ (il se peut qu'on tombe sur la matrice opposée, mais dans ce cas il suffit de remplacer (a_1, b_1) par $(a_1, -b_1)$ pour changer les signes, et on peut donc se ramener à $\alpha_1 > 0$). On raisonne maintenant par récurrence sur la dimension de E, en considérant une décomposition orthogonale $E = \Pi_1 \oplus S$. Le noyau K

est nécessairement contenu dans S, et on applique l'hypothèse de récurrence pour voir que S admet une décomposition $S = \Pi_2 \oplus \ldots \oplus \Pi_s \oplus K$ comme souhaité. □

7. Coniques et quadriques

On va maintenant appliquer la théorie des formes quadratiques à l'étude des coniques et des quadriques. On se place dans l'espace $\mathbb{R}^n$ muni de son produit scalaire usuel, avec $n = 2$ ou $n = 3$ (sauf dans le tout dernier paragraphe où n sera quelconque).

Définition 7.1. *Une* conique $\mathcal{C}$ *est le lieu géométrique de* $\mathbb{R}^2$ *défini par une équation polynomiale* $P(x, y)$ *du second degré en les coordonnées* (x, y)*, c'est-à-dire une équation de la forme*

$$P(x, y) = ax^2 + bxy + cy^2 + dx + ey + f = 0.$$

On peut écrire une telle équation sous la forme

$$q(x, y) + \ell(x, y) + f = 0$$

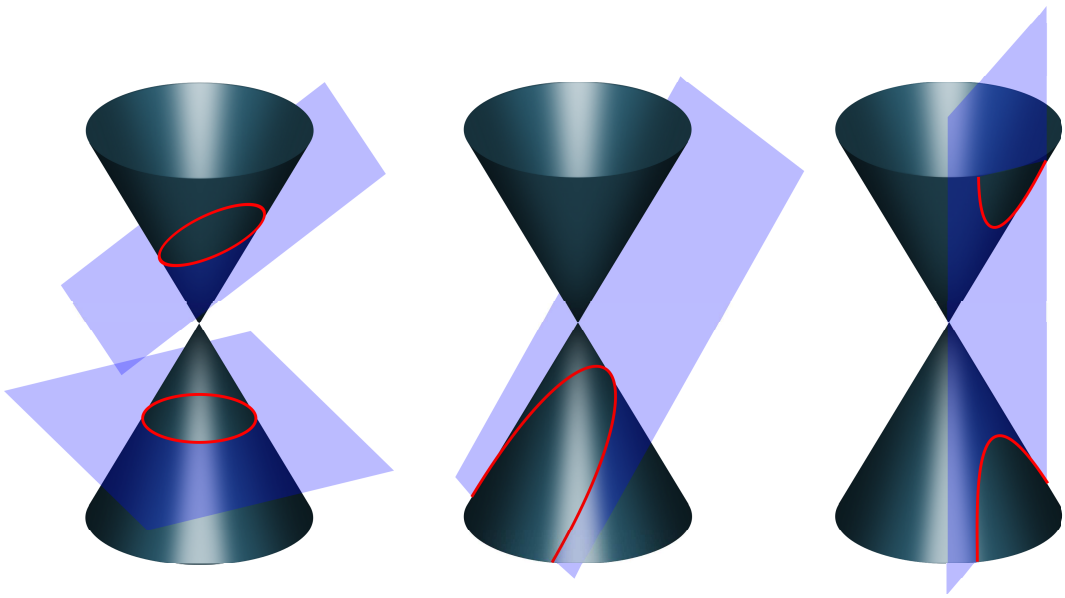
où $q(x, y) = ax^2 + bxy + cy^2$ est une forme quadratique sur $\mathbb{R}^2$, $\ell(x, y) = dx + ey$ une forme linéaire et f une constante. Il est utile également de plonger le plan dans l'espace $\mathbb{R}^3$ et d'introduire le polynôme "homogénéïsé" de degré 2

$$Q(x, y, z) = ax^2 + bxy + cy^2 + dxz + eyz + fz^2.$$

Ce polynôme est une forme quadratique sur $\mathbb{R}^3$, et on voit que $P(x, y) = Q(x, y, 1)$. La conique $\mathcal{C}$ est donc l'intersection du cône isotrope $Q(x, y, z) = 0$ avec le plan affine $\Pi = \{z = 1\}$. Or, en diagonalisant Q par un calcul de valeurs propres, on trouve de nouvelles coordonnées orthonormées $(\widetilde{x}, \widetilde{y}, \widetilde{z})$ dans lesquelles les équations deviennent

$$Q(x, y, z) = \alpha\widetilde{x}^2 + \beta\widetilde{y}^2 + \gamma\widetilde{z}^2 = 0, \qquad \Pi : u\widetilde{x} + v\widetilde{y} + w\widetilde{z} = 1.$$

Ceci permet de représenter classiquement les coniques comme des sections planes de cônes ; en fonction de la position relative du plan et du cône, on peut obtenir cercles, ellipses, paraboles et hyperboles (cette description était déjà connue des Grecs…) :



Cercle, Ellipse **Parabole** **Hyperbole**

Nous allons maintenant classer les coniques à isométrie près, au moyen de diagonalisations et de réductions successives de l'équation $P(x, y) = 0$.

Classification euclidienne des coniques. On cherche pour cela une **équation réduite** de la conique. On commence par diagonaliser la forme quadratique $q(x,y) = ax^2 + bxy + cy^2$ dans un repère orthonormé ; on pourrait calculer valeurs propres et vecteurs propres, mais en dimension 2 il est plus rapide d'effectuer directement une rotation sous la forme

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} u & -v \\ v & u \end{pmatrix} \begin{pmatrix} \widetilde{x} \\ \widetilde{y} \end{pmatrix} \iff \begin{cases} x = u\widetilde{x} - v\widetilde{y} \\ y = v\widetilde{x} + u\widetilde{y} \end{cases} .$$

Après substitution dans $P(x,y) = q(x,y) + dx + ey + f$ il vient

$$\begin{aligned} P(x,y) &= a(u\widetilde{x} - v\widetilde{y})^2 + b(u\widetilde{x} - v\widetilde{y})(v\widetilde{x} + u\widetilde{y}) + c(v\widetilde{x} + u\widetilde{y})^2 + d(u\widetilde{x} - v\widetilde{y}) + e(v\widetilde{x} + u\widetilde{y}) + f \\ &= (au^2 + buv + cv^2)\widetilde{x}^{\,2} + ((c-a)2uv + b(u^2 - v^2))\widetilde{x}\,\widetilde{y} + (av^2 - buv + cu^2)\widetilde{y}^{\,2} \\ &\qquad + (du + ev)\widetilde{x} + (-dv + eu)\widetilde{y} + f. \end{aligned}$$

On veut rendre nul le coefficient $(c-a)2uv + b(u^2 - v^2)$ de $\widetilde{x}\,\widetilde{y}$. En posant $u = \cos\theta$, $v = \sin\theta$, ceci donne la condition $(c-a)\sin(2\theta) + b\cos(2\theta) = 0$, soit $\tan(2\theta) = -b/(c-a)$ si $a \neq c$, et (par exemple) $u = v = 1/\sqrt{2}$ si $a = c$. On est ainsi ramené à une expression de la forme

$$(*) \qquad P(x,y) = \alpha\widetilde{x}^{\,2} + \beta\widetilde{y}^{\,2} + \gamma\widetilde{x} + \delta\widetilde{y} + f.$$

On a ici $\alpha\beta = \det(q) = ac - b^2/4$, et α, β sont précisément les valeurs propres de q. Si q est *non dégénérée*, c'est-à-dire $\det(q) = \alpha\beta \neq 0$, on peut éliminer la partie linéaire $\ell(x,y) = \gamma\widetilde{x} + \delta\widetilde{y}$ en observant que

$$(**) \qquad P(x,y) = \alpha\Big(\widetilde{x} + \frac{\gamma}{2\alpha}\Big)^2 + \beta\Big(\widetilde{y} + \frac{\delta}{2\beta}\Big)^2 + f - \frac{\gamma^2}{4\alpha} - \frac{\delta^2}{4\beta} = \alpha X^2 + \beta Y^2 + \varepsilon$$

et en posant

$$\widetilde{x}_0 = -\frac{\gamma}{2\alpha}, \quad \widetilde{y}_0 = -\frac{\delta}{2\beta}, \qquad X = \widetilde{x} - \widetilde{x}_0, \quad Y = \widetilde{y} - \widetilde{y}_0.$$

On aboutit ainsi à **l'équation réduite**

$$\alpha X^2 + \beta Y^2 + \varepsilon = 0.$$

La conique $\alpha X^2 + \beta Y^2 + \varepsilon = 0$ admet le point $\omega : (X,Y) = (0,0)$ comme centre de symétrie et les axes $X = 0$, $Y = 0$ comme axes de symétries, on dit qu'il s'agit d'une *conique à centre*. Dans les anciennes coordonnées, le centre est donné par $(\widetilde{x}_0, \widetilde{y}_0) = (-\frac{\gamma}{2\alpha}, -\frac{\delta}{2\beta})$, soit $(x_0, y_0) = (u\widetilde{x}_0 - v\widetilde{y}_0, v\widetilde{x}_0 + u\widetilde{y}_0)$, est les axes sont les droites de vecteurs directeurs $(-v,u)$ et (u,v), soit

$$u(x - x_0) + v(y - y_0) = 0, \qquad -v(x - x_0) + u(y - y_0) = 0.$$

Pour calculer directement le centre dans les coordonnées initiales, on peut aussi observer que

$$dP = \frac{\partial P}{\partial x}dx + \frac{\partial P}{\partial y}dy = (2ax + by + d)dx + (bx + 2cy + e)dy = 2\alpha X\,dX + 2\beta Y\,dY,$$

le centre $(X,Y) = (0,0)$ est caractérisé par la propriété $dP = 0$, ce qui amène à résoudre le système d'équations

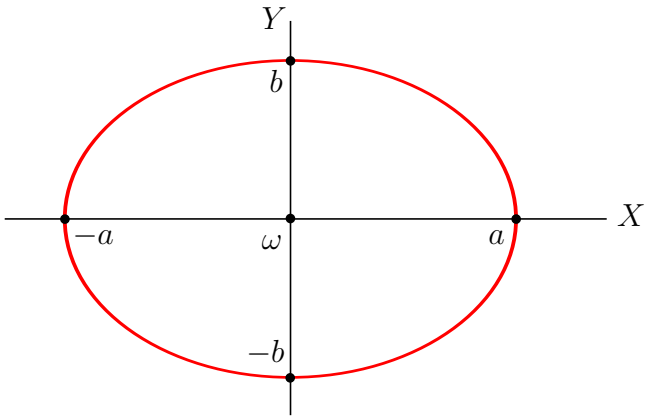
$$\begin{cases} \dfrac{\partial P}{\partial x} = 2ax + by + d = 0 \\ \dfrac{\partial P}{\partial y} = bx + 2cy + e = 0, \end{cases}$$

dont la solution fournit le centre $\omega : (x_0, y_0)$ cherché.

Maintenant, si on divise $(**)$ par $\pm\varepsilon$ (en supposant $\varepsilon \neq 0$), on est ramené à l'un des cas suivants.

Cas où q est définie positive ou négative $\Leftrightarrow \alpha\beta = \det(q) > 0$**, ce qui équivaut à** $\alpha, \beta > 0$ **[signature (2,0)] ou** $\alpha, \beta < 0$ **[signature (0,2)]**. On trouve alors une équation de l'un des 3 types suivants.

(a) $\dfrac{X^2}{a^2} + \dfrac{Y^2}{b^2} = 1$: c'est une ellipse de demi-axes a, b, ou un cercle si $a = b$.



Équations paramétriques :

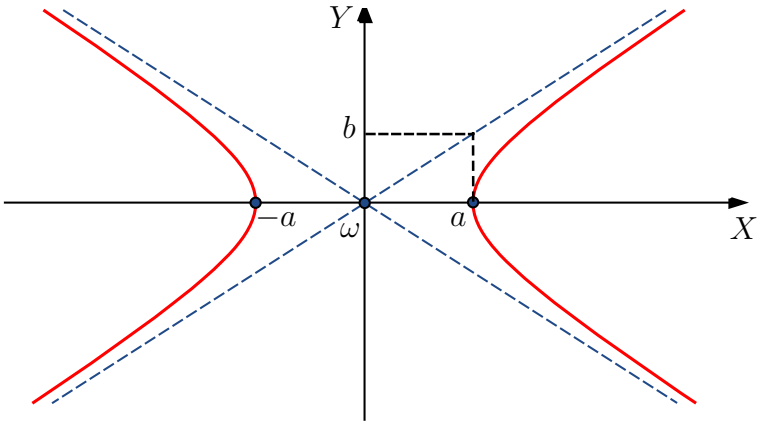
$$\begin{cases} X = a\cos(t) \\ Y = b\sin(t). \end{cases}$$

(b) $\dfrac{X^2}{a^2} + \dfrac{Y^2}{b^2} = -1$: c'est l'ensemble vide $\emptyset$.

(c) $\dfrac{X^2}{a^2} + \dfrac{Y^2}{b^2} = 0$: cas dégénéré d'une ellipse réduite à son centre.

Cas où q est de signature (1,1). Quitte à permuter les coordonnées, on aboutit à une équation de l'un des 2 types suivants.

(a) $\dfrac{X^2}{a^2} - \dfrac{Y^2}{b^2} = 1$: c'est une hyperbole de paramètres a, b, équilatère si $a = b$.



Équations paramétriques :

$$\begin{cases} X = a\cosh(t) \\ Y = b\sinh(t). \end{cases}$$

Le changement de coordonnées (non orthonormé) $U = \frac{X}{a} + \frac{Y}{b}$, $V = \frac{X}{a} - \frac{Y}{b}$ ramène l'hyperbole à l'équation classique $UV = 1$, les asymptotes sont les axes $U = \frac{X}{a} + \frac{Y}{b} = 0$, $V = \frac{X}{a} - \frac{Y}{b} = 0$.

(b) $\dfrac{X^2}{a^2} - \dfrac{Y^2}{b^2} = 0$: cas d'une hyperbole dégénérée en ses 2 asymptotes.

Cas où q est de rang 1. Dans ce cas $q(x, y) = \alpha\widetilde{x}^2$, $\alpha \neq 0$, et on peut éliminer le terme linéaire $\gamma\widetilde{x}$ par changement d'origine. Ceci conduit au deux cas suivants suivant que le terme linéaire restant contient Y ou non :

(a) $\alpha X^2 + \gamma Y = 0$, $\gamma \neq 0$: il s'agit d'une parabole $Y = \lambda X^2$.

(b) $\alpha X^2 + \varepsilon = 0$: il s'agit de deux droites parallèles $X = \pm\mu$, éventuellement confondues, ou de l'ensemble vide $\emptyset$.

Cas où q est de rang 0. Dans le cas dégénéré où $q(x, y) = 0$, l'équation se réduit à une équation affine $dx + ey + f = 0$. On obtient une droite si $(d, e) \neq (0, 0)$, et $\emptyset$ ou $\mathbb{R}^2$ si $(d, e) = (0, 0)$, suivant que $f \neq 0$ ou $f = 0$.

Remarque 7.2. Considérons le cône de révolution $\alpha^2(x^2 + y^2) - z^2 = 0$, $\alpha > 0$. Le lecteur vérifiera facilement que l'intersection de ce cône avec le plan $z = \beta x + \gamma$ de pente $\beta \geq 0$ est une ellipse si $\beta < \alpha$, une parabole si $\beta = \alpha$ et une hyperbole si $\beta > \alpha$ (si $\gamma = 0$, il s'agit de cas dégénérés, la conique se réduit à un point, une droite ou deux droites sécantes).

Description des coniques à l'aide des foyers.

Considérons deux points distincts F, F' du plan, et notons $c = \frac{1}{2}d(F, F')$ leur demi-distance. Si l'on choisit un repère dont l'origine ω est le milieu du segment $[F, F']$ et dont l'axe ωx est porté par la droite (FF'), on peut supposer que $F = (0, c)$ et $F' = (0, -c)$.

Cherchons d'abord l'ensemble des points $M = (x, y) \in \mathbb{R}^2$ tels que $d(M, F) + d(M, F') = 2a$, avec $a > c$. La condition s'écrit

$$\sqrt{(x-c)^2 + y^2} + \sqrt{(x+c)^2 + y^2} = 2a,$$

ce qui implique

$$(x+c)^2 + y^2 = \big(2a - \sqrt{(x-c)^2 + y^2}\big)^2 = 4a^2 - 4a\sqrt{(x-c)^2 + y^2} + (x-c)^2 + y^2,$$

ou encore (après transposition et division par 4) :

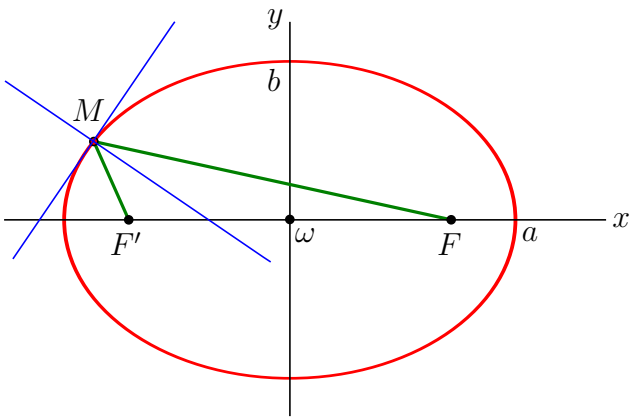
$$a\sqrt{(x-c)^2 + y^2} = a^2 - cx \quad \Longrightarrow a^2\big((x-c)^2 + y^2\big) = (a^2 - cx)^2 = a^4 - 2a^2cx + c^2x^2.$$

Cette dernière équation se ramène à celle d'une **ellipse**, car en posant $b^2 = a^2 - c^2$, c'est-à-dire $b = \sqrt{a^2 - c^2}$, on peut la récrire

$$(a^2 - c^2)x^2 + a^2y^2 = a^4 - a^2c^2 = a^2(a^2 - c^2) \quad \Longleftrightarrow \quad \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1.$$

[On notera que cette équation entraîne en fait $|x| \leq a$ et $|y| \leq b$, donc $a^2 - cx \geq a^2 - ac > 0$ et $\sqrt{(x-c)^2 + y^2} \leq \sqrt{(a+c)^2 + b^2} = \sqrt{2a^2 + 2ac} < 2a$, et les deux implications invoquées précédemment sont alors bien des équivalences]. L'**excentricité** de l'ellipse est par définition

$$e = \frac{c}{a} = \frac{\sqrt{a^2 - b^2}}{a} = \sqrt{1 - b^2/a^2} \quad \in \ [0, 1[.$$



On peut démontrer que la tangente et la normale à l'ellipse au point M sont les bissectrices des droites (MF), (MF').

Cherchons maintenant l'ensemble des points $M=(x,y)$ tels que $\big|d(M,F)-d(M,F')\big|=2a$, avec $a<c$ (l'inégalité triangulaire impose en fait $2a\le d(F,F')=2c$). La condition s'écrit

$$\sqrt{(x+c)^2+y^2}=\sqrt{(x-c)^2+y^2}\pm 2a$$

ce qui implique(†)

$$(x+c)^2+y^2=\big(\sqrt{(x-c)^2+y^2}\pm 2a\big)^2=(x-c)^2+y^2\pm 4a\sqrt{(x-c)^2+y^2}+4a^2,$$

ou encore (après transposition et division par 4) :

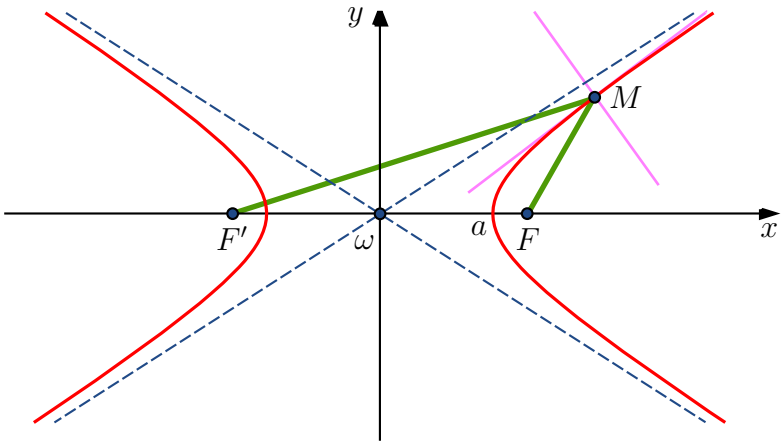
$$cx-a^2=\pm a\sqrt{(x-c)^2+y^2}\quad\Longleftrightarrow a^2\big((x-c)^2+y^2\big)=(cx-a^2)^2=c^2x^2-2a^2cx+a^4.$$

Cette dernière équation se ramène à celle d'une **hyperbole**, car en posant $b^2=c^2-a^2$, c'est-à-dire $b=\sqrt{c^2-a^2}$, on peut la récrire

$$(c^2-a^2)x^2-a^2y^2=a^2c^2-a^4=a^2(c^2-a^2)\quad\Longleftrightarrow\quad\frac{x^2}{a^2}-\frac{y^2}{b^2}=1.$$

[On notera que si par exemple $x\ge 0$, cette équation entraîne en fait $x\ge a$ et donc $\sqrt{(x+c)^2+y^2}\ge a+c>2a$, de sorte que la première implication (†) est bien une équivalence : on ne peut avoir l'égalité parasite éventuelle $\sqrt{(x+c)^2+y^2}=2a-\sqrt{(x-c)^2+y^2}\le 2a$].
L'**excentricité** de l'hyperbole est par définition

$$e=\frac{c}{a}=\frac{\sqrt{a^2+b^2}}{a}=\sqrt{1+b^2/a^2}\quad\in\,]1,+\infty[.$$



Ici encore, la tangente et la normale en M à l'hyperbole sont les bissectrices des droites (MF), (MF') (exercice !)

Définition monofocale des coniques. Étant donné une droite Δ (directrice), un point $F\notin\Delta$ (foyer) et un réel $e>0$ (excentricité), l'ensemble des points M tels que

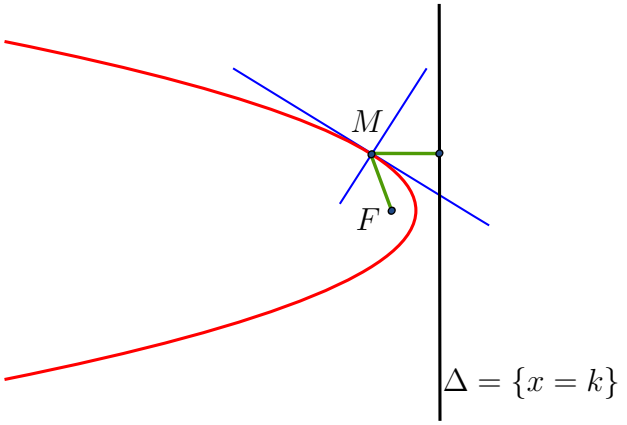
$$d(M,F)=e\,d(M,\Delta)$$

est une conique. Pour le voir, on peut par exemple choisir un repère dans lequel $F=O$ est l'origine et $\Delta=\{x=k\}$, $k>0$. La condition devient $\sqrt{x^2+y^2}=e|x-k|$, soit $x^2+y^2=e^2(x-k)^2$ ou encore

$$(1-e^2)x^2+2e^2kx+y^2=e^2k^2.$$

Si $e = 1$, il s'agit d'une parabole $2kx + y^2 = k^2 \Leftrightarrow x = \frac{k}{2} - \frac{1}{2k}y^2$. Si $e \neq 1$, c'est une conique de centre $\omega = (-\frac{e^2k}{1-e^2}, 0)$, à savoir une ellipse de demi-axes $a = ek/(1-e^2)$, $b = ek/\sqrt{1-e^2}$ si $e < 1$, et une hyperbole de demi-axes $a = ek/(e^2-1)$, $b = ek/\sqrt{e^2-1}$ si $e > 1$. On vérifie facilement que l'équation polaire de la conique est donnée par

$$r = \frac{p}{1 + e\cos\theta} \quad \text{où} \quad p = ek.$$



Parabole

Excentricité $e = 1$

Description des quadriques de $\mathbb{R}^3$.

Une quadrique est par définition l'ensemble des points $(x, y, z) \in \mathbb{R}^3$ défini par une équation du second degré en 3 variables, c'est-à-dire

$$P(x, y, z) = q(x, y, z) + \ell(x, y, z) + c = 0$$

où q est une forme quadratique, ℓ une forme linéaire sur $\mathbb{R}^3$ et c une constante.

Comme précédemment, on va classer les quadriques à isométrie près, et pour cela, on cherche à obtenir une équation réduite. On commence par diagonaliser q dans une base orhonormée : cette fois, il convient de calculer les valeurs propres et α, β, γ et les vecteurs propres par la méthode générale. On trouve alors de nouvelle coordonnées $(\widetilde{x}, \widetilde{y}, \widetilde{z})$ dans lesquelles

$$P(x, y, z) = \alpha\widetilde{x}^2 + \beta\widetilde{y}^2 + \gamma\widetilde{z}^2 + \delta\widetilde{x} + \varepsilon\widetilde{y} + \theta\widetilde{z} + c.$$

Cas où $\det(q) = \alpha\beta\gamma \neq 0$ **(forme quadratique q non dégénérée).** On peut alors éliminer la partie linéaire et il s'agit d'une quadrique à centre. Le centre (x_0, y_0, z_0) peut s'obtenir directement en résolvant les équations linéaires

$$\frac{\partial P}{\partial x} = \frac{\partial P}{\partial y} = \frac{\partial P}{\partial z} = 0,$$

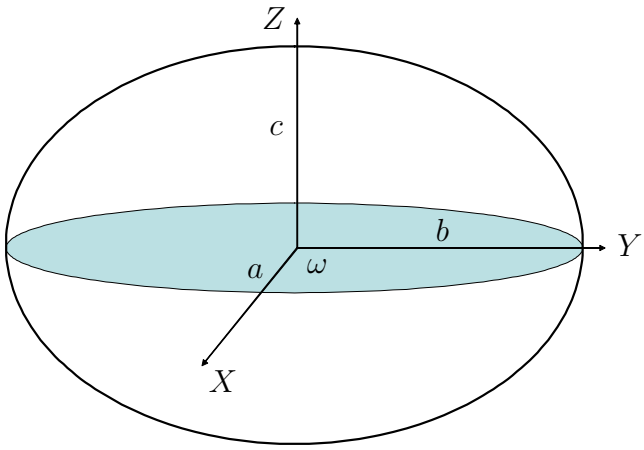
et on introduit alors les nouvelles coordonnées

$$X = \widetilde{x} - \widetilde{x}_0, \quad Y = \widetilde{y} - \widetilde{y}_0, \quad Z = \widetilde{z} - \widetilde{z}_0.$$

Après division par la constante restante (si elle est non nulle), changement de signe et permutation éventuelle des coordonnées X, Y, Z, on obtient une **équation réduite** d'un des types suivants.

(1) Signature $(3, 0)$ ou $(0, 3)$. On a 3 cas :

(a) $\dfrac{X^2}{a^2} + \dfrac{Y^2}{b^2} + \dfrac{Z^2}{c^2} = 1$: c'est un ellipsoïde (ou une sphère si $a = b = c$).



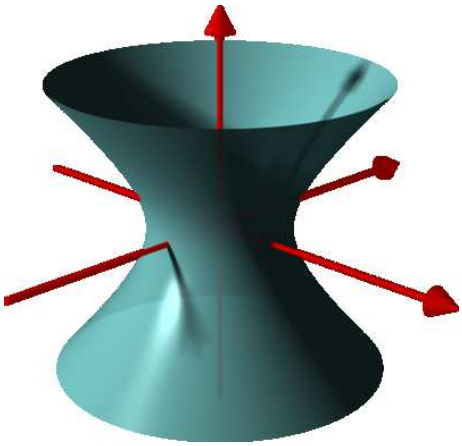
(b) $\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}+\dfrac{Z^2}{c^2}=-1$: c'est l'ensemble vide $\emptyset$.

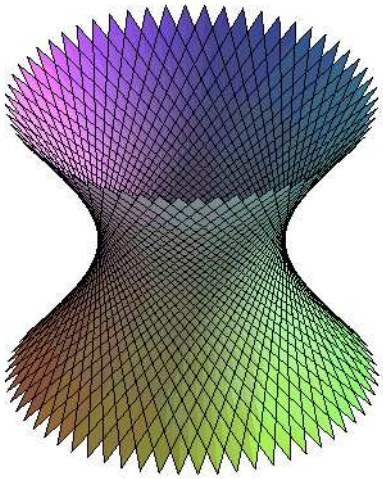
(c) $\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}+\dfrac{Z^2}{c^2}=0$: cas dégénéré d'un ellipsoïde réduit à son centre.

(2) Signature $(2,1)$ **ou** $(1,2)$**.** On a 3 cas :

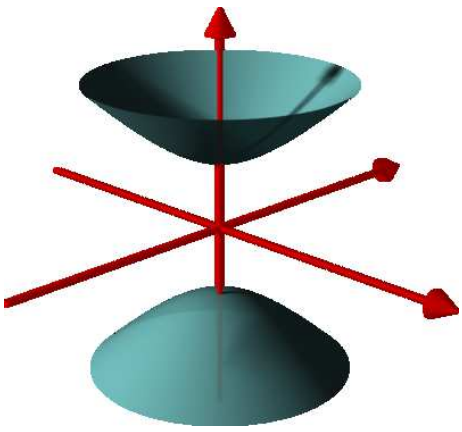
(a) $\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}-\dfrac{Z^2}{c^2}=1$: c'est un hyperboloïde à une nappe

On peut vérifier qu'un tel hyperboloïde est engendré par deux familles de droites :





(b) $\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}-\dfrac{Z^2}{c^2}=-1$: c'est un hyperboloïde à deux nappes $Z=\pm c\sqrt{1+\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}}$

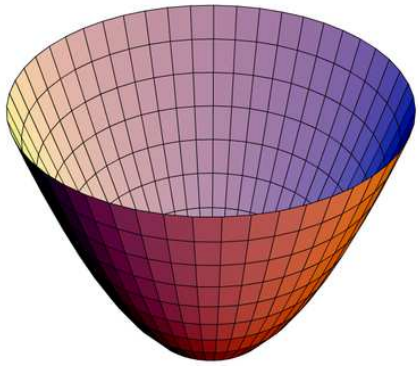


(c) $\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}-\dfrac{Z^2}{c^2}=0$ (constante nulle) : c'est un cône elliptique ou circulaire.

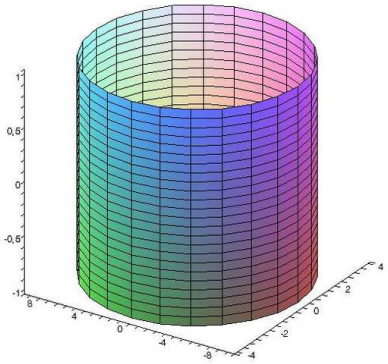
Cas où la forme quadratique q est de rang <3. On distingue ces cas en fonction de la signature de q et de la présence d'une partie linéaire ℓ ou d'une constante c non nulle. Par changement d'origine, on peut éliminer dans ℓ celles des coordonnées $\widetilde{x},\widetilde{y},\widetilde{z}$ qui correspondent à des valeurs propres non nulles de q. Après isométrie et permutation éventuelle des coordonnées on aboutit aux situations suivantes.

(3) Signature $(2,0)$ **ou** $(0,2)$. On a 4 cas :

(a) $\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}=Z$: c'est un paraboloïde elliptique (ou de révolution si $a=b$).



(b) $\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}=1$: c'est un cyclindre elliptique



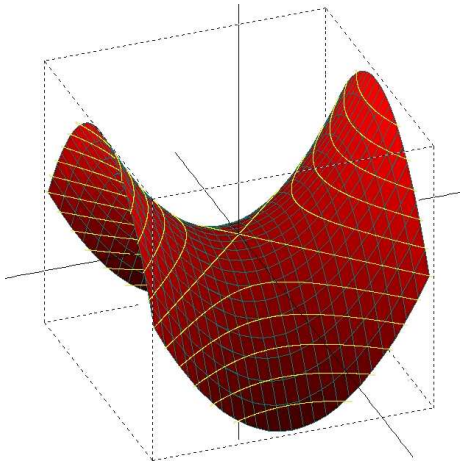
(c) $\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}=0$ (cas dégénéré) : c'est l'axe OZ.

(d) $\dfrac{X^2}{a^2}+\dfrac{Y^2}{b^2}=-1$: c'est l'ensemble vide $\emptyset$.

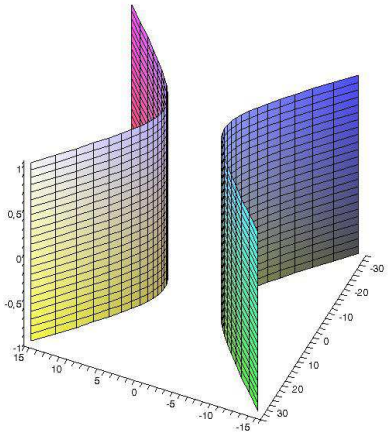
(4) Signature $(1,1)$. Il y a 3 cas :

(a) $\dfrac{X^2}{a^2}-\dfrac{Y^2}{b^2}=Z$:

c'est un paraboloïde hyperbolique.



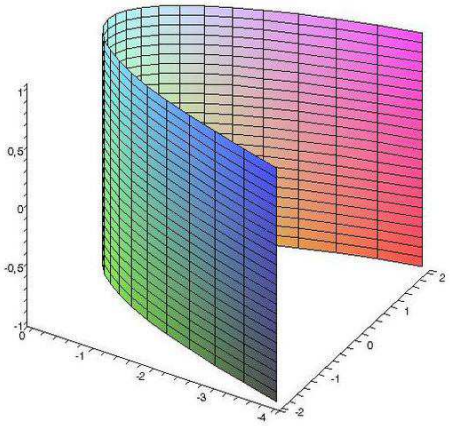
(b) $\dfrac{X^2}{a^2} - \dfrac{Y^2}{b^2} = 1$: c'est un cyclindre hyperbolique.



(c) $\dfrac{X^2}{a^2} - \dfrac{Y^2}{b^2} = 0$: il s'agit de la réunion des 2 plans verticaux sécants $\dfrac{X}{a} \pm \dfrac{Y}{b} = 0$.

(5) Signature $(1,0)$ **ou** $(0,1)$. Nous avons deux cas :

(a) $X^2 = 2pY$: cylindre parabolique.



(b) $X^2 = \lambda$: plan parallèles ou confondus ou $\emptyset$.

(6) Signature (0,0). Dans ce cas très dégénéré on retombe sur une équation affine

$$\ell(x, y, z) + c = 0.$$

Il s'agit d'un plan, de $\mathbb{R}^3$ tout entier ou de l'ensemble vide $\emptyset$.

Exemple 7.3. Soit la quadrique

$$x^2 + y^2 + z^2 + 2xy + 2xz + 2yz + \sqrt{3}x + \sqrt{3}y + 2 = 0.$$

La partie quadratique de l'équation est $q(x, y, z) = x^2+y^2+z^2+2xy+2xz+2yz$; on s'aperçoit aussitôt qu'il s'agit d'une forme quadratique de rang 1 donnée par $q(x, y, z) = (x + y + z)^2$. Pour trouver une base orthonormée de vecteurs propres il suffit de considérer le vecteur

$$\widetilde{e}_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

normal au plan $\Pi = \{x+y+z=0\}$ et une base orthonormée $(\widetilde{e}_2, \widetilde{e}_3)$ de Π, ce qui donne

$$\widetilde{e}_1 = \frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}, \quad \widetilde{e}_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}, \quad \widetilde{e}_3 = \sqrt{\frac{2}{3}}\begin{pmatrix}\frac{1}{2}\\\frac{1}{2}\\-1\end{pmatrix},$$

les vecteurs étant respectivement des vecteurs propres pour les valeurs propres 3, 0 et 0. Soient $\widetilde{x}, \widetilde{y}, \widetilde{z}$ les coordonnées dans la nouvelle base. Par construction de cette base, la forme $q(x,y,z)$ s'écrit $3\,\widetilde{x}^2$. Elle est donc de signature $(1,0)$.

Comme $x = \frac{1}{\sqrt{3}}\widetilde{x} + \frac{1}{\sqrt{2}}\widetilde{y} + \frac{1}{\sqrt{6}}\widetilde{z}$ et $y = \frac{1}{\sqrt{3}}\widetilde{x} - \frac{1}{\sqrt{2}}\widetilde{y} + \frac{1}{\sqrt{6}}\widetilde{z}$, on vérifie que l'équation de cette quadrique dans la nouvelle base est

$$3\,\widetilde{x}^2 + 2\widetilde{x} + \sqrt{2}\,\widetilde{z} + 2 = 0,$$

soit

$$3\Big(\widetilde{x} + \frac{1}{3}\Big)^2 + \sqrt{2}\,\widetilde{z} + \frac{5}{3} = 0.$$

Si on pose $X = \widetilde{x} + \frac{1}{3}$, $Y = \widetilde{z} + \frac{5}{3\sqrt{2}}$, $Z = \widetilde{y}$, on obtient l'équation réduite

$$X^2 = -\sqrt{2}Y,$$

il s'agit d'un cylindre parabolique.

Généralisation : quadriques de $\mathbb{R}^n$. De nouveau, on définit une quadrique de $\mathbb{R}^n$ comme étant l'ensemble des points $x = (x_1, \ldots, x_n)$ satisfaisant une équation du second degré

$$P(x) = q(x) + \ell(x) + c,$$

où q est une forme quadratique, ℓ une forme linéaire et c une constante. Pour trouver une équation réduite, on commence par diagonaliser q en recherchant une base orthonormée $(\widetilde{e}_1, \ldots, \widetilde{e}_n)$ de $\mathbb{R}^n$ dans laquelle les nouvelles coordonnées $\widetilde{x} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$ fournissent

$$q(x) = \widetilde{q}(\widetilde{x}) = \sum_{i=1}^{r} \alpha_i \widetilde{x}_i^2, \qquad \text{où } r = \operatorname{rang}(q), \quad \alpha_i \neq 0.$$

On peut décrire géométriquement cette situation à l'aide d'une somme directe

$$\mathbb{R}^n = S \oplus K \quad \text{où} \quad K = \operatorname{Ker}(q) = \operatorname{Vect}(\widetilde{e}_{r+1}, \ldots, \widetilde{e}_n), \quad S = K^\perp = \operatorname{Vect}(\widetilde{e}_1, \ldots, \widetilde{e}_r).$$

Après avoir réexprimé $\ell(x)$ en termes des $\widetilde{x}_i$, on voit comme précédemment que grâce à une translation $X_i = \widetilde{x}_i - \widetilde{x}_i^0$ dans S $(1 \leq i \leq r)$, il est possible d'éliminer les variables $\widetilde{x}_1, \ldots, \widetilde{x}_r$ de $\ell(x)$ en les intégrant dans les carrés de $\widetilde{q}(\widetilde{x})$. Ceci fournit une forme simplifiée

$$P(x) = \widetilde{q}(X) + \widetilde{l}(\widetilde{x}) + \widetilde{c} \qquad \text{avec} \qquad \widetilde{\ell}(\widetilde{x}) = \beta_{r+1}\widetilde{x}_{r+1} + \ldots + \beta_n \widetilde{x}_n$$

dans les nouvelles coordonnées $(X_1, \ldots, X_r, \widetilde{x}_{r+1}, \ldots, \widetilde{x}_n)$. On considère ici $\widetilde{\ell}$ comme une forme linéaire sur K.

Si le rang r est égal à n, on a nécessairement $K = \{0\}$ et $\widetilde{\ell} = 0$, et on obtient alors après division éventuelle par la constante résiduelle (si elle est non nulle) une équation réduite

$$(a) \qquad \alpha_1' X_1^2 + \ldots + \alpha_n' X_n^2 = \pm 1 \qquad \text{ou} \qquad (b) \qquad \alpha_1 X_1^2 + \ldots + \alpha_n X_n^2 = 0.$$

Dans le cas (a), en fonction de la signature, il s'agit d'une quadrique de type ellipsoïde ou hyperboloïde (ou $\emptyset$), dans le cas (b) il s'agit d'un cône, éventuellement réduit à son sommet.

Supposons maintenant que $r \leq n-1$, de sorte que $K \neq \{0\}$. Si $\widetilde{l} \neq 0$, on peut trouver une nouvelle base orthonormée $(\widehat{e}_{r+1}, \ldots, \widehat{e}_n)$ de K de sorte que $(\widehat{e}_{r+2}, \ldots, \widehat{e}_n)$ constitue une base

de $\operatorname{Ker}\widetilde{\ell}$ et $\beta = \widetilde{l}(\widehat{e}_{r+1}) \neq 0$. Soient $(X_{r+1}, \ldots, X_n)$ les coordonnées de K relatives à la base $(\widehat{e}_{r+1}, \ldots, \widehat{e}_n)$ (les coordonnées $(X_1, \ldots, X_r)$ de S restant inchangées). On trouve alors

$$\widetilde{\ell}(\widetilde{x}) = \beta\, X_{r+1}.$$

Comme $\beta \neq 0$, on peut éliminer la constante résiduelle par une translation de la coordonnée X_{r+1}, ce qui, après division par $-\beta$, mène à l'équation réduite

$$(c) \qquad \alpha'_1 X_1^2 + \ldots + \alpha'_r X_r^2 - X_{r+1} = 0.$$

Il s'agit d'un paraboloïde si $r = n-1$, d'un cylindre à base paraboloïdale si $r \leq n-2$ (produit d'un paraboloïde de $\mathbb{R}^{r+1}$ par $\mathbb{R}^{n-r-1}$).

Il reste enfin le cas plus simple où $\widetilde{l} = 0$. Dans ce cas, après division éventuelle par la constante résiduelle $\widetilde{c}$ si $\widetilde{c} \neq 0$, on aboutit à l'équation réduite

$$(d) \qquad \alpha'_1 X_1^2 + \ldots + \alpha'_r X_r^2 = \pm 1 \qquad \text{ou} \qquad (e) \qquad \alpha_1 X_1^2 + \ldots + \alpha_r X_r^2 = 0.$$

Dans le cas (d), en fonction de la signature, il s'agit d'un cylindre à base ellipsoïde ou hyperboloïde (éventuellement dégénéré en un sous-espace linéaire ou en $\emptyset$), et dans le cas (e), il s'agit d'un cylindre à base conique (= produit d'un cône de $\mathbb{R}^r$ par $\mathbb{R}^{n-r}$).

8. Un bref aperçu de la vie de Fourier

Singulière destinée que celle de Jean Baptiste Joseph Fourier : né en 1768 à Auxerre dans une famille modeste – son père est garçon-tailleur – il est orphelin de mère à 8 ans et orphelin de père à 10 ans. Envoyé au pensionnat par l'organiste de la ville, il fait ses études à l'École militaire d'Auxerre alors tenue par les Bénédictins. Il étudie le latin, la rhétorique, la théologie, mais aussi les sciences et les mathématiques, qui deviennent rapidement son principal centre d'intérêt – Fourier découvre ainsi à la bibliothèque d'Auxerre des ouvrages écrits par quelques mathématiciens de premier plan comme Clairaut. Fourier se révèle vite être un élève hors norme, et il collectionne les premiers prix. Ses progrès sont si rapides que le directeur de l'école militaire d'Auxerre lui demande bientôt d'assurer la fonction de professeur de mathématiques, bien qu'il n'ait que 16 ans et demi !

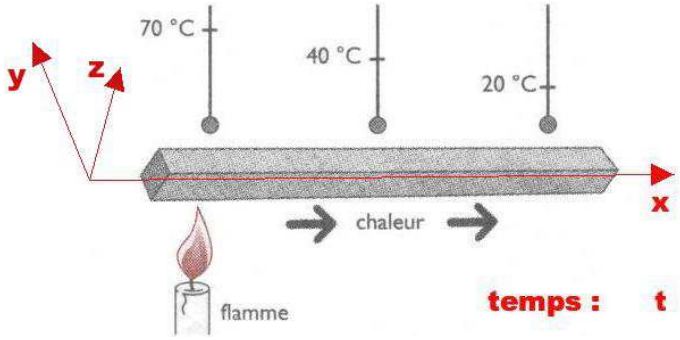
Un peu plus tard, Fourier prend une part active à la Révolution ; en 1792, il devient ainsi président de la société populaire d'Auxerre. Mais en 1794, c'est la Terreur, et Fourier est emprisonné. Il échappe de peu à l'échafaud, grâce à la la chute de Robespierre qui intervient juste avant la date prévue pour son jugement définitif. L'ouverture sociale consécutive à la

révolution permet au citoyen modeste qu'est Fourier d'entrer comme élève à l'École Normale de Paris nouvellement fondée en 1795. Ses travaux sur les équations algébriques le font très vite remarquer par le mathématicien Gaspard Monge. En 1796, celui-ci lui confie une charge de cours à l'École Normale, et le fait également nommer professeur à l'École Polytechnique : à moins de trente ans, Fourier est déjà ainsi un scientifique reconnu. En 1798, Bonaparte organise l'expédition d'Égypte, et Fourier y est associé comme l'un des principaux conseillers scientifiques. Il s'occupe de travaux de topographie, puis est nommé secrétaire de l'Institut d'Égypte au Caire, où il effectue de nombreuses missions de nature scientifique, administrative ou diplomatique. Après la débâcle française en Égypte, Fourier revient en France en 1801, et Napoléon le nomme peu après préfet de l'Isère. C'est à Grenoble, où il restera de 1802 à 1815, que Fourier entame ses travaux fondamentaux sur la propagation de la chaleur. Son activité est débordante – il crée l'université et le lycée de Grenoble, dirige les travaux de drainage de la vallée, fait construire la route de Grenoble à Briançon par le col du Lautaret. Parallèlement à ses travaux sur la propagation de la chaleur qui aboutissent à la publication d'un mémoire de l'Académie des Sciences en 1807, il rédige son importante "préface historique pour la description de l'Égypte", parue en 1809. Fourier fait la connaissance des frères Champollion, d'abord de l'aîné Jacques-Joseph puis de son frère cadet Jean-François ; les encouragements de Fourier et ses travaux précurseurs sur la civilisation égyptienne joueront ainsi un rôle décisif dans le déchiffrage des hiéroglyphes par Jean-François Champollion en 1822. Après la chute de Napoléon en 1815, Fourier connaît quelques difficultés (il avait été nommé baron d'Empire en 1809 !), mais c'est finalement la consécration avec son élection à l'Académie des Sciences en 1817. En 1822, il publie son monumental traité sur la "Théorie analytique de la chaleur" qui contient aussi en germe la théorie des séries et des transformées de Fourier, et il devient secrétaire perpétuel de l'Académie des sciences. Ses talents d'écrivain lui valent d'être élu simultanément à l'Académie Française en 1826, quelques années avant son décès en 1830.

L'Analyse de Fourier est devenue aujourd'hui un très vaste champ d'études. Les séries de Fourier réelles $\sum a_n \cos n\omega x + b_n \sin nx$ ou complexes $\sum c_n e^{inx}$ permettent de représenter tous les phénomènes périodiques et sont donc fondamentales en théorie des ondes et en théorie du signal. Mais il se trouve qu'il existe aussi un algorithme numérique rapide de calcul des séries de Fourier, connu sous le nom de FFT ou "Fast Fourier Transform". Celui-ci a des applications aussi bien en arithmétique que dans de nombreux domaines technologiques. L'algorithme de compression des images photographiques au format JPEG utilise quant à lui une "transformée de Fourier" discrète en cosinus, portant sur un échantillonnage par carrés de 8×8 pixels. C'est ainsi que Fourier est peut-être devenu le mathématicien le plus cité au monde, aucune branche de la science ne pouvant échapper aux séries et transformées de Fourier. Mais, bien que cela soit nettement moins connu, Fourier a aussi été un précurseur en statistiques ; encore plus fort, il a été le premier à imaginer et décrire l'effet de serre dû à la rétention de chaleur dans l'atmosphère des planètes gazeuses, dans un mémoire prémonitoire publié en 1824 dans les Annales de chimie et de physique. Fourier peut à bon droit être considéré comme le fondateur de la physique mathématique !

9. L'équation de la chaleur

Nous allons ici expliquer à la lumière des connaissances modernes le cheminement qui conduit à l'équation de la propagation de la chaleur. Considérons un objet matériel soumis à un échauffement initial, par exemple un barreau métallique, comme figuré ci-dessous :



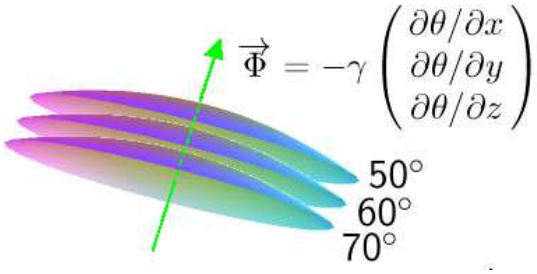
Le problème est de déterminer la température $\theta = \theta(x, y, z, t)$ de cet objet au point de coordonnées (x, y, z) et au temps t. Nous savons aujourd'hui que la température mesure l'énergie cinétique moyenne de vibration des atomes ou molécules constituant l'objet (mais la théorie atomique de Dalton est à peine née et Fourier n'y fait pas référence). Du fait que les particules s'entrechoquent, l'énergie cinétique se propage, d'autant plus vite que la différence de température entre des points voisins est plus grande. Notons $\overrightarrow{\Phi} = \overrightarrow{\Phi}(x, y, z, t)$ la *densité de flux de chaleur* traversant l'objet : c'est par définition la grandeur vectorielle dont la norme vaut $\Phi = dQ/(dS\,dt)$, si dQ est la quantité de chaleur traversant une surface infinitésimale dS pendant le temps dt, pour une surface dS perpendiculaire à la direction de propagation portée par $\overrightarrow{\Phi}$. Ces considérations conduisent à la loi physique

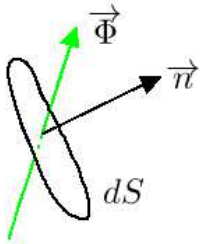
$$\overrightarrow{\Phi} = -\gamma \overrightarrow{\operatorname{grad}}\,\theta = -\gamma \begin{pmatrix} \partial\theta/\partial x \\ \partial\theta/\partial y \\ \partial\theta/\partial z \end{pmatrix} \tag{9.1}$$

où γ est la conductivité thermique du matériau ; la densité de flux Φ s'exprime en $Js^{-1}m^{-2} = Wm^{-2}$, et l'unité de conductivité thermique est donc le $WK^{-1}m^{-1}$ (où J = Joule, $W = Js^{-1}$ = Watt, K = Kelvin). On prendra garde au signe moins devant γ, qui traduit le fait que la chaleur se déplace des points chauds vers les points froids. Si on pose $\overrightarrow{dS} = dS\,\overrightarrow{n}$ où $\overrightarrow{n}$ est le vecteur normal à l'élément de surface, alors la quantité de chaleur dQ le traversant est en général donnée par

$$dQ = \overrightarrow{\Phi} \cdot \overrightarrow{dS}\,dt,$$

et ce, quelles que soient les orientations de $\overrightarrow{\Phi}$ et de $\overrightarrow{dS}$.





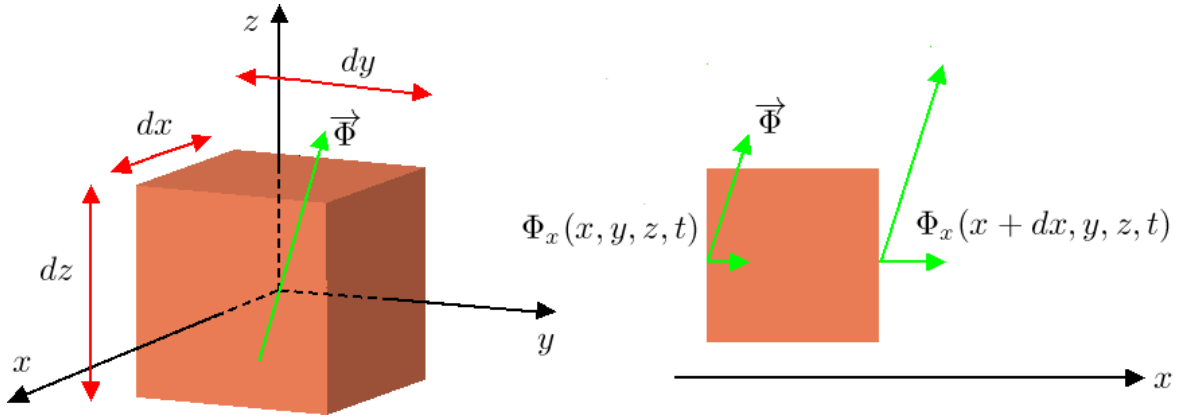
Considérons maintenant un petit parallélépipède $[x, x+dx] \times [y, y+dx] \times [z, z+dz]$ d'arêtes parallèles aux axes, de volume $dV = dx\,dy\,dz$. La quantité de chaleur qui entre dans ce parallélépipède pendant le temps dt est la somme algébrique des quantités de chaleur qui entrent ou sortent par chacune des 6 faces, soit par exemple

$$\begin{aligned} dQ_x &= \overrightarrow{\Phi}(x, y, z, t) \cdot (dy\,dz\ \overrightarrow{i})\,dt - \overrightarrow{\Phi}(x+dx, y, z, t) \cdot (dy\,dz\ \overrightarrow{i})\,dt \\ &= \big(\Phi_x(x, y, z, t) - \Phi_x(x+dx, y, z, t)\big) dy\,dz\,dt \end{aligned}$$

pour les faces d'abscisses x et $x+dx$, et en notant Φ_x la composante de $\overrightarrow{\Phi}$ suivant Ox. Si dx est très petit, on trouve

$$\Phi_x(x,y,z,t) - \Phi_x(x+dx,y,z,t) \sim -\frac{\partial \Phi_x}{\partial x}(x,y,z,t)\,dx,$$

du fait que l'on peut approximer le taux d'accroissement par la dérivée.



On obtient ainsi

$$dQ_x = -\frac{\partial \Phi_x}{\partial x}(x,y,z,t)\,dx\,dy\,dz\,dt.$$

En prenant la somme suivant les 3 paires de faces, on obtient que le bilan des "entrées-sorties" de chaleur est

$$dQ = dQ_x + dQ_y + dQ_z = -\left(\frac{\partial \Phi_x}{\partial x} + \frac{\partial \Phi_x}{\partial y} + \frac{\partial \Phi_z}{\partial z}\right) dx\,dy\,dz\,dt.$$

En combinant ce résultat avec l'expression (9.1) du flux, on obtient

$$dQ = \gamma\left(\frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2} + \frac{\partial^2 \theta}{\partial z^2}\right) dx\,dy\,dz\,dt. \tag{9.2}$$

Une autre loi physique fondamentale est la loi donnant la variation de température $d\theta$ produite par l'apport d'une quantité de chaleur dQ à un élément de matière de masse m. C'est une relation de proportionnalité, que l'on peut écrire

$$dQ = mc\,d\theta \tag{9.3}$$

où c est une constante appelée capacité calorifique, s'exprimant en $JK^{-1}kg^{-1}$ (Joules par degré et par kg). Ici la masse du petit parallélépipède vaut $m = \rho\,dx\,dy\,dz$ où ρ est la masse volumique (en $kg\,m^{-3}$). En comparant (9.2) et (9.3), on obtient l'égalité

$$dQ = (\rho\,dx\,dy\,dz)\,c\,d\theta = \gamma\left(\frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2} + \frac{\partial^2 \theta}{\partial z^2}\right) dx\,dy\,dz\,dt.$$

Après division par $\rho c\,dx\,dy\,dz\,dt$, on trouve l'équation fondamentale

$$\frac{\partial \theta}{\partial t} = \frac{\gamma}{\rho c}\left(\frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2} + \frac{\partial^2 \theta}{\partial z^2}\right). \tag{9.4}$$

Cette équation ne vaut que si le solide considéré ne reçoit pas de chaleur de l'extérieur – disons, par contact ou par rayonnement – et s'il ne produit pas lui-même de chaleur – ce qui est le cas s'il est le siège d'une réaction chimique ou nucléaire interne. Dans ce cas plus général, on note $P = P(x,y,z,t)$ la *production (ou l'apport) volumique* de chaleur par unité de volume et de temps à la position (x,y,z) et au temps t, en $W\,m^{-3}$; il va alors s'ajouter

à dQ une quantité de chaleur supplémentaire $dQ' = P(x,y,z,t)\,dx\,dy\,dz\,dt$, ce qui donne l'équation générale

$$\frac{\partial\theta}{\partial t} = \frac{\gamma}{\rho c}\left(\frac{\partial^2\theta}{\partial x^2} + \frac{\partial^2\theta}{\partial y^2} + \frac{\partial^2\theta}{\partial z^2}\right) + \frac{P}{\rho c}, \tag{9.5}$$

appelée *équation de propagation de la chaleur*. Aux notations près, c'est exactement l'équation proposée par Fourier en 1807 ! Il est commode d'introduire la constante $D = \gamma/\rho c$ spécifique du matériau, qu'on appelle le coefficient de *diffusivité thermique* (en $m^2 s^{-1}$). Avec cette notation, l'équation de la chaleur prend la forme usuelle

$$\frac{\partial\theta}{\partial t} = D\left(\frac{\partial^2\theta}{\partial x^2} + \frac{\partial^2\theta}{\partial y^2} + \frac{\partial^2\theta}{\partial z^2}\right) + \frac{P}{\rho c}. \tag{9.6}$$

Lien avec les séries trigonométriques. L'équation de la chaleur ne peut en général être résolue explicitement. Considérons le cas d'un barreau métallique de longueur L, qui a été chauffé initialement de manière non uniforme, et dont on étudie l'évolution de la température. On l'assimile à un segment $[0,L]$ dans la direction $0x$, en négligeant son épaisseur et les variations de température en y et z, de sorte que la température est juste une fonction $\theta(x,t)$ de l'abscisse et du temps. En l'absence de production interne de chaleur, l'équation (9.6) prend la forme très simplifiée

$$\frac{\partial\theta}{\partial t} = D\frac{\partial^2\theta}{\partial x^2}. \tag{9.7}$$

À cette équation, il faut ajouter le fait que le flux de chaleur $\Phi(x,t) = -\gamma\frac{\partial\theta}{\partial x}(x,t)$ est nul aux extrêmités du barreau, donc lorsque $x = 0$ ou $x = L$, ce qui donne les conditions supplémentaires (dites *conditions aux limites*)

$$\frac{\partial\theta}{\partial x}(0,t) = 0, \qquad \frac{\partial\theta}{\partial x}(L,t) = 0. \tag{9.8}$$

Ces équations sont déjà bien assez difficiles à résoudre ! La méthode de Fourier consiste à rechercher des solutions sous forme de fonctions à variables séparées $\theta(x,t) = f(x)\,g(t)$. L'équation (9.7) devient dans ce cas $f(x)g'(t) = Df''(x)g(t)$, soit encore

$$g'(t)/g(t) = D\,f''(x)/f(x)$$

si les fonctions ne s'annulent pas. On voit que ces quotients doivent être constants, disons $f''(x)/f(x) = a$ et $g'(t)/g(t) = aD$. La deuxième relation donne aussitôt $g(t) = C\,e^{aDt}$, ce qui est physiquement inacceptable si $a > 0$ (la température augmenterait de manière exponentielle avec le temps). La constante a est donc négative ou nulle, et on peut poser $a = -\omega^2$. Ceci fournit alors les équations différentielles bien connues

$$f''(x) = -\omega^2 f(x), \qquad g'(t) = -(\omega^2 D)g(t),$$

d'où les solutions

$$f(x) = \alpha\cos\omega x + \beta\sin\omega x, \qquad g(t) = e^{-\omega^2 Dt} \quad \text{à constante près,}$$
$$\theta(x,t) = f(x)\,g(t) = (\alpha\cos\omega x + \beta\sin\omega x)e^{-\omega^2 Dt}.$$

La première condition (9.8) impose $\frac{\partial\theta}{\partial x}(0,t) = \beta\omega\,e^{-\omega^2 Dt} = 0$, donc $\beta = 0$, et la seconde donne alors $\frac{\partial\theta}{\partial x}(L,t) = -\alpha\omega\,\sin\omega L\,e^{-\omega^2 Dt} = 0$, donc $\sin\omega L = 0$ (on remarquera que si $\omega = 0$, la solution est $f(x) = \alpha + \beta x$ qui ne permet de réaliser (9.8) que si $\beta = 0$, auquel cas

$\theta(x,t) = \alpha = \text{Cte}$, et sinon on peut supposer $\omega > 0$). Cette dernière condition montre que l'on doit avoir $\omega L = n\pi$ et donc $\omega = n\pi/L$, $n \in \mathbb{N}$. Ceci fournit la solution

$$\theta(x,t) = \alpha \cos\left(\frac{n\pi}{L}x\right)\, e^{-(n^2\pi^2/L^2)Dt},$$

encore valable pour $n = 0$. Comme l'équation (9.7) est linéaire, on peut ajouter ces solutions entre elles, ce qui donne également comme solutions toutes les sommes

$$\theta(x,t) = \sum_{0\le n\le N} \alpha_n \cos\left(\frac{n\pi}{L}x\right)\, e^{-(n^2\pi^2/L^2)Dt},$$

et par passage à la limite quand $N \to +\infty$ toutes les *séries trigonométriques convergentes*

$$\theta(x,t) = \sum_{n=0}^{+\infty} \alpha_n \cos\left(\frac{n\pi}{L}x\right)\, e^{-(n^2\pi^2/L^2)Dt}. \tag{9.9}$$

Il faut supposer ici que les coefficients (α_n) soient choisis en sorte que l'on puisse dériver terme à terme la série deux fois par rapport à x et une fois par rapport à t, ce qui est le cas par exemple si $\sum_{n=0}^{+\infty} n^2|\alpha_n| < +\infty$. À ce point, la grande audace de Fourier a été de prétendre que l'on obtenait ainsi toutes les solutions ! Si l'on étudie l'évolution de la température à partir du temps $t = 0$, il faudrait déjà que

$$\theta(x,0) = \sum_{n=0}^{+\infty} \alpha_n \cos\left(\frac{n\pi}{L}x\right) \tag{9.10}$$

puisse représenter au temps initial n'importe quelle fonction température suffisamment régulière sur l'intervalle $[0, L]$, disons par exemple n'importe quelle fonction dérivable à dérivée continue, de dérivée nulle en $x = 0$ et en $x = L$. Face aux objections de ses contemporains, Fourier tente de se justifier, et il y parvient en partie, mais une preuve vraiment indiscutable de cette propriété ne viendra que plus de 20 ans plus tard, avec le mathématicien allemand Gustav Lejeune-Dirichlet (1805–1859), en 1829. Il nous reste donc encore bien du travail à accomplir pour détailler et justifier mathématiquement tous ces résultats et toutes ces intuitions ! Ce sera l'objet des sections suivantes.

10. Séries de Fourier, introduction

Comme nous l'avons déjà entrevu, l'une des idées essentielles de Fourier est l'approximation des fonctions périodiques par des séries trigonométriques. On introduit à cet effet la définition suivante.

Définition 10.1. *On appelle polynôme trigonométrique de degré $\le N$ et de pulsation $\omega > 0$ toute fonction de la forme*

$$P(x) = \sum_{n=-N}^{+N} c_n e^{in\omega x}, \qquad c_n \in \mathbb{C}.$$

On notera $\mathcal{P}_{N,\omega}$ le $\mathbb{C}$-espace vectoriel des polynômes trigonométriques de degré $\le N$ et de pulsation ω.

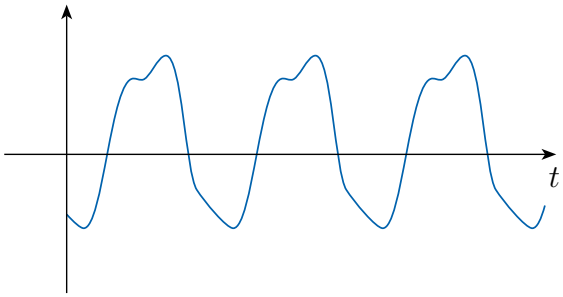
On voit immédiatement qu'il s'agit de **fonctions périodiques** de période

$$T = \frac{2\pi}{\omega}.$$

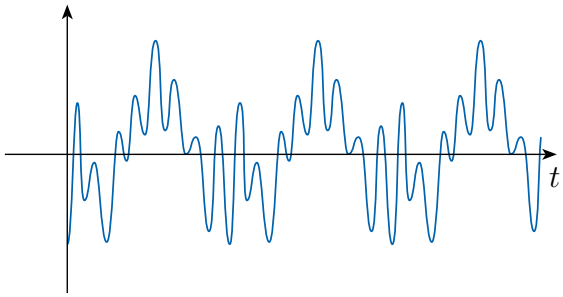
On a en effet $e^{2\pi i} = 1$, d'où

$$e^{in\omega(x+T)} = e^{in\omega x + in\cdot 2\pi} = e^{in\omega x}(e^{2\pi i})^n = e^{in\omega x},$$

et donc $P(x+T) = P(x)$; plus généralement, on a $P(x+kT) = P(x)$ pour tout $k \in \mathbb{Z}$.



Son de la flûte



Son du violon

En théorie du signal, on s'intéresse aux fonctions périodiques de période T donnée, et la question fondamentale est de savoir si on peut approcher tout signal périodique f par une suite $f_N \in \mathcal{P}_{N,\omega}$ de polynômes trigonométriques, avec des composantes dont les fréquences $n\omega$ sont multiples de la fréquence fondamentale $\omega = 2\pi/T$. Par référence aux signaux sonores, les composantes $c_n e^{in\omega x}$ s'appellent les harmoniques de la fréquence fondamentale.

La formule d'Euler

$$e^{in\omega x} = \cos(n\omega x) + i\sin(n\omega x)$$

permet d'écrire alternativement tout polynôme $P(x) = \sum_{|n|\leq N} c_n e^{in\omega x}$ sous la forme

$$P(x) = c_0 + \sum_{n=1}^{+N} a_n \cos(n\omega x) + b_n \sin(n\omega x)$$

avec $a_n, b_n \in \mathbb{C}$ reliés aux coefficients c_n par les relations

$$\begin{cases} a_n = c_n + c_{-n} \\ b_n = ic_n - ic_{-n} \end{cases} \qquad \text{pour tout } n \geq 1.$$

Lorsque l'on travaille avec des fonctions réelles, il peut en effet s'avérer plus commode de travailler avec les fonctions cos et sin (mais pas toujours !). Réciproquement, partant d'une écriture en cos et sin, les formules

$$\cos(n\omega x) = \frac{1}{2}\big(e^{in\omega x} + e^{-in\omega x}\big), \qquad \sin(n\omega x) = \frac{1}{2i}\big(e^{in\omega x} - e^{-in\omega x}\big)$$

permettent de calculer c_n en fonctions des coefficients a_n et b_n :

$$c_n = \frac{1}{2}a_n + \frac{1}{2i}b_n, \qquad c_{-n} = \frac{1}{2}a_n - \frac{1}{2i}b_n.$$

Introduisons quelques définitions et notations utiles pour la suite.

Notation 10.2. *Soit $p \geq 0$ un entier et $I \subset \mathbb{R}$ un intervalle. Pour $\mathbb{K} = \mathbb{R}$ ou $\mathbb{K} = \mathbb{C}$, on désignera par*

(1) *$C^p(I, \mathbb{K})$ l'ensemble des fonctions $f : I \to \mathbb{K}$ de classe C^p, c'est-à-dire l'ensemble des fonctions f admettant des dérivées f, f', …, $f^{(p)}$ continues sur I.*

(2) *$C^p(\mathbb{R}/T\mathbb{Z}, \mathbb{K})$ l'ensemble des fonctions $f : \mathbb{R} \to \mathbb{K}$ de classe C^p qui sont en outre périodiques de période T : $f(x+kT) = f(x)$ pour tout $k \in \mathbb{Z}$.*

Il arrive cependant assez fréquemment que certains signaux périodiques f utilisés en électronique ou en physique ne soient pas continus ("signal carré" par exemple). Il est donc commode d'introduire aussi les fonctions de classe C^p par morceaux.

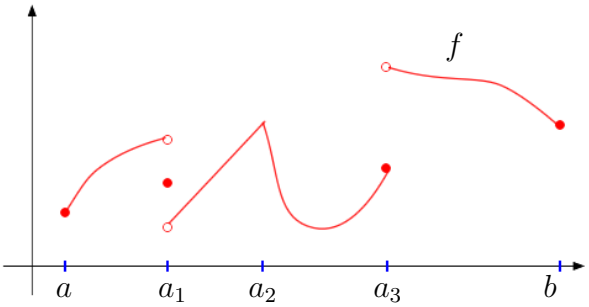
Définition 10.3. *Soit $p \geq 0$ un entier.*

(1) *Une fonction $f : [a,b] \to \mathbb{K}$ est dite de classe C^p* **par morceaux** *s'il existe une subdivision finie*

$$a = a_0 < a_1 < \ldots < a_s = b$$

de $[a,b]$ telle que f soit de classe C^p sur chaque intervalle $]a_{k-1}, a_k[$, et qu'en outre les limites à droite et à gauche $f^{(j)}(a_k \pm 0) = \lim_{x \to a_k \pm 0} f^{(j)}(x)$ existent pour tout k et tout $j = 0, 1, \ldots, p$ (avec la convention $f^{(0)} = f$; on prend seulement la limite à droite en a et la limite à gauche en b). On notera $C^p_{\text{morc}}([a,b], \mathbb{K})$ l'ensemble des fonctions C^p par morceaux.

(2) *On notera de même $C^p_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{K})$ l'ensemble des fonctions périodiques de période T qui sont de classe C^p par morceaux sur tout intervalle $[x_0, x_0 + T]$ (compte tenu de la périodicité, le choix de x_0 n'a pas d'importance).*



Fonction de classe C^p par morceaux

Sur le schéma précédent, on voit que $x = a_2$ est un point de continuité pour f, mais comme c'est un point anguleux du graphe, il s'agit d'un point de discontinuité pour f', seules les dérivées à droite et à gauche sont définies.

Définition 10.4. *Si f est une fonction continue par morceaux, on définit la* **valeur principale** *de f en tout point par*

$$\mathrm{VP}(f)(x) = \frac{1}{2}\big(f(x+0) + f(x-0)\big).$$

On a donc en particulier $\mathrm{VP}(f)(x) = f(x)$ en tout point où f est continue.

Sur le schéma ci-dessus, on a par exemple $f(a_1) = \mathrm{VP}(f)(a_1)$, mais ce n'est pas le cas pour le point a_3 (dans cette définition, on ne se préoccupe pas des dérivées de f ...).

Notation 10.5. *Dans l'espace $C^p_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{K})$, on considère le sous-espace*

$$C^{p\,[\mathrm{VP}]}_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{K})$$

des fonctions périodiques de classe C^p par morceaux qui coïncident avec leur valeur principale en tout point ; autrement dit, on suppose que

$$f(x) = \mathrm{VP}(f)(x) = \frac{1}{2}\big(f(x+0) + f(x-0)\big)$$

en tout point x de discontinuité. Il est immédiat que $C^{p\,[\mathrm{VP}]}_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{K})$ est un sous-espace vectoriel de $C^p_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{K})$.

Pour $f \in C^0_{\mathrm{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$ et $q > 0$, on définit la semi-norme L^q de f comme étant le nombre réel ≥ 0

$$\|f\|_q = \left(\frac{1}{T} \int_{x_0}^{x_0+T} |f(x)|^q \, dx \right)^{1/q}.$$

L'intégrale est bien définie grâce à l'hypothèse que f est continue par morceaux. On notera que l'intégrale ne dépend pas du choix du point initial x_0 du fait de la périodicité de f ; on choisira souvent $x_0 = 0$ ou $x_0 = -T/2$, en fonction de la commodité des calculs à effectuer. (Les espaces L^q ont été introduits à la suite des travaux du mathématicien français Henri-Léon Lebesgue (1875–1941), fondateur d'une théorie de l'intégration très améliorée entre 1902 et 1904 ; c'est la raison pour laquelle on utilise la lettre L ...). Dans toute la suite, on considérera surtout la norme L^2 ($q = 2$), car dans ce cas il s'agit d'une (semi)-norme hermitienne associée à la forme sesquilinéaire hermitienne semi-positive

$$\langle f, g \rangle = \frac{1}{T} \int_{x_0}^{x_0+T} \overline{f(x)} g(x) \, dx, \qquad f, g \in C^0_{\mathrm{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C}).$$

Cette forme sesquilinéaire n'est pas définie positive sur $C^0_{\mathrm{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$ car les fonctions f nulles partout en dehors d'un ensemble fini $\{a_k\}$ modulo T sont d'intégrale nulle (et sont bien continues par morceaux) ; elles constituent donc des vecteurs isotropes. Comme une fonction continue ≥ 0 d'intégrale nulle est nécessairement nulle, on voit que ces fonctions f sont les seuls vecteurs isotropes présents dans $C^0_{\mathrm{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$. Si l'on impose que f coïncide avec sa valeur principale, cet ennui disparaît, car les valeurs $f(a_k)$ sont alors nécessairement nulles elles aussi :

Corollaire 10.6. *Le produit scalaire L^2 est défini positif sur $C^{0\,[\mathrm{VP}]}_{\mathrm{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$.*

L'une des observations essentielles de la théorie des séries de Fourier est la suivante.

Théorème 10.7. *Pour $a \in \mathbb{R}$, considérons les fonctions e_a, $\cos_a$, $\sin_a$ telles que*

$$e_a(x) = e^{iax}, \qquad \cos_a(x) = \cos(ax), \qquad \sin_a(x) = \sin(ax).$$

Alors

(1) *la famille $(e_{n\omega})_{n\in\mathbb{Z}}$ est une famille de vecteurs orthonormée pour le produit scalaire hermitien L^2, c'est-à-dire que $\langle e_{n\omega}, e_{m\omega} \rangle = \delta_{nm}$ (indice de Kronecker).*

(2) *la famille $(e_0, \cos_{n\omega}, \sin_{n\omega})_{n\geq 1}$ est une famille orthogonale pour le produit scalaire L^2 (sur $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$).*

Les normes L^2 de ces fonctions sont données par

$$\|e_{n\omega}\|_2 = 1, \qquad \|\cos_{n\omega}\|_2^2 = \|\sin_{n\omega}\|_2^2 = \frac{1}{2}.$$

Démonstration. Comme $\overline{e_{n\omega}(x)} = \overline{e^{in\omega x}} = e^{-in\omega x}$, on obtient par définition

$$\langle e_{n\omega}, e_{m\omega} \rangle = \frac{1}{T} \int_0^T \overline{e^{in\omega x}} \, e^{im\omega x} \, dx = \frac{1}{T} \int_0^T e^{i(m-n)\omega x} \, dx.$$

Si $n = m$, on a $e^{i(m-n)\omega x} = 1$ et donc $\langle e_{n\omega}, e_{m\omega} \rangle = 1$. Si $n \neq m$, une primitive de $e^{i(m-n)\omega x}$ est $\frac{1}{i(m-n)} e^{i(m-n)\omega x}$, et compte tenu de la périodicité on voit que l'intégrale est nulle, i.e. $\langle e_{n\omega}, e_{m\omega} \rangle = 0$. Ceci démontre déjà le point (1). On vérifie aisément à partir de là que les produits scalaires mutuels des fonctions

$$\cos_{n\omega} = \frac{1}{2}\big(e_{n\omega} + e_{-n\omega}\big), \qquad \sin_{m\omega} = \frac{1}{2i}\big(e_{m\omega} - e_{-m\omega}\big), \qquad n, m \geq 1,$$

sont nuls, sauf les normes L^2 $\| \cos_{m\omega} \|_2^2 = |\frac{1}{2}|^2 + |\frac{1}{2}|^2 = \frac{1}{2}$ et $\| \sin_{m\omega} \|_2^2 = |\frac{1}{2i}|^2 + |\frac{1}{2i}|^2 = \frac{1}{2}$. En d'autres termes, on a des "valeurs moyennes quadratiques" égales à $\frac{1}{2}$:

$$\frac{1}{T}\int_0^T (\cos n\omega x)^2\, dx = \frac{1}{T}\int_0^T (\sin n\omega x)^2\, dx = \frac{1}{2}.$$

Il est facile de voir aussi que $e_0 \perp \cos_{n\omega}$ et $e_0 \perp \sin_{n\omega}$. Le Théorème en résulte. □

En utilisant la Proposition 5.5, on en déduit déjà le corollaire suivant.

Corollaire 10.8. *Les familles* $(e_{n\omega})_{n\in\mathbb{Z}}$ *et* $(e_0, \cos_{n\omega}, \sin_{n\omega})_{n\geq 1}$ *sont* **libres**. *En particulier, l'espace* $\mathcal{P}_{N,\omega}$ *des polynômes trigonométriques de degré* $\leq N$ *est de dimension* $2N+1$ *sur* $\mathbb{C}$; *il admet*

$(e_{n\omega})_{-N\leq n\leq +N}$ *comme base orthonormée*

et

$(e_0, \cos_{n\omega}, \sin_{n\omega})_{1\leq n\leq N}$ *comme base orthogonale.*

Étant donné une fonction $f \in C^0_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$, on cherche maintenant un polynôme trigonométrique $f_N \in \mathcal{P}_{N,\omega}$ qui approxime f le mieux possible en norme L^2. D'après la Proposition 5.6, la fonction f_N qui répond à la question est la **projection orthogonale**

$$f_N = \pi_N(f) \quad \text{de } f \text{ sur } \mathcal{P}_{N,\omega}.$$

D'après les résultats vus à la Section 4, nous avons la formule

$$f_N = \sum_{n=-N}^{+N} \langle e_{n\omega}, f\rangle\, e_{n\omega},$$

soit encore

$$f_N(x) = \sum_{n=-N}^{+N} c_n(f)\, e^{in\omega x}$$

avec par définition

$$c_n(f) = \langle e_{n\omega}, f\rangle = \frac{1}{T}\int_{x_0}^{x_0+T} f(x)\, e^{-in\omega x}\, dx.$$

Notation 10.9. Pour $f \in C^0_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$, on appelle **coefficients de Fourier** de f les nombres complexes $\widehat{f}(n) = \langle e_{n\omega}, f\rangle$, encore notés $c_n(f)$, c'est-à-dire

$$\widehat{f}(n) = c_n(f) = \frac{1}{T}\int_{x_0}^{x_0+T} f(x)\, e^{-in\omega x}\, dx.$$

Si l'on préfère travailler avec des fonctions réelles, on utilisera plutôt la base orthogonale $(e_0, \cos_{n\omega}, \sin_{n\omega})_{1\leq n\leq N}$. Compte tenu de la norme $= \frac{1}{\sqrt{2}}$ de $\cos_{n\omega}$ et $\sin_{n\omega}$, ceci conduit aux formules alternatives

$$f_N = \langle e_0, f\rangle\, e_0 + \sum_{n=1}^{N} 2\langle \cos_{n\omega}, f\rangle\, \cos_{n\omega} + 2\langle \sin_{n\omega}, f\rangle\, \sin_{n\omega},$$

ou encore

$$f_N(x) = c_0(f) + \sum_{n=1}^{N} a_n(f)\, \cos(n\omega x) + b_n(f)\, \sin(n\omega x)$$

avec

$$c_0(f) = \langle e_0, f\rangle = \frac{1}{T}\int_{x_0}^{x_0+T} f(x)\,dx = \text{valeur moyenne de } f,$$
$$a_n(f) = 2\langle \cos_{n\omega}, f\rangle = \frac{2}{T}\int_{x_0}^{x_0+T} f(x)\,\cos(n\omega x)\,dx,$$
$$b_n(f) = 2\langle \cos_{n\omega}, f\rangle = \frac{2}{T}\int_{x_0}^{x_0+T} f(x)\,\sin(n\omega x)\,dx.$$

Lorsque f est paire ou impaire, on peut utiliser le choix symétrique $x_0 = -T/2$ qui conduit à considérer des intégrales $\int_{-T/2}^{T/2}(...)$, et on en déduit l'observation pratique importante :

Proposition 10.10.

(1) *Si f est paire, alors $b_n(f) = 0$ et la projection $f_N = \pi_N(f)$ est donnée par*

$$f_N(x) = c_0(f) + \sum_{n=1}^{N} a_n(f)\,\cos(n\omega x), \qquad a_n(f) = \frac{4}{T}\int_0^{T/2} f(x)\,\cos(n\omega x)\,dx, \quad n \geq 1.$$

(2) *Si f est impaire, alors $a_n(f) = 0$ et la projection $f_N = \pi_N(f)$ est donnée par*

$$f_N(x) = \sum_{n=1}^{N} b_n(f)\,\sin(n\omega x), \qquad b_n(f) = \frac{4}{T}\int_0^{T/2} f(x)\,\sin(n\omega x)\,dx, \quad n \geq 1.$$

En général, comme $f = f_N + (f - f_N)$ avec $f_N \in \mathcal{P}_{N,\omega}$ et $(f - f_N) \in \mathcal{P}_{N,\omega}^{\perp}$, le théorème de Pythagore donne la formule

$$\|f\|_2^2 = \|f_N\|_2^2 + \|f - f_N\|_2^2,$$

en particulier $\|f_N\|_2^2 \leq \|f\|_2^2$. Cette inégalité, attribuée au mathématicien Friedrich Wilhelm Bessel (1784–1846) peut encore s'énoncer sous la forme suivante.

Théorème 10.11 (Inégalité de Bessel)**.** *Pour tout $f \in C^0_{\text{morc}}(\mathbb{R}/\mathbb{Z}, \mathbb{C})$ et tout entier $N \in \mathbb{N}$, on a l'inégalité*

$$\sum_{n=-N}^{+N} |c_n(f)|^2 \leq \|f\|_2^2 = \frac{1}{T}\int_{x_0}^{x_0+T} |f(x)|^2 dx.$$

La question essentielle qui subsiste est de savoir si l'on a bien $\lim_{N\to+\infty} \|f - f_N\|_2 = 0$ (et donc $\lim_{N\to+\infty} \|f_N\|_2 = \|f\|_2$). Ce problème est lié à l'étude de la convergence de la **série de Fourier** de f, à savoir la série trigonométrique

$$\sum_{n=-\infty}^{+\infty} c_n(f)\, e^{in\omega x}.$$

Avant d'entamer l'étude de la convergence d'une telle série, nous aurons besoin de quelques résultats préliminaires généraux concernant les séries numériques (sommes infinies de nombres réels ou complexes).

11. Notions de base sur les séries numériques et les séries de fonctions

Une série numérique est une somme infinie de la forme

$$(*) \qquad \sum_{n=n_0}^{+\infty} u_n, \qquad u_n \in \mathbb{R} \text{ ou } \mathbb{C}.$$

Dans la théorie des séries de Fourier on a plutôt affaire à des sommes infinies sur $\mathbb{Z}$ tout entier, du type $\sum_{n=-\infty}^{+\infty} u_n = \sum_{n=-\infty}^{+\infty} c_n e^{in\omega x}$; on les regroupera en général sous la forme

$$c_0 + \sum_{n=1}^{+\infty} (c_n e^{in\omega x} + c_{-n} e^{-in\omega x})$$

(on pourrait aussi calculer séparément les deux sommes $\sum_{n=0}^{+\infty} u_n$, $\sum_{n=1}^{+\infty} u_{-n}$ et les ajouter, mais les conditions de convergence ne sont pas exactement identiques en général car les termes u_n et u_{-n} peuvent se compenser entre eux). Pour donner sens à une somme infinie telle que $(*)$, on calcule les **sommes partielles**

$$S_N = \sum_{n=n_0}^{N} u_n = u_{n_0} + u_{n_0+1} + \ldots + u_n + \ldots + u_{N-1} + u_N$$

et on pose la définition suivante.

Définition 11.1. *On dit que la série* $\sum_{n=n_0}^{+\infty} u_n$ *est* **convergente** *si la limite* $S = \lim_{N\to+\infty} S_N$ *existe dans* $\mathbb{C}$. *Dans ce cas on écrit*

$$\sum_{n=n_0}^{+\infty} u_n = S$$

et on dit que S **est la somme de la série**. *Dans le cas contraire, on dit que la série est* **divergente**.

Il est clair que toute combinaison linéaire $\sum_{n=n_0}^{+\infty} (\lambda u_n + \mu v_n)$ de séries convergentes $\sum u_n$, $\sum v_n$ est convergente. Une autre observation immédiate importante est la suivante :

Proposition 11.2. *Pour qu'une série* $\displaystyle\sum_{n=n_0}^{+\infty} u_n$ *converge, il est nécessaire que* $\lim\limits_{n\to+\infty} u_n = 0$.

Démonstration. En effet, on a $u_n = S_n - S_{n-1}$, donc l'existence d'une limite $S = \lim_{N\to+\infty} S_N$ implique $\lim_{n\to+\infty} u_n = S - S = 0$. □

Si la série converge, on appelle **reste d'ordre** N de la série la différence $R_N = S - S_N$. Il est évident que l'on a alors

$$R_N = \sum_{n=N+1}^{+\infty} u_n \quad \text{et} \quad \lim_{N\to+\infty} R_N = 0.$$

Exemple 11.3. Considérons la "série géométrique de raison a"

$$\sum_{n=n_0}^{+\infty} a^n, \qquad a \in \mathbb{C}.$$

D'après une identité remarquable bien connue on obtient

$$S_N = a^{n_0}(1 + a + a^2 + \ldots + a^{N-n_0}) = a^{n_0} \frac{1 - a^{N-n_0+1}}{1-a} = \frac{a^{n_0} - a^{N+1}}{1-a} \qquad \text{si } a \neq 1.$$

• Si $a = 1$, on trouve $S_N = N - n_0 + 1$, donc $\lim_{N\to+\infty} S_N = +\infty$, la série est divergente.

Supposons maintenant $a \neq 1$.

• Si $|a| < 1$, nous avons $\lim_{N\to+\infty} a^{N+1} = 0$, donc la série géométrique est convergente et

$$\sum_{n=n_0}^{+\infty} a^n = \lim_{N\to+\infty} S_N = \frac{a^{n_0}}{1-a}.$$

• Si $|a| > 1$, nous avons $\lim_{N\to+\infty} |a^{N+1}| = +\infty$, donc $\lim_{N\to+\infty} |S_N| = +\infty$ et la série géométrique est divergente.

• Si $|a| = 1$, c'est-à-dire $a = e^{i\theta}$, le nombre complexe $a^{N+1} = e^{i(N+1)\theta}$ décrit le cercle trigonométrique sans avoir de limite si $a \neq 1$, la suite (S_N) n'a donc pas de limite et la série est divergente. On remarquera cependant que l'on a dans ce cas

$$|S_N| \leq \frac{2}{|1-a|},$$

les sommes partielles sont bornées. Par exemple, dans le cas $a = -1$ et n_0 pair, on voit aussitôt que les sommes partielles S_n successives sont $1, 0, 1, 0, 1, 0, \ldots$.

En résumé, la série géométrique de raison a converge si et seulement si $|a| < 1$, et dans ce cas le reste de la série est donné par

$$R_N = \sum_{n=N+1}^{+\infty} a^n = \frac{a^{N+1}}{1-a}.$$

Série à termes positifs. Les séries les plus simples à étudier sont les séries à termes réels $u_n \geq 0$. En effet les sommes partielles S_N forment alors une **suite croissante** ≥ 0, puisque $S_N = S_{N-1} + u_N \geq S_{N-1}$. On a donc deux éventualités :

• la suite (S_N) est majorée. Dans ce cas elle admet une limite S égale à sa borne supérieure, et la série $\sum_{n=n_0}^{+\infty} u_n$ converge vers $S \in \mathbb{R}_+$.

• la suite (S_N) n'est pas majorée. Dans ce cas $\lim_{N\to+\infty} S_N = +\infty$ et la série est divergente. On convient d'écrire encore

$$\sum_{n=n_0}^{+\infty} u_n = +\infty$$

(cette notation étant réservée exclusivement au cas des séries à termes positifs).

Le critère de comparaison suivant est très utile, même s'il est quasiment évident.

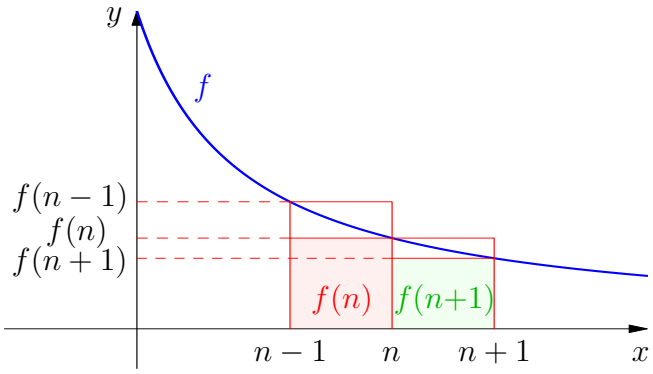
Proposition 11.4 (Critère de comparaison)**.** *Soient $\sum_{n=n_0}^{+\infty} u_n$ et $\sum_{n=n_0}^{+\infty} v_n$ deux séries telles que $u_n \geq v_n \geq 0$ pour $n \geq n_1$ assez grand.*

- *Si $\sum_{n=n_0}^{+\infty} u_n$ converge, alors $\sum_{n=n_0}^{+\infty} v_n$ converge.*
- *Si $\sum_{n=n_0}^{+\infty} v_n$ diverge, alors $\sum_{n=n_0}^{+\infty} u_n$ diverge (vers $+\infty$).*

Démonstration. Comme la nature (convergence/divergence) d'une série ne change pas si on change l'indice de sommation initial n_0, on peut toujours supposer que $n_1 = n_0$ (quitte à remplacer n_0 et n_1 par $\max(n_0, n_1)$). Soit S_N (resp. S'_N) les sommes partielles de $\sum u_n$ (resp. de $\sum v_n$). L'hypothèse $u_n \geq v_n \geq 0$ entraîne $S_N \geq S'_N \geq 0$. Par conséquent, si $S = \sum_{n=n_0}^{+\infty} u_n$ converge, la suite (S'_N) admet la majoration $S'_N \leq S_n \leq S$, donc $\sum_{n=n_0}^{+\infty} v_n$ converge. Si $\sum_{n=n_0}^{+\infty} v_n$ diverge, alors $\lim_{N\to+\infty} S'_N = +\infty$, donc $\sum_{n=n_0}^{+\infty} u_n = \lim_{N\to+\infty} S_N = +\infty$ diverge. □

Remarque 11.5. Si $u_n > 0$ et $v_n > 0$ sont telles que $u_n \sim v_n$ quand $n \to +\infty$, on en déduit que les séries $\sum u_n$ et $\sum v_n$ sont de même nature, puisque pour tout $\varepsilon > 0$ on a $0 \leq u_n \leq (1+\varepsilon)v_n$ et $0 \leq v_n \leq (1+\varepsilon)u_n$ pour $n \geq n_\varepsilon$ assez grand.

Méthode de comparaison des séries et des intégrales. Soit $f : [n_0, +\infty[\to \mathbb{R}$ une fonction positive décroissante. Nous allons comparer la série $\sum_{n=n_0}^{+\infty} f(n)$ et l'intégrale $\int_{n_0}^{+\infty} f(x)\,dx$.



Comme f est décroissante, on a les inégalités

$$\int_n^{n+1} f(x)\,dx \le f(n) \le \int_{n-1}^{n} f(x)\,dx.$$

En effectuant la somme de $n = n_0$ à N, il vient

$$\int_{n_0}^{N+1} f(x)\,dx \le \sum_{n=n_0}^{N} f(n) \le f(n_0) + \int_{n_0}^{N} f(x)\,dx.$$

Si l'on pose par définition

$$\int_{n_0}^{+\infty} f(x)\,dx = \lim_{A\to+\infty} \int_{n_0}^{A} f(x)\,dx,$$

on obtient en définitive l'encadrement :

Théorème 11.6. *Pour toute fonction $f : [n_0, +\infty[\to \mathbb{R}$ positive décroissante on a*

$$\int_{n_0}^{+\infty} f(x)\,dx \le \sum_{n=n_0}^{+\infty} f(n) \le f(n_0) + \int_{n_0}^{+\infty} f(x)\,dx,$$

en particulier la série $\sum_{n=n_0}^{+\infty} f(n)$ et l'intégrale $\int_{n_0}^{+\infty} f(x)\,dx$ sont simultanément convergentes ou divergentes.

Nous allons utiliser ce critère pour déterminer la nature de la série de Riemann

$$\zeta(\alpha) = \sum_{n=1}^{+\infty} \frac{1}{n^\alpha}, \qquad \alpha \in \mathbb{R},$$

étdiée en profondeur par le mathématicien Bernhard Riemann (1826–1866). À l'aide d'une définition alternative, Riemann montra en fait la possibilité d'étendre $\zeta(\alpha)$ à tout $\alpha \in \mathbb{C}\setminus\{1\}$. "L'hypothèse de Riemann" est la question de savoir si les zéros $\zeta(\alpha) = 0$ de ζ vérifiant $\mathrm{Re}(\alpha) > 0$ sont bien tous situés sur la droite $\mathrm{Re}(\alpha) = \frac{1}{2}$. C'est une des questions majeures non résolues des mathématiques contemporaines, qui aurait d'importantes conséquences en arithmétique et en cryptographie ; c'est d'ailleurs l'un des sept "problèmes du millénaire", dont la solution est récompensée par un prix d'un million de Dollars !

Notons que $\frac{1}{n^\alpha} = n^{-\alpha} \geq 1$ lorsque $\alpha \leq 0$, donc nous trouvons $\zeta(\alpha) = +\infty$ avec la définition ci-dessus. Lorsque $\alpha > 0$, la fonction $f(x) = x^{-\alpha}$ est décroissante, et on déduit du Théorème 11.6 que la série $\zeta(\alpha)$ est de même nature que l'intégrale $\int_1^{+\infty} x^{-\alpha}\,dx$. Or

$$\int_1^A x^{-\alpha}\,dx = \Big[\frac{1}{1-\alpha}x^{1-\alpha}\Big]_1^A = \frac{1-A^{1-\alpha}}{\alpha-1} \qquad \text{si } \alpha \neq 1$$

et $\int_1^A x^{-\alpha}\,dx = \ln A$ si $\alpha = 1$, de sorte que la convergence a lieu si et seulement si $\alpha > 1$, avec $\int_1^{+\infty} x^{-\alpha}\,dx = \frac{1}{\alpha-1}$, tandis que $\int_1^{+\infty} x^{-\alpha}\,dx = +\infty$ si $\alpha \leq 1$. Par conséquent **la série de Riemann $\zeta(\alpha)$ converge si et seulement si** $\alpha > 1$. Le Théorème 11.6 donne l'encadrement

$$\frac{1}{\alpha-1} \leq \zeta(\alpha) \leq 1 + \frac{1}{\alpha-1} = \frac{\alpha}{\alpha-1}.$$

On notera en particulier que la "série harmonique" $1 + \frac{1}{2} + \ldots + \frac{1}{n} + \ldots$ est divergente. D'après ce que nous avons vu précédemment, la somme partielle $S_N = 1 + \frac{1}{2} + \ldots + \frac{1}{N}$ satisfait l'encadrement

$$\ln(N) \leq \ln(N+1) \leq S_N \leq 1 + \ln(N),$$

on a donc $S_N \sim \ln(N)$ quand $N \to +\infty$.

Séries absolument convergentes. Lorsque la série $\sum u_n$ n'est plus à termes positifs, on peut essayer de s'y ramener en étudiant la série $\sum |u_n|$, qui est quant à elle à termes positifs.

Définition 11.7. *On dit que la série $\sum u_n$ est absolument convergente si la série des modules $\sum |u_n|$ est convergente.*

Considérons par exemple la série

$$\sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots + \frac{1}{2p-1} - \frac{1}{2p} + \ldots\ .$$

Comme $\big|\frac{(-1)^{n-1}}{n}\big| = \frac{1}{n}$ et que la série harmonique est divergente, on en conclut que la série $\sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{n}$ n'est pas absolument convergente. Cependant, en regroupant les termes consécutifs deux par deux, on observe que ces termes "se compensent presque" :

$$0 < \frac{1}{2p-1} - \frac{1}{2p} = \frac{1}{2p(2p-1)} \sim \frac{1}{4p^2} \qquad \text{quand } p \to +\infty.$$

Comme la série de Riemann $\zeta(2) = \sum_{p=1}^{+\infty} \frac{1}{p^2}$ converge, on en déduit facilement que la série $\sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{n}$ est en fait convergente (on peut montrer que sa somme vaut $\ln(2)$). En général, nous avons le résultat suivant.

Théorème 11.8. *Toute série $\sum u_n$ absolument convergente est convergente.*

Démonstration. L'hypothèse est que $\sum |u_n|$ converge. Lorsque $u_n \in \mathbb{R}$, on utilise la décomposition d'un réel x en sa partie positive $x^+ = \max(x, 0)$ et sa partie négative $x^- = \max(-x, 0)$ (de sorte que par exemple $(-3)^+ = 0$, $(-3)^- = 3$, tandis que la partie négative d'un réel $x \geq 0$ est nulle). Avec cette notation, on vérifie aussitôt que l'on peut écrire

$$x = x^+ - x^-$$

pour tout réel x. Comme $0 \leq (u_n)^+ \leq |u_n|$ et $0 \leq (u_n)^- \leq |u_n|$, l'hypothèse de convergence absolue et le principe de comparaison entraînent que les séries $\sum (u_n)^+$ et $\sum (u_n)^-$ convergent, mais alors

$$\sum u_n = \sum ((u_n)^+ - (u_n)^-)$$

est bien convergente. Lorsque $u_n \in \mathbb{C}$, on écrit $u_n = x_n + iy_n$ avec $x_n, y_n \in \mathbb{R}$. Comme $0 \leq |x_n| \leq |u_n|$ et $0 \leq |y_n| \leq |u_n|$, on en conclut que les séries $\sum x_n$ et $\sum y_n$ sont absolument convergentes (et donc convergentes, puisque le cas réel est déjà démontré). Par conséquent

$$\sum u_n = \sum x_n + iy_n$$

est bien convergente dans $\mathbb{C}$. □

Une application importante à la théorie des séries de Fourier est la suivante.

Théorème 11.9. *Si $\sum_{n\in\mathbb{Z}} c_n$ est une série absolument convergente de nombres complexes, la série trigonométrique*

$$S(x) = \sum_{n=-\infty}^{+\infty} c_n e^{in\omega x}$$

converge vers une fonction $S : \mathbb{R} \to \mathbb{C}$ **continue**. *De plus, si $S_N(x) = \sum_{n=-N}^{+N} c_n e^{in\omega x}$, on a* **convergence uniforme** *de S_N vers S, c'est-à-dire que la "norme L^∞"*

$$\|S - S_N\|_\infty := \sup_{x\in\mathbb{R}} |S(x) - S_N(x)|$$

tend vers 0 quand N tend vers $+\infty$.

Démonstration. L'hypothèse est que la série $\sum_{n\in\mathbb{Z}} |c_n|$ converge. Comme $|c_n e^{in\omega x}| = |c_n|$, il résulte du théorème précédent que les séries $\sum_{n=1}^{+\infty} c_{\pm n} e^{\pm in\omega x}$ convergent en tout point. De plus

$$|S(x) - S_N(x)| = \Big| \sum_{n=N+1}^{+\infty} (c_n e^{in\omega x} + c_{-n} e^{-in\omega x}) \Big| \leq R_N := \sum_{n=N+1}^{+\infty} (|c_n| + |c_{-n}|)$$

et le reste R_N tend bien vers 0 quand N tend vers l'infini puisqu'il s'agit du reste d'une série convergente ; nous avons ici par définition $\|S - S_N\|_\infty \leq R_N$. Fixons maintenant un point $x_0 \in \mathbb{R}$ et $\varepsilon > 0$. Il existe alors N (dépendant de ε) tel que $0 \leq R_N \leq \varepsilon$. Comme la fonction S_N est une somme finie de fonctions trigonométriques, elle est continue au point x_0 et il existe $\delta > 0$ tel que $|x - x_0| \leq \delta$ implique $|S_N(x) - S_N(x_0)| \leq \varepsilon$. Pour $|x - x_0| \leq \delta$ on obtient par conséquent

$$|S(x) - S(x_0)| \leq |S(x) - S_N(x)| + |S_N(x) - S_N(x_0)| + |S_N(x_0) - S(x_0)| \leq R_N + \varepsilon + R_N \leq 3\varepsilon.$$

Ceci montre que la fonction S est bien continue en x_0. □

Séries alternées. Considérons une "série alternée" de terme général $u_n = (-1)^n c_n$ avec une suite (c_n) **positive décroissante tendant vers 0** quand $n \to +\infty$. On symbolise souvent cette situation en écrivant

$$\sum_{n=n_0}^{+\infty} (-1)^n c_n, \qquad c_n \searrow 0.$$

Si on regroupe les termes d'indices $2p$ et $2p+1$, on obtient $c_{2p} - c_{2p+1}$, qui vérifie l'encadrement

$$0 \leq c_{2p} - c_{2p+1} \leq c_{2p} - c_{2p+2}$$

du fait de la décroissance de la suite (c_n). Maintenant, si l'on fait la somme de ces inégalités à partir d'un indice initial pair $n_0 = 2p_0$, on voit que

$$\begin{aligned} S_{2p+1} = (c_{2p_0} - c_{2p_0+1}) + \ldots + (c_{2p} - c_{2p+1}) &\leq (c_{2p_0} - c_{2p_0+2}) + \ldots + (c_{2p} - c_{2p+2}) \\ &= c_{2p_0} - c_{2p+2} \leq c_{2p_0}. \end{aligned}$$

Ceci entraîne que (S_{2p+1}) est une suite croissante majorée, donc la limite $S = \lim_{p\to+\infty} S_{2p+1}$ existe, avec d'ailleurs $0 \leq S \leq c_{2p_0}$. On notera que l'on a aussi $S_{2p} = S_{2p-1} + c_{2p} \to S$, puisque

$\lim_{p\to+\infty} c_{2p} = 0$ par hypothèse. Ceci entraîne immédiatement qu'une série alternée comme ci-dessus est toujours convergente, et que la somme $\sum_{n=n_0}^{+\infty}(-1)^n c_n$ est majorée en valeur absolue par le terme initial c_{n_0} et du même signe $(-1)^{n_0}$ que lui (si n_0 est pair, c'est ce qu'on a vu, et si n_0 est impair, il suffit de changer tous les signes et de décaler les indices d'une unité). En particulier le reste $R_N = S - S_N = \sum_{n=N+1}(-1)^n c_n$ est du signe de $(-1)^{N+1}$ et majoré par c_{N+1}, ce qui permet d'estimer la vitesse de convergence. On déduit par exemple de ce qui précède que la série de Riemann alternée

$$\widetilde{\zeta}(\alpha) = \sum_{n=1}^{+\infty} \frac{(-1)^{n-1}}{n^\alpha}$$

est convergente pour tout $\alpha > 0$. Comme les termes d'indices pairs donnent la contribution

$$-\Big(\frac{1}{2^\alpha} + \frac{1}{4^\alpha} + \ldots + \frac{1}{(2p)^\alpha} + \ldots\Big) = -\frac{1}{2^\alpha}\Big(1 + \frac{1}{2^\alpha} + \ldots + \frac{1}{p^\alpha} + \ldots\Big) = -\frac{1}{2^\alpha}\zeta(\alpha)$$

et qu'il faut l'ajouter deux fois à $\zeta(\alpha)$ pour obtenir $\widetilde{\zeta}(\alpha)$, on voit que

$$\widetilde{\zeta}(\alpha) = \zeta(\alpha) - 2 \times \frac{1}{2^\alpha}\zeta(\alpha) = \Big(1 - \frac{1}{2^{\alpha-1}}\Big)\zeta(\alpha)$$

pour $\alpha > 1$.

12. Séries de Fourier, théorèmes fondamentaux de convergence

Pour $f \in C^0_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$, nous étudions ici la convergence des sommes partielles de la série de Fourier introduite dans la Section 10 :

$$f_N(x) = \sum_{n=-N}^{+N} c_n(f)\, e^{in\omega x} \qquad \text{où} \quad c_n(f) = \frac{1}{T}\int_{x_0}^{x_0+T} f(x)\, e^{-in\omega x}\, dx.$$

Les résultats fondamentaux sont les suivants.

Théorème 12.1 (Dirichlet, 1829)**.** *Supposons que $f \in C^0_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$ admette des dérivées à droite et à gauche en un point $x \in \mathbb{R}$ (il suffit pour cela que f soit de classe C^1 par morceaux). Alors en ce point*

$$\lim_{N\to+\infty} f_N(x) = \sum_{n=-\infty}^{+\infty} c_n(f)\, e^{in\omega x} = \mathrm{VP}(f)(x) = \frac{1}{2}\big(f(x+0) + f(x-0)\big).$$

Théorème 12.2 (Convergence L^∞)**.** *Si f est de classe C^1 par morceaux sans discontinuités, alors f_N converge uniformément vers f (i.e. $\|f - f_N\|_\infty$ tend vers 0 quand $N \to +\infty$).*

Le résultat suivant est dû à Marc-Antoine Parseval (1755–1836).

Théorème 12.3 (Parseval)**.** *Pour toute fonction $f \in C^0_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$ on a*

$$\|f\|_2^2 = \frac{1}{T}\int_{x_0}^{x_0+T} |f(x)|^2 dx = \sum_{n=-\infty}^{+\infty} |c_n(f)|^2.$$

En termes des coefficients $a_n(f)$, $b_n(f)$, ceci s'écrit

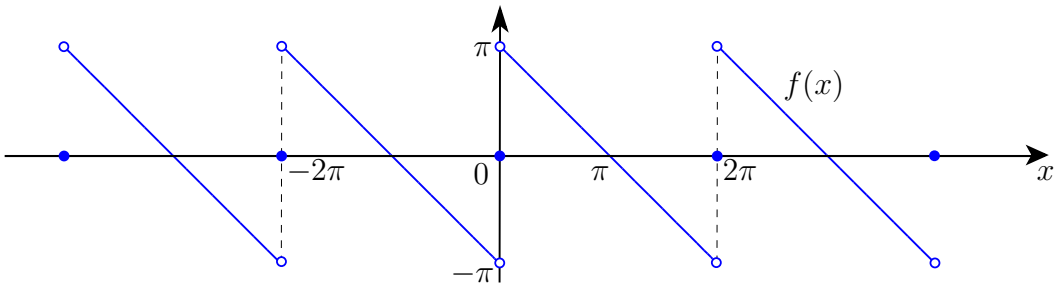
$$\|f\|_2^2 = |c_0(f)|^2 + \frac{1}{2}\sum_{n=1}^{+\infty}\big(|a_n(f)|^2 + |b_n(f)|^2\big).$$

Corollaire 12.4 (Convergence L^2)**.** *Pour toute fonction $f \in C^0_{\mathrm{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$, le polynôme trigonométrique f_N converge en norme L^2 vers f, c'est-à-dire*

$$\lim_{N\to+\infty} \|f - f_N\|_2 = 0.$$

Avant de passer aux démonstrations de ces résultats, donnons d'abord quelques exemples. Lorsqu'on travaille avec des fonctions continues par morceaux, on notera que l'on peut commencer par remplacer f par $\mathrm{VP}(f)$ dans les calculs, car ceci ne change f qu'en un nombre fini de points, par conséquent les coefficients de Fourier $c_n(f)$ ne sont pas modifiés et les conclusions des énoncés de convergence ne sont pas affectés.

Exemple 12.5 (Signal en dent de scie discontinu)**.** Soit f la fonction périodique de période $T = 2\pi$ (i.e. de pulsation $\omega = 1$) telle que $f(x) = \pi - x$ pour $x \in]0, 2\pi[$. On a alors $f(0+0) = \pi$, $f(0-0) = f(2\pi - 0) = -\pi$, donc on pose $f(0) = f(2k\pi) = 0$ afin que f coïncide avec sa valeur principale aux points $2k\pi$, $k \in \mathbb{Z}$.



La fonction f étant impaire, on est amené à calculer uniquement les coefficients $b_n(f)$:

$$b_n(f) = \frac{4}{2\pi}\int_0^\pi (\pi - x)\sin(nx)\,dx = \frac{2}{\pi}\Big[-(\pi - x)\frac{\cos(nx)}{n}\Big]_0^\pi - \frac{2}{\pi}\int_0^\pi \frac{\cos(nx)}{n}\,dx = \frac{2}{n}.$$

Le théorème de Dirichlet implique

$$\sum_{n=1}^{+\infty} \frac{2}{n}\sin(nx) = \mathrm{VP}(f)(x) = f(x) \qquad \text{pour tout } x \in \mathbb{R}.$$

En particulier, pour $x = \pi/2$ et $n = 2p$ pair, on a $\sin(n\pi/2) = \sin(p\pi) = 0$, tandis que pour $n = 2p+1$ impair on a $\sin(n\pi/2) = \sin((2p+1)\pi/2) = (-1)^p$. L'identité précédente nous donne une formule déjà connue du mathématicien indien Mādhava de Sañgamāgrama (1350–1425), et redécouverte par Gregory et Leibniz entre 1670 et 1680 :

$$\sum_{p=0}^{+\infty} \frac{(-1)^p}{2p+1} = 1 - \frac{1}{3} + \frac{1}{5} + \ldots + \frac{(-1)^p}{2p+1} + \ldots = \frac{\pi}{4}.$$

L'identité de Parseval implique d'autre part

$$\frac{1}{2}\sum_{n=1}^{+\infty} |b_n(f)|^2 = \frac{1}{2}\sum_{n=1}^{+\infty} \frac{4}{n^2} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |f(x)|^2 dx = \frac{1}{\pi}\int_0^\pi (\pi - x)^2 dx = \frac{\pi^2}{3},$$

et on en déduit une autre formule célèbre :

$$\zeta(2) = \sum_{n=1}^{+\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \qquad \text{(Euler 1735)}.$$

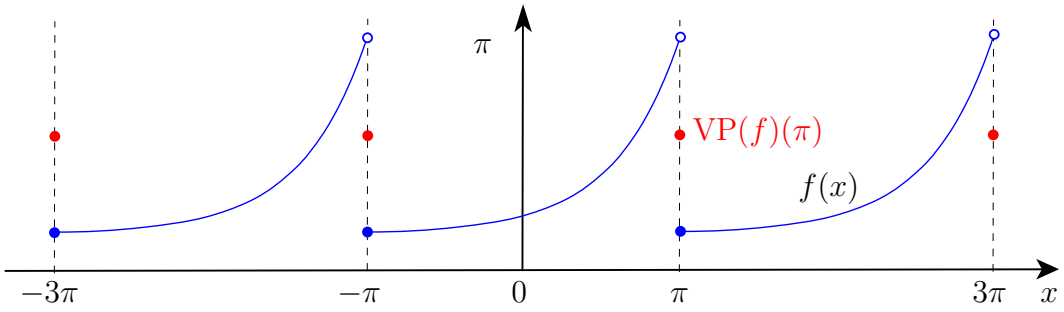
Exemple 12.6. On considère maintenant la fonction f périodique de période 2π telle que $f(x) = e^{ax}$ sur $[-\pi, \pi[$, où a est un paramètre réel non nul. Elle est bien de classe C^1 par morceaux, avec des discontinuités aux points $\pi + 2k\pi$, $k \in \mathbb{Z}$. La fonction f n'est ni paire ni impaire, et on s'aperçoit tout de suite qu'il est beaucoup plus facile de calculer les coefficients complexes $c_n(f)$. En effet, comme $a^{ax}e^{-inx} = e^{(a-in)x}$ et $e^{\pm in\pi} = (-1)^n$, on trouve

$$c_n(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ax}e^{-inx}dx = \frac{1}{2\pi}\Big[\frac{e^{(a-in)x}}{a-in}\Big]_{-\pi}^{\pi} = \frac{1}{2\pi(a-in)}(-1)^n(e^{a\pi} - e^{-a\pi}).$$

On en déduit

$$\begin{aligned}\sum_{n=-\infty}^{+\infty} c_n(f)e^{inx} &= \frac{e^{a\pi}-e^{-a\pi}}{2\pi}\left(\frac{1}{a} + \sum_{n=1}^{+\infty}(-1)^n\Big(\frac{e^{inx}}{a-in} + \frac{e^{-inx}}{a+in}\Big)\right)\\ &= \frac{e^{a\pi}-e^{-a\pi}}{2\pi}\left(\frac{1}{a} + \sum_{n=1}^{+\infty}(-1)^n\frac{2a\,\cos(nx) - 2n\,\sin(nx)}{a^2+n^2}\right),\end{aligned}$$

et cette série a pour somme $\mathrm{VP}(f)(x)$ en tout point, soit $f(x)$ pour $x \neq \pi + 2k\pi$, et $\mathrm{VP}(f)(\pi + 2k\pi) = \frac{1}{2}(e^{a\pi} + e^{-a\pi})$.



En particulier, pour $x = 0$ et $x = \pi$, nous obtenons les formules

$$\frac{e^{a\pi}-e^{-a\pi}}{2\pi}\left(\frac{1}{a}+\sum_{n=1}^{+\infty}(-1)^n\frac{2a}{a^2+n^2}\right) = 1, \qquad \frac{e^{a\pi}-e^{-a\pi}}{2\pi}\left(\frac{1}{a}+\sum_{n=1}^{+\infty}\frac{2a}{a^2+n^2}\right) = \frac{1}{2}\Big(e^{a\pi}+e^{-a\pi}\Big).$$

L'identité de Parseval donne quant à elle

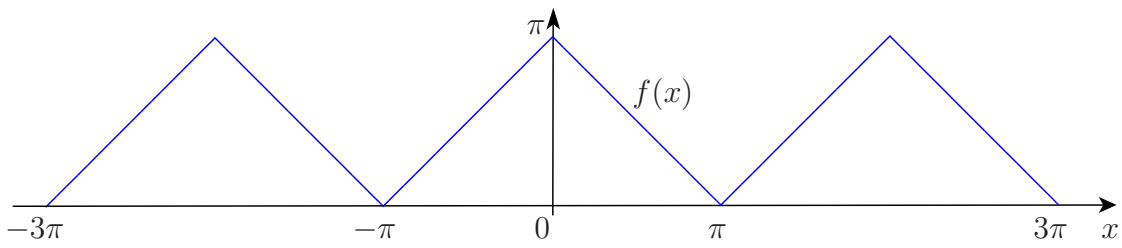
$$\sum_{n=-\infty}^{+\infty}|c_n(f)|^2 = \sum_{n=-\infty}^{+\infty}\frac{(e^{a\pi}-e^{-a\pi})^2}{4\pi^2(a^2+n^2)} = \frac{1}{2\pi}\int_{-\pi}^{\pi}e^{2ax}dx = \frac{e^{2a\pi}-e^{-2a\pi}}{4a\pi}.$$

Ceci fournit la formule

$$\sum_{n=-\infty}^{+\infty}\frac{1}{a^2+n^2} = \frac{\pi}{a}\,\frac{e^{a\pi}+e^{-a\pi}}{e^{a\pi}-e^{-a\pi}}$$

(qui se trouve fortuitement être équivalente à celle trouvée plus haut pour $x = \pi$).

Exemple 12.7 (Signal en dent de scie continu)**.** On considère ici la fonction périodique de période 2π telle que $f(x) = \pi - |x|$ pour $x \in [-\pi, \pi]$.



Il s'agit d'une fonction continue, de classe C^1 par morceaux. La fonction f est paire, on calcule donc les coefficients $a_n(f)$. La valeur moyenne de f est $c_0(f) = \pi/2$ (calcul immédiat), et pour $n \geq 1$ on obtient

$$a_n(f) = \frac{4}{2\pi}\int_0^\pi (\pi - x)\cos(nx)\,dx = \frac{2}{\pi}\Big[(\pi - x)\frac{\sin(nx)}{n}\Big]_0^\pi - \frac{2}{\pi}\int_0^\pi -\frac{\sin(nx)}{n}\,dx$$
$$= \frac{2}{\pi}\Big[-\frac{\cos(nx)}{n^2}\Big]_0^\pi = \frac{2}{\pi}\,\frac{(1-(-1)^n)}{n^2}.$$

Seuls les coefficients $a_n(f)$ d'indices impairs sont non nuls, et on trouve $a_{2p+1}(f) = \frac{4}{\pi}\frac{1}{(2p+1)^2}$. On a $\mathrm{VP}(f) = f$ et le théorème de Dirichlet implique

$$\frac{\pi}{2} + \sum_{p=0}^{+\infty}\frac{4}{\pi}\,\frac{1}{(2p+1)^2}\cos((2p+1)x) = f(x) \qquad \text{pour tout } x \in \mathbb{R}.$$

En particulier, pour $x = 0$ on trouve

$$\pi = \frac{\pi}{2} + \sum_{p=0}^{+\infty}\frac{4}{\pi}\,\frac{1}{(2p+1)^2},$$

donc

$$\sum_{p=0}^{+\infty}\frac{1}{(2p+1)^2} = \frac{\pi^2}{8}.$$

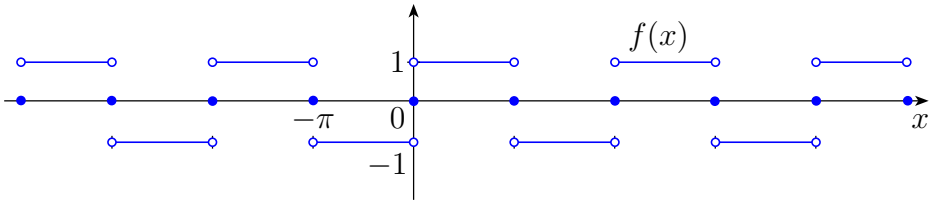
Ce résultat est bien cohérent avec celui déjà trouvé pour $\zeta(2)$, puisque la somme $\sum_{p=0}^{+\infty}\frac{1}{(2p+1)^2}$ est égale à $(1-\frac{1}{4})\zeta(2) = \frac{\pi^2}{8}$ (les termes d'indices pairs dans la série $\zeta(2)$ ayant précisément pour somme $\frac{1}{4}\zeta(2)$). L'identité de Parseval donne ici

$$|c_0(f)|^2 + \frac{1}{2}\sum_{n=1}^{+\infty}|a_n(f)|^2 = \frac{\pi^2}{4} + \frac{1}{2}\sum_{n=1}^{+\infty}\frac{16}{\pi^2}\frac{1}{(2p+1)^4}$$
$$= \frac{1}{2\pi}\int_{-\pi}^{\pi}|f(x)|^2dx = \frac{1}{\pi}\int_0^\pi(\pi - x)^2dx = \frac{\pi^2}{3},$$

et on en déduit

$$\sum_{p=0}^{+\infty}\frac{1}{(2p+1)^4} = \frac{\pi^4}{96} = \Big(1 - \frac{1}{16}\Big)\zeta(4) \qquad \text{d'où} \qquad \zeta(4) = \sum_{n=1}^{+\infty}\frac{1}{n^4} = \frac{\pi^4}{90}.$$

Exemple 12.8. Signal carré (ou signal en créneau). Il s'agit de la fonction f de période 2π telle que $f(x) = 1$ si $x \in]0, \pi[$, $f(x) = -1$ si $x \in]-\pi, 0[$ et $f(k\pi) = 0$.



La fonction f est de classe C^1 par morceaux, impaire, et

$$b_n(f) = \frac{4}{2\pi}\int_0^\pi \sin(nx)\,dx = \frac{2}{\pi}\Big[-\frac{\cos(nx)}{n}\Big]_0^\pi = \frac{2}{\pi}\,\frac{1-(-1)^n}{n},$$

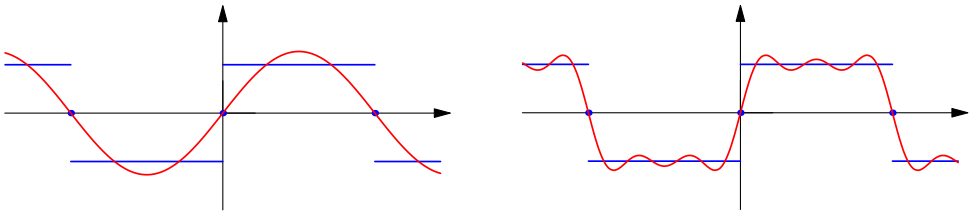
d'où $b_{2p+1}(f) = \frac{4}{\pi}\frac{1}{2p+1}$ et $b_n(f) = 0$ si $n = 2p$ est pair. Comme $f = \mathrm{VP}(f)$, on trouve

$$\sum_{p=0}^{+\infty} \frac{4}{\pi}\frac{1}{2p+1}\sin((2p+1)x) = f(x) \qquad \text{pour tout } x \in \mathbb{R},$$

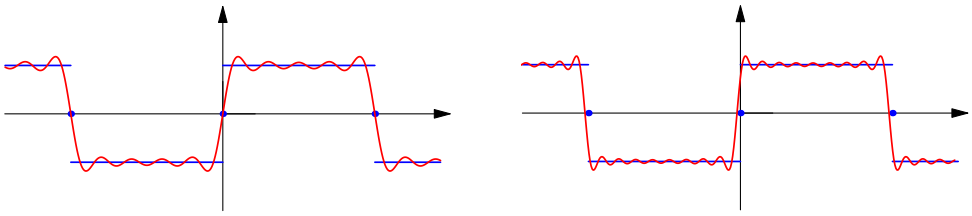
et l'identité de Parseval implique

$$\frac{1}{2}\sum_{n=1}^{+\infty}|b_n(f)|^2 = \frac{1}{2}\sum_{p=0}^{+\infty}\frac{16}{\pi^2}\frac{1}{(2p+1)^2} = \|f\|_2^2 = 1$$

(qui redonne la valeur $\sum_{p=0}^{+\infty}\frac{1}{(2p+1)^2} = \frac{\pi^2}{8}$ déjà connue). Il est particulièrement intéressant dans ce cas d'observer les sommes partielles $f_N(x)$ approchant la fonction en créneau :



$0 \le p \le 0, \quad N = 1$ $\qquad$ $0 \le p \le 2, \quad N = 5$



$0 \le p \le 4, \quad N = 9$ $\qquad$ $0 \le p \le 8, \quad N = 17$

Henry Wilbraham a remarqué en 1848 que des oscillations persistent à proximité des discontinuités, même lorsque l'entier N est pris très grand : les bosses se rapprochent de la discontinuité, mais la hauteur des oscillations au delà de l'intervalle de saut ne tend pas vers 0. Cette observation resta cependant lettre morte jusqu'en 1898, année où Albert Michelson utilisa une table traçante pour obtenir des représentations graphiques plus précises. Il crut d'abord que les oscillations provenaient d'imprécisions de tracé, mais Josiah Willard Gibbs comprit en 1899 qu'il s'agissait bel et bien d'un phénomène mathématique. Celui-ci est connu maintenant sous le nom de **phénomène de Gibbs**.

Pour expliciter le phénomène, calculons la somme partielle $f_N(x)$ pour $N = 2m-1$ et $x = a/2m$. On trouve

$$f_{2m-1}(x) = \sum_{p=0}^{m-1}\frac{4}{\pi}\frac{\sin((2p+1)x)}{2p+1} \quad\Longrightarrow\quad f_{2m-1}\Big(\frac{a}{2m}\Big) = \frac{a}{m}\sum_{p=0}^{m-1}\frac{2}{\pi}\frac{\sin((p+\frac{1}{2})a/m)}{(p+\frac{1}{2})a/m}.$$

On observe que l'expression de droite n'est autre qu'une somme de Riemann de la fonction continue $g(x) = \frac{2}{\pi}\frac{\sin x}{x}$, intégrée sur l'intervalle $[0,a]$, avec une subdivision en m sous-intervalles de longueur a/m et des valeurs prises au milieu $(p+\frac{1}{2})a/m$ de chaque sous-intervalle. Comme la fonction g est continue, il en résulte que

$$\lim_{m\to+\infty} f_{2m-1}\Big(\frac{a}{2m}\Big) = \int_0^a g(x)\,dx = \frac{2}{\pi}\int_0^a \frac{\sin x}{x}\,dx.$$

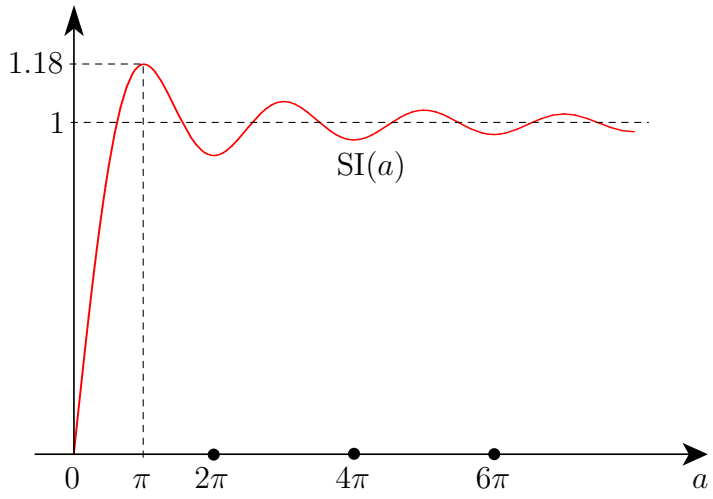
La fonction “sinus intégral” définie par

$$\mathrm{SI}(a) = \frac{2}{\pi}\int_0^a \frac{\sin x}{x}\,dx$$

est une fonction impaire (puisque $x \mapsto \frac{\sin x}{x}$ est paire). Comme $\sin(x+n\pi) = (-1)^n \sin x$, on voit que

$$u_n = \frac{2}{\pi}\int_{n\pi}^{(n+1)\pi} \frac{\sin x}{x}\,dx = (-1)^n \frac{2}{\pi}\int_0^{\pi} \frac{\sin x}{n\pi + x}\,dx$$

définit une série alternée $\sum_{n=0}^{+\infty} u_n$ dont la valeur absolue du terme général $|u_n|$ est décroissante et tend vers 0. On déduit du théorème des séries alternées que la série $\sum_{n=0}^{+\infty} u_n$ est convergente, et donc que l'intégrale $\frac{2}{\pi}\int_0^{+\infty} \frac{\sin x}{x}\,dx$ converge elle aussi ; on montrera plus loin que $\lim_{a\to+\infty} \mathrm{SI}(a) = 1$. La représentation graphique de SI (donnée ici seulement sur $[0,+\infty[$) fournit de façon précise l'allure des oscillations quand $N \to +\infty$:



Un calcul numérique montre que la fonction SI passe par un maximum $\mathrm{SI}(\pi) = u_0 = \frac{2}{\pi}\int_0^{\pi} \frac{\sin x}{x}\,dx \simeq 1.17898\ldots$ (là encore, cela résulte du fait que le signe de u_n est alterné). En général, on peut obtenir le résultat suivant que nous démontrerons plus loin.

Théorème 12.9 (Phénomène de Gibbs)**.** *Les sommes partielles $f_N(x) = \sum_{|n|\le N} c_n(f)\, e^{in\omega x}$ de la série de Fourier d'une fonction $f \in C^1_{\mathrm{morc}}(\mathbb{R}/T\mathbb{Z},\mathbb{C})$ vérifient les estimations suivantes. Si $x_0 < x_1 < \ldots < x_s = x_0 + 2\pi$ sont les points de discontinuité modulo 2π, il existe des constantes $C_1, C_2, C_3, C_4 \ge 0$ telles que pour tout $\delta \in]0, T/4]$, on ait*

- *convergence uniforme de f_N vers f à l'écart des points de discontinuité*:

$$\Big|f_N(x) - f(x)\Big| \le \frac{C_1}{\delta N} + \frac{C_2}{\sqrt{N}} \qquad \text{pour } x \in [x_k+\delta, x_{k+1}-\delta],\ \ 0 \le k \le s-1,$$

- *les “estimées de saut” suivantes au voisinage de chaque point de discontinuité x_k :*

$$\Big|f_N(x_k+h) - \big(\mathrm{VP}(f)(x_k) + \frac{1}{2}\big(f(x_k+0) - f(x_k-0)\big)\,\mathrm{SI}(N\omega h)\Big| \le C_3 h + \frac{C_4}{\sqrt{N}} \quad \text{si } |h| \le \delta.$$

Il en résulte que pour tout $x \in \mathbb{R}$ et tout $a \in \mathbb{R}$ on a

$$\lim_{N\to+\infty} f_N(x + a/N) = \mathrm{VP}(f)(x) + \frac{1}{2}\big(f(x+0) - f(x-0)\big)\,\mathrm{SI}(\omega a).$$

Cette valeur tend vers $f(x \pm 0)$ quand $a \to \pm\infty$ et se réduit simplement à $f(x)$ si x est un point de continuité de f.

Preuve des résultats de convergence fondamentaux. Pour $f \in C^0_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$, on sait que la série $\sum_{n=-\infty}^{+\infty} |c_n(f)|^2$ est convergente et que sa somme est majorée par $\|f\|_2^2$ (inégalité de Bessel). La Proposition 11.2 entraîne déjà le résultat important suivant.

Théorème 12.10 (Lemme de Riemann-Lebesgue)**.** *Pour tout $f \in C^0_{\text{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$ on a*

$$\lim_{n\to\pm\infty} c_n(f) = \lim_{n\to\pm\infty} a_n(f) = \lim_{n\to\pm\infty} b_n(f) = 0.$$

Démonstration. En effet, la convergence de la série $\sum_{n=-\infty}^{+\infty} |c_n(f)|^2$ exige que

$$\lim_{n\to\pm\infty} c_n(f) = 0.$$

Les relations entre les coefficients $a_n(f)$, $b_n(f)$, $c_n(f)$ impliquent que $a_n(f)$, $b_n(f)$ tendent également vers 0 quand $n \to +\infty$ ou $n \to -\infty$. □

On cherche maintenant une expression intégrale explicite pour le polynôme trigonométrique

$$f_N(x) = \sum_{n=-N}^{+N} c_n(f) e^{in\omega x}.$$

Comme

$$c_n(f) = \frac{1}{T}\int_{x_0}^{x_0+T} f(y)\, e^{-in\omega y} dy,$$

nous obtenons

$$f_N(x) = \sum_{n=-N}^{+N} c_n(f) e^{in\omega x} = \frac{1}{T}\int_{x_0}^{x_0+T} f(y) \sum_{n=-N}^{+N} e^{in\omega x}\, e^{-in\omega y} dy,$$

soit encore

$$f_N(x) = \int_{x_0}^{x_0+T} f(y) D_N(x-y)\, dy$$

avec une fonction D_N appelée "noyau de Dirichlet" :

$$D_N(y) = \frac{1}{T}\sum_{n=-N}^{+N} e^{in\omega y}.$$

Comme $\sum_{n=-N}^{+N} e^{in\omega y} = e^{-iN\omega y}(1 + e^{i\omega y} + \ldots + e^{2Ni\omega y})$, on trouve

$$D_N(y) = \frac{1}{T} e^{-iN\omega y} \frac{e^{(2N+1)i\omega y} - 1}{e^{i\omega y} - 1} = \frac{1}{T} e^{-iN\omega y} \frac{e^{(N+1/2)i\omega y}}{e^{i\omega y/2}} \frac{e^{(N+1/2)i\omega y} - e^{-(N+1/2)i\omega y}}{e^{i\omega y/2} - e^{-i\omega y/2}},$$

soit encore plus simplement

$$D_N(y) = \frac{1}{T} \frac{\sin((N+1/2)\omega y)}{\sin(\omega y/2)}.$$

Nous aurons besoin des propriétés suivantes du noyau de Dirichlet :

(1) D_N est une fonction paire, périodique de période T, réelle, et de classe C^∞.

(2) $\displaystyle\int_{-T/2}^{T/2} D_N(y)\, dy = 1, \quad \int_0^{T/2} D_N(y) = \frac{1}{2}.$

Cette deuxième propriété résulte immédiatement de la définition initiale de D_N, car seul le terme constant $e_0(y) = 1$ contribue à l'intégrale par périodicité. Si l'on effectue le changement

de variable $t \mapsto y = x + t \Leftrightarrow t = y - x$ (et donc $dy = dt$) dans la formule donnant $f_N(x)$ on aboutit, par utilisation de la parité et de la périodicité, à

$$\begin{aligned} f_N(x) &= \int_{x_0-x}^{x_0-x+T} f(x+t)D_N(-t)\,dt = \int_{-T/2}^{T/2} f(x+y)D_N(y)\,dy \\ &= \int_0^{T/2} \big(f(x+y)+f(x-y)\big)D_N(y)\,dy. \end{aligned}$$

Si l'on retranche $\mathrm{VP}(f)(x) = \frac{1}{2}\big(f(x+0)+f(x-0)\big)$, il vient

$$f_N(x) - \mathrm{VP}(f)(x) = \int_0^{T/2} \big(f(x+y)+f(x-y)-(f(x+0)+f(x-0))\big)D_N(y)\,dy$$

grâce à la propriété (2).

Supposons maintenant que f admette des dérivées à droite et à gauche au point x. Alors la fonction g définie sur $[-T/2, T/2] \smallsetminus \{0\}$ par

$$\begin{aligned} g(y) &= \frac{f(x+y)+f(x-y)-(f(x+0)+f(x-0))}{y} \\ &= \frac{f(x+y)-f(x+0)}{y} - \frac{f(x-y)-f(x-0)}{-y}, \qquad y \neq 0 \end{aligned}$$

admet une limite à droite en $y = 0$, égale à la différence des dérivées à droite et à gauche de f en x. Puisque g est impaire, elle admet aussi une limite à gauche en 0, et il est clair qu'elle est (comme f) continue par morceaux partout ailleurs qu'en 0. On en conclut que g se prolonge en une fonction continue par morceaux sur $[-T/2, T/2]$. D'après ce qui précède, nous avons

$$f_N(x) - \mathrm{VP}(f)(x) = \frac{1}{2}\int_{-T/2}^{T/2} g(y)\,yD_N(y)\,dy$$

(on a de nouveau utilisé la parité). Or

$$\begin{aligned} yD_N(y) &= \frac{1}{T}\frac{y}{\sin \omega y/2}\sin((N+1/2)\omega y) \\ &= \frac{1}{T}\frac{y}{\sin \omega y/2}\Big(\sin(\omega y/2)\cos(N\omega y) + \cos(\omega y/2)\sin(N\omega y)\Big). \end{aligned}$$

On observe ici que $\frac{y}{\sin \omega y/2}$ se prolonge par continuité en $y = 0$ (puisque $\sin(\omega y/2) \sim \omega y/2$, la limite vaut $2/\omega$) ; comme $\sin(\omega y/2)$ s'annule seulement aux points $y = 2k\pi/\omega = kT$, on voit que $y \mapsto \frac{y}{\sin \omega y/2}$ se prolonge en une fonction continue sur $[-T/2, T/2]$. On en déduit qu'on peut écrire

$$f_N(x) - \mathrm{VP}(f)(x) = \int_{-T/2}^{T/2} \big(u(y)\cos(N\omega y) + v(y)\sin(N\omega y)\big)\,dy$$

avec des fonctions u, v continues par morceaux sur $[-T/2, T/2]$. Le lemme de Riemann-Lebesgue implique alors que ces coefficients de Fourier tendent vers 0. On a donc bien $\lim_{N\to+\infty} f_N(x) = \mathrm{VP}(f)(x)$ et le théorème de Dirichlet 12.1 est démontré. En définitive, la preuve est extrêmement subtile, on peut excuser Fourier que celle-ci lui ait échappé !

Étude de la convergence L^∞. On suppose ici que $f \in C^1_{\mathrm{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$ est de classe C^1 par morceaux et **sans discontinuités**. Alors la dérivée f' est bien définie sauf peut-être en un nombre fini de points anguleux a_k ; si on pose arbitrairement $f'(a_k) = 0$ en ces points,

on obtient une fonction continue par morceaux. Grâce à une intégration par parties, les coefficients de Fourier de f' s'expriment par

$$\widehat{f'}(n) = \frac{1}{T}\int_{x_0}^{x_0+T} f'(x)\, e^{-in\omega x}\, dx = \frac{1}{T}\left(\Big[f(x)\, e^{-in\omega x} \Big]_{x_0}^{x_0+T} - \frac{1}{T}\int_{x_0}^{x_0+T} f(x)\, (-in\omega)\, e^{-in\omega x}\, dx \right).$$

(Pour justifier cette formule, on peut intégrer sur chaque intervalle $[x_k, x_{k+1}]$ où f est vraiment C^1, et on applique la formule de Chasles pour les intégrales pour faire la somme). En utilisant la continuité de f et la périodicité, on voit que le terme $\big[\ \dots \ \big]_{x_0}^{x_0+T}$ est nul, donc

Théorème 12.11. *Si $f \in C^1_{\mathrm{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$ est de classe C^1 par morceaux et* **continue***, on a la formule*

$$\widehat{f'}(n) = in\omega\, \widehat{f}(n).$$

Sous ces hypothèses, on obtient par conséquent

$$\big|\widehat{f}(n)\big| = \frac{1}{n\omega}\, \big|\widehat{f'}(n)\big| \qquad \text{si } n \neq 0.$$

On utilise maintenant l'inégalité élémentaire $ab \leq \frac{1}{2}(a^2 + b^2)$ pour tous $a, b \geq 0$. Il vient

$$\big|\widehat{f}(n)\big| \leq \frac{1}{2\omega^2 n^2} + \frac{1}{2}\big|\widehat{f'}(n)\big|^2,$$

et comme la série $\sum_{n\in\mathbb{Z}} \big|\widehat{f'}(n)\big|^2$ est convergente d'après l'inégalité de Bessel, on en déduit que la série $\sum_{n\in\mathbb{Z}} \big|\widehat{f}(n)\big|$ est convergente. D'après le Théorème 11.9, la suite des sommes partielles $f_N(x) = \sum_{|n|\leq N} \widehat{f}(n)\, e^{inx}$ converge uniformément vers sa somme $S(x)$, qui coïncide avec $f(x)$ par le théorème de Dirichlet, i.e. $\|f - f_N\|_\infty \to 0$ quand $N \to +\infty$, et le Théorème 12.2 est démontré. Mais on a de façon évidente

$$\|f - f_N\|_2^2 = \frac{1}{T}\int_{x_0}^{x_0+T} |f(x) - f_N(x)|^2\, dx \leq \sup_{x\in\mathbb{R}} |f(x) - f_N(x)|^2 = \|f - f_N\|_\infty^2$$

donc $\|f - f_N\|_2 \to 0$. En utilisant l'égalité de Pythagore $\|f\|_2^2 = \|f_N\|_2^2 + \|f - f_N\|_2^2$, ce dernier résultat montre le Théorème 12.3 et le Corollaire 12.4 lorsque f est de classe C^1 par morceaux et continue. Pour généraliser ces résultats à des fonctions f seulement continues par morceaux, on peut utiliser un procédé de moyennisation.

Un procédé de moyennisation. Étant donné une fonction f intégrable et périodique de période T, on lui associe la fonction

$$\mu_\varepsilon f(x) = \frac{1}{2\varepsilon}\int_{x-\varepsilon}^{x+\varepsilon} f(y)\, dy = \frac{1}{2\varepsilon}\int_{-\varepsilon}^{\varepsilon} f(x+y)\, dy,$$

à savoir la valeur moyenne de f sur l'intervalle $[x - \varepsilon, x + \varepsilon]$. Il est clair que $\mu_\varepsilon f$ est continue, et elle est encore périodique de période T. Lorsque $\varepsilon \to 0$, $\mu_\varepsilon f(x) \to f(x)$ en tout point x où f est continue. Si f est seulement continue par morceaux, on voit aisément que $\lim_{\varepsilon\to 0} \mu_\varepsilon f(x) = \mathrm{VP}(f)(x)$ en tout point. De plus $\mu_\varepsilon f$ est C^1 par morceaux puisqu'elle a en tout point des dérivées à droite et à gauche

$$(\mu_\varepsilon f)'(x \pm 0) = \frac{1}{2\varepsilon}(f(x + \varepsilon \pm 0) - f(x - \varepsilon \pm 0))$$

qui sont C^0 par morceaux. D'après ce qui précède, la formule de Parseval s'applique à $\mu_\varepsilon f$; on va en déduire qu'elle s'applique aussi à f par passage à la limite. Des résultats standards de théorie de l'intégration permettent de démontrer que $\lim_{\varepsilon\to 0} \|\mu_\varepsilon f - f\|_2 = 0$ du fait que

$\mu_\varepsilon f \to f$ presque partout ; nous admettrons ici ce résultat dû à Lebesgue. Le théorème de Fubini pour les intégrales doubles donne quant à lui les coefficients de Fourier de $\mu_\varepsilon f$:

$$\begin{aligned}
\widehat{\mu_\varepsilon f}(n) &= \frac{1}{T}\int_{x_0}^{x_0+T} \mu_\varepsilon f(x)\, e^{-in\omega x}\, dx = \frac{1}{2\varepsilon}\frac{1}{T}\int_{x_0\le x\le x_0+T}\int_{-\varepsilon\le y\le\varepsilon} f(x+y)\, e^{-in\omega x}\, dx\, dy \\
&= \frac{1}{2\varepsilon}\frac{1}{T}\int_{-\varepsilon\le y\le\varepsilon} e^{in\omega y}\left(\int_{x_0\le x\le x_0+T} f(x+y)\, e^{-in\omega(x+y)}\, dx\right) dy \\
&= \frac{1}{2\varepsilon}\frac{1}{T}\int_{-\varepsilon\le y\le\varepsilon} e^{in\omega y}\left(\int_{0\le x\le T} f(x)\, e^{-in\omega x}\, dx\right) dy \qquad \text{(par périodicité)} \\
&= \frac{1}{2\varepsilon}\widehat{f}(n)\int_{-\varepsilon\le y\le\varepsilon} e^{in\omega y} dy = \frac{\sin(n\omega\varepsilon)}{n\omega\varepsilon}\widehat{f}(n).
\end{aligned}$$

Puisque $|\sin x| \le x$ pour tout $x \ge 0$, on en déduit de façon quelque peu miraculeuse que $\left|\widehat{\mu_\varepsilon f}(n)\right| \le |\widehat{f}(n)|$ pour tout $n \in \mathbb{Z}$. Par conséquent, comme l'égalité de Parseval est connue pour $\mu_\varepsilon f$ qui est continue et C^1 par morceaux, on obtient

$$\|\mu_\varepsilon f\|_2^2 = \sum_{n\in\mathbb{Z}} \left|\widehat{\mu_\varepsilon f}(n)\right|^2 \le \sum_{n\in\mathbb{Z}} \left|\widehat{f}(n)\right|^2.$$

Par passage à la limite quand $\varepsilon \to 0$, on voit que $\|f\|_2^2 \le \sum_{n\in\mathbb{Z}} \left|\widehat{f}(n)\right|^2$ et les théorèmes fondamentaux 12.3 et 12.4 en résultent.

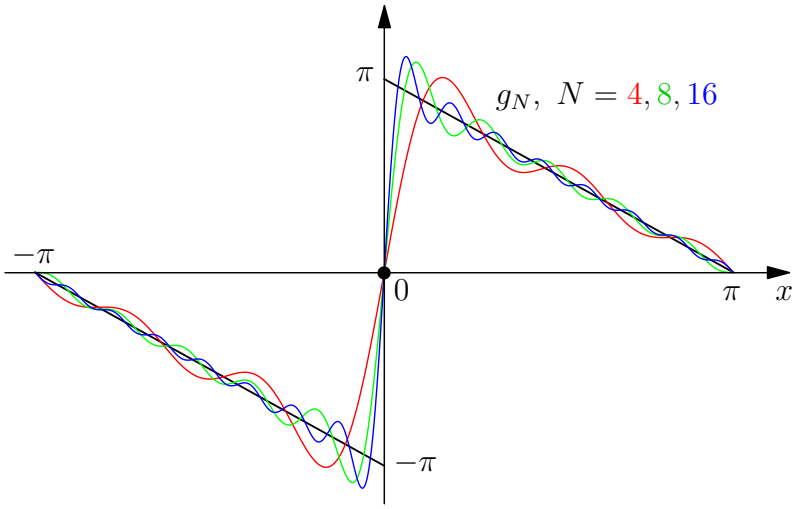
Preuve du Théorème 12.9 (phénomène de Gibbs). Quitte à changer l'unité de temps, il n'est pas restrictif de supposer $T = 2\pi$ et $\omega = 1$. L'étape principale est de vérifier explicitement les inégalités voulues dans le cas de la fonction en dent de scie

$$g(x) = \pi - x \quad x \in]0, 2\pi[\,, \quad g(2k\pi) = 0.$$

D'après les calculs de l'exemple 12.5, nous avons $b_n(g) = \frac{2}{n}$ pour $n \ge 1$ et $a_n(g) = 0$. Pour $x \in]0, 2\pi[$, ceci implique

$$g_N(x) - g(x) = \sum_{n=1}^{N} \frac{2}{n}\sin(nx) - (\pi - x) = \int_\pi^x \Big(1 + 2\sum_{n=1}^{N}\cos(ny)\Big) dy = \int_\pi^x \frac{\sin(\frac{2N+1}{2}y)}{\sin(\frac{y}{2})}\, dy,$$

car la somme entre parenthèses n'est autre que le noyau de Dirichlet.



On effectue maintenant une intégration par parties pour obtenir

$$g_N(x) - g(x) = -\frac{2}{2N+1}\left(\Big[\frac{\cos(\frac{2N+1}{2}y)}{\sin(\frac{y}{2})}\Big]_\pi^x + \int_\pi^x \frac{1}{2}\frac{\cos(\frac{2N+1}{2}y)\cos(\frac{y}{2})}{\sin^2(\frac{y}{2})}\, dy\right).$$

Comme $|\cos(\frac{2N+1}{2}y)| \le 1$, une majoration brutale des différents termes donne

$$|g_N(x) - g(x)| \le \frac{2}{2N+1}\left(\frac{1}{\sin(\frac{x}{2})} + \left|\int_\pi^x \frac{1}{2}\frac{\cos(\frac{y}{2})}{\sin^2(\frac{y}{2})}\,dy\right|\right) = \frac{2}{2N+1}\left(\frac{2}{\sin(\frac{x}{2})} - 1\right)$$

(la fonction $y \mapsto \cos(\frac{y}{2})$ ne change pas de signe sur $[\pi, x]$). Comme $\sin y \ge \frac{2}{\pi}y$ sur $[0, \frac{\pi}{2}]$ par concavité de la fonction sin, et comme $\delta \le \frac{T}{4} = \frac{\pi}{2}$, ceci montre déjà que

$$\sup_{x\in[\delta,2\pi-\delta]} |g_N(x) - g(x)| \le \frac{2}{2N+1}\,\frac{2}{\sin(\frac{\delta}{2})} \le \frac{1}{N}\,\frac{2}{\frac{2}{\pi}\frac{\delta}{2}} = \frac{2\pi}{N\delta}.$$

Pour $f = g$ et $\omega = 1$, on veut maintenant estimer les quantités

$$f_N(x_k + h) - \big(\mathrm{VP}(f)(x_k) + \frac{1}{2}\big(f(x_k+0) - f(x_k-0)\big)\,\mathrm{SI}(N\omega h)$$

aux points de discontinuité x_k. Modulo 2π, la fonction g admet seulement le point de discontinuité $x_k = 0$, avec $g(x_k+0) - g(x_k-0) = 2\pi$ et $\mathrm{VP}(g)(x_k) = 0$. Il nous faut donc estimer

$$\begin{aligned}
g_N(h) - \pi\ \mathrm{SI}(Nh) &= \sum_{n=1}^{N} \frac{2}{n}\sin(nh) - 2\int_0^{Nh} \frac{\sin y}{y}\,dy \\
&= \int_0^h \sum_{n=1}^{N} 2\cos(ny)\,dy - 2\int_0^{Nh}\frac{\sin y}{y}\,dy \\
&= \int_0^h \left(\frac{\sin(\frac{2N+1}{2}y)}{\sin(\frac{y}{2})} - 1\right)dy - 2\int_0^{(N+1/2)h}\frac{\sin y}{y}\,dy + 2\int_{Nh}^{(N+1/2)h}\frac{\sin y}{y}\,dy \\
&= \int_0^h \sin\Big(\frac{2N+1}{2}y\Big)\left(\frac{1}{\sin(\frac{y}{2})} - \frac{1}{\frac{y}{2}}\right)dy - 2\int_{Nh}^{(N+1/2)h}\frac{\sin y}{y}\,dy - h
\end{aligned}$$

où la dernière ligne s'obtient à la suite du changement de variable $y \mapsto \frac{2N+1}{2}y$ dans l'intégrale $\int_0^{(N+1/2)h}\frac{\sin y}{y}\,dy$. Comme $|\sin y| \le y$, la dernière intégrale est majorée par $2(h/2) = h$. On obtient donc

$$\big|g_N(h) - \pi\ \mathrm{SI}(Nh)\big| \le 2|h| + \left|\int_0^h \sin\Big(\frac{2N+1}{2}y\Big)\,u(y)\,dy\right| \qquad \text{où} \quad u(y) = \frac{1}{\sin(\frac{y}{2})} - \frac{1}{\frac{y}{2}}.$$

On peut vérifier au moyen d'un développement limité que la fonction u se prolonge en une fonction de classe C^1 (et même de classe C^∞) sur $]-2\pi, 2\pi[$, en posant $u(0) = 0$. En utilisant une intégration par parties on voit alors que $\big|\int_0^h \sin\big(\frac{2N+1}{2}y\big)\,u(y)\,dy\big| \le \frac{C}{2N+1} \le \frac{C}{N}$, donc

$$(*) \qquad \big|g_N(h) - \pi\ \mathrm{SI}(Nh)\big| \le 2|h| + \frac{C}{N}.$$

Le cas de la fonction $f = g$ est établi. Dans le cas où la fonction f est de classe C^1 par morceaux et **continue**, on sait qu'il y a convergence uniforme sur tout intervalle $[x_0, x_0+2\pi]$. Une majoration à l'aide des coefficients de Fourier de f et f' donne

$$\begin{aligned}
\big|f_N(x) - f(x)\big| &\le \left|\sum_{|n|\ge N+1} \widehat{f}(n)\,e^{inx}\right| \le \sum_{|n|\ge N+1} |\widehat{f}(n)| = \sum_{|n|\ge N+1} \frac{1}{n}\,|\widehat{f'}(n)| \\
&\le \sqrt{\sum_{|n|\ge N+1}\frac{1}{n^2}}\sqrt{\sum_{|n|\ge N+1}|\widehat{f'}(n)|^2} \\
&\le \sqrt{\frac{1}{N}}\,\|f'\|_2
\end{aligned}$$

où la deuxième ligne résulte de l'inégalité de Cauchy-Schwarz et la dernière de l'estimation du reste de la série de Riemann par l'intégrale $\int_N^{+\infty} \frac{1}{x^2}\,dx$ combinée avec l'inégalité de Bessel pour f'. On a donc bien $\big|f_N(x) - f(x)\big| \leq C' \frac{1}{\sqrt{N}}$ et les estimations voulues sont vérifiées pour f.

Dans le cas général, lorsque $f \in C^1_{\mathrm{morc}}(\mathbb{R}/T\mathbb{Z}, \mathbb{C})$ présente des points de discontinuités x_k, elle admet des "sauts" d'amplitude $f(x_k+0)-f(x_k-0)$, et on peut lui retrancher la fonction en dent de scie $x \mapsto \frac{1}{2\pi}\big(f(x_k+0) - f(x_k-0)\big)\, g(x-x_k)$ pour faire disparaître ces sauts, puisque ceux de $x \mapsto g(x-x_k)$ sont d'amplitude 2π en $x = x_k$. Ceci montre que

$$\widetilde{f}(x) = f(x) - \sum_k \frac{1}{2\pi}\big(f(x_k+0) - f(x_k-0)\big)\, g(x-x_k)$$

n'a plus de sauts, et donc que $\widetilde{f}$ est continue (quitte à la redéfinir convenablement en x_k). La fonction $\widetilde{f}$ est aussi clairement de classe C^1 par morceaux. Si l'on écrit

$$f(x) = \widetilde{f}(x) + \sum_k \frac{1}{2\pi}\big(f(x_k+0) - f(x_k-0)\big)\, g(x-x_k),$$

les estimations de Gibbs sont vérifiées pour $\widetilde{f}$ et pour chaque terme de la sommation, il ne reste plus qu'à en faire la somme pour obtenir le résultat désiré : même si x_k n'est pas un point de discontinuité de $\widetilde{f}$ ou de $g(x-x_\ell)$, $\ell \neq k$, il faut quand même s'assurer que $\widetilde{f}_N$ et les $g_N(x-x_\ell)$ vérifient les estimées de saut en x_k ; cela résulte de l'estimée uniforme prise au point $x_k + h$ et du fait qu'on a

$$|\widetilde{f}(x_k+h) - \widetilde{f}(x_k)| \leq C''|h|$$

d'après le théorème des accroissements finis. □

Remarque 12.12. Posons $\mathrm{SI}(+\infty) = \lim_{x\to+\infty} \mathrm{SI}(x)$. Comme

$$\lim_{N\to+\infty} g_N(h) = g(h) = \pi - h \qquad \text{pour } h \in\,]0, 2\pi[,$$

l'estimation $(*)$ implique $|\pi - h - \pi\,\mathrm{SI}(+\infty)| \leq 2h$, donc $|1 - \mathrm{SI}(+\infty)| \leq 3h/\pi$ et puisque ceci est vrai pour tout $h > 0$ on en déduit que $\mathrm{SI}(+\infty) = 1$. Ce résultat n'est pas du tout évident à trouver directement, car la fonction $x \mapsto \frac{\sin x}{x}$ n'a pas de primitive explicitable !

Remarque 12.13. Lorsque la fonction f est de classe C^{p-1} et de classe C^p par morceaux avec $p \geq 2$, un raisonnement par récurrence montre que $\widehat{f^{(p)}}(n) = (in)^p \widehat{f}(n)$, de sorte qu'on a la majoration d'erreur améliorée

$$\begin{aligned}
\big|f_N(x) - f(x)\big| &\leq \Big|\sum_{|n|\geq N+1} \widehat{f}(n)\, e^{inx}\Big| \leq \sum_{|n|\geq N+1} |\widehat{f}(n)| = \sum_{|n|\geq N+1} \frac{1}{n^p}\, |\widehat{f^{(p)}}(n)| \\
&\leq \sqrt{\sum_{|n|\geq N+1} \frac{1}{n^{2p}}} \sqrt{\sum_{|n|\geq N+1} |\widehat{f^{(p)}}(n)|^2} \\
&\leq \sqrt{\frac{1}{2p-1}\,\frac{1}{N^{2p-1}}}\, \|f^{(p)}\|_2 \leq \frac{C_p}{N^{p-1/2}}.
\end{aligned}$$

On voit que la convergence est d'autant plus rapide que f est régulière, autrement dit les harmoniques tendent vers 0 d'autant plus vite que p est grand.

Ceci a également pour conséquence que l'on peut améliorer les termes $C_*/\sqrt{N}$ en C_*/N dans les estimées de Gibbs lorsque la fonction f est de classe C^2 par morceaux. Pour le prouver, il convient d'éliminer également les points anguleux, ce qu'on peut faire en soustrayant à $\widetilde{f}$ des translatées de fonctions du type $h(x) = x^2$ sur $[-\pi, \pi]$, pour lesquelles on a convergence

uniforme en $|h_N - h| \leq C/N$ (exercice !) ; la fonction restante $\widetilde{f}$ devient alors de classe C^1 et C^2 par morceaux, et on a donc d'après ce qu'on vient de voir $|\widetilde{f}_N - \widetilde{f}| \leq C'/N^{3/2}$.

Séries de Fourier des fonctions L^2. Nous indiquons ici brièvement des développements plus modernes de la théorie des séries de Fourier, postérieurs aux années 1900. Mathématiquement, il n'est en effet pas très naturel de supposer que les fonctions sont continues par morceaux, puisque les coefficients de Fourier d'une fonction f se calculent dès lors que la fonction f est intégrable. L'intégrale de Riemann n'est cependant pas satisfaisante dans ce cadre ; pour obtenir des résultats complets on doit s'appuyer sur l'intégrale de Lebesgue. À la fin des années 1950, Jaroslav Kurzweil et Ralph Henstock ont montré qu'il suffisait d'une très petite amélioration de la définition de l'intégrale de Riemann pour y parvenir. C'est celle que nous donnerons ici (afin de rester parfaitement rigoureux et de ne pas donner d'énoncés mathématiquement incorrects, mais nous ne développerons pas vraiment la théorie...).

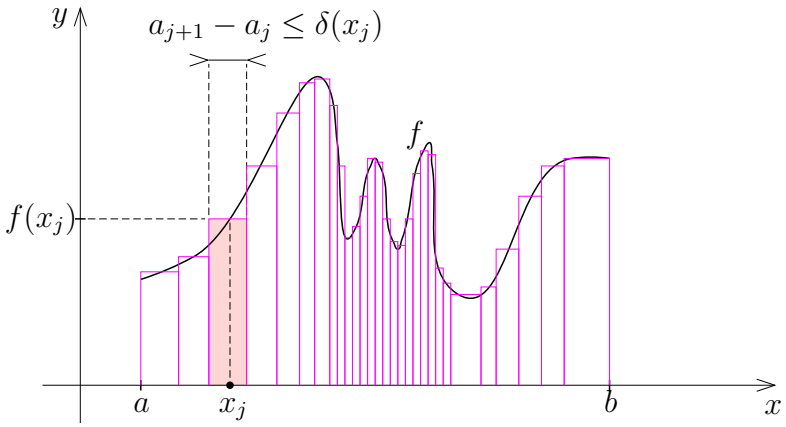
Définition 12.14 (Intégrale de Kurzweil-Henstock)**.** *Soit $f : [a,b] \to \mathbb{C}$ une fonction quelconque. Étant donné une* **subdvision pointée** $D = \{([a_j, a_{j+1}], x_j)\}_{0 \leq j < m}$ *de $[a,b]$, c'est-à-dire des sous-intervalles $[a_j, a_{j+1}]$*

$$a = a_0 < a_1 < a_2 < \ldots < a_{p-1} < a_m = b,$$

et des points intermédiaires $x_j \in [a_j, a_{j+1}]$, on introduit la somme de Riemann

$$S_D(f) = \sum_{j=0}^{m-1} (a_{j+1} - a_j) f(x_j)$$

qui représente la somme des aires algébriques des rectangles s'appuyant sur le graphe :



On dit que f est intégrable au sens de Kurzweil-Henstock (ou KH-intégrable) s'il existe un réel A (qui va représenter l'aire algébrique exacte située sous le graphe de f), tel que pour toute marge d'erreur $\varepsilon > 0$ donnée a priori, on peut trouver une fonction positive $\delta : [a,b] \to \mathbb{R}_+^*$ en sorte que pour toute subdivision pointée D de $[a,b]$ on ait

$$\Big(D \text{ est } \delta\text{-fine, i.e. } \forall j,\ a_{j+1} - a_j \leq \delta(x_j)\Big) \quad \Longrightarrow \quad |S_D(f) - A| \leq \varepsilon.$$

Le nombre réel A de la définition précédente est appelé intégrale de f sur $[a,b]$, on écrit

$$A = \int_a^b f(x)\,dx = \lim_{\text{KH, } D} S_D(f) = \lim_{\text{KH, } D} \sum_{j=0}^{m-1} (a_{j+1} - a_j) f(x_j)$$

et on dit que $\int_a^b f(x)\,dx$ est la limite (au sens de Kurzweil-Henstock) des sommes de Riemann, lorsque la subdivision D devient de plus en plus fine.

La différence avec la définition usuelle de l'intégrale de Riemann est que l'on s'autorise à prendre des pas $a_{j+1}-a_j$ variables, contrôlés par une fonction $\delta(x)>0$ quelconque, que l'on peut prendre très petite aux endroits où la fonction f est "mauvaise" (fortes oscillations). En fait rien n'interdit non plus de prendre f non bornée, par exemple l'intégrale $\int_c^1 x^{-\alpha}dx$ converge quand $c\to 0_+$ et $\alpha<1$; il suffira de prendre des pas $a_{j+1}-a_j$ de plus en plus petits quand on se rapproche de 0 (et donc une fonction de contrôle des pas δ telle que $\lim_{x\to 0}\delta(x)=0$ très vite, choisie pour avoir une précision ε suffisante).

On peut maintenant introduire les espaces L^p de Lebesgue pour tout $p\geq 1$.

Définition 12.15. *On dit qu'une fonction $f:[a,b]\to\mathbb{C}$ est L^p (au sens de Lebesgue) si f est KH-intégrable et si $|f|^p$ est également KH-intégrable. On définit la semi-norme L^p de f par*

$$\|f\|_p=\Big(\int_a^b |f(x)|^p\Big)^{1/p},$$

et on note $\widetilde{L}^p([a,b],\mathbb{C})$ l'ensemble des fonctions L^p à valeurs complexes sur $[a,b]$.

L'espace $\widetilde{L}^2([a,b],\mathbb{C})$ se trouve être muni d'une forme sesquilinéaire hermitienne semi-positive

$$\langle f,g\rangle=\int_a^b \overline{f(x)}g(x)\,dx,\qquad f,g\in\widetilde{L}^2([a,b],\mathbb{C}),$$

l'intégrabilité de $\overline{f}g$ étant garantie par le fait que $|\,\overline{f}g\,|\leq\frac{1}{2}(|f|^2+|g|^2)$. Une petite difficulté est qu'il y a beaucoup de vecteurs isotropes ; on dit qu'un ensemble $E\subset[a,b]$ est **négligeable** si sa fonction caractéristique $\mathbf{1}_E$ (telle que $\mathbf{1}_E(x)=1$ si $x\in E$ et $\mathbf{1}_E(x)=0$ si $x\notin E$) est KH-intégrable et vérifie $\int_a^b \mathbf{1}_E(x)\,dx=0$. Alors :

Théorème 12.16 (Admis)**.** *Une fonction f dans $\widetilde{L}^p$ est de semi-norme $\|f\|_p=0$ si et seulement si il existe un ensemble négligeable E telle que $f(x)=0$ pour tout $x\in[a,b]\smallsetminus E$ (f pouvant prendre des valeurs quelconques sur E du fait que E est "très petit").*

On notera qu'un ensemble négligeable peut tout de même avoir une infinité de points, par exemple on peut montrer que $[a,b]\cap\mathbb{Q}$ est négligeable (et même que tout ensemble dénombrable est négligeable). Un autre exemple classique d'ensemble négligeable infini est **l'ensemble triadique de Cantor** $E\subset[0,1]$, défini comme l'ensemble des nombres réels de l'intervalle $[0,1]$ pouvant s'écrire en base 3 sous la forme

$$x=\sum_{n=1}^{+\infty} a_n 3^{-n},\qquad a_n=0 \text{ ou } a_n=2,$$

autrement dit l'ensemble des réels de $[0,1]$ n'ayant en base 3 que les chiffres 0 ou 2. Il est clair par définition que $E=\bigcap E_N$ où E_N est la réunion des 2^N intervalles de la forme $[r,r+3^{-N}]$, $r=\sum_{n=1}^N a_n3^{-n}$ décrivant l'ensemble des nombres triadiques à N chiffres ne s'écrivant qu'avec des chiffres 0 et 2. Les ensembles E_N s'obtiennent itérativement en partant de l'intervalle $E_0=[0,1]$ et en otant le tiers médian de chacun des intervalles constituant E_{N-1}. La longueur totale des intervalles constituant E_N est $2^N\cdot 3^{-N}=(2/3)^N\to 0$ quand $N\to+\infty$, ce qui implique que E est négligeable $\big($on a $0\leq\int_0^1\mathbf{1}_E(x)\,dx\leq\int_0^1\mathbf{1}_{E_N}(x)\,dx=(2/3)^N\big)$.

Les ensembles triadiques E_N, $0\leq N\leq 6$.

Pour éviter d'avoir à considérer des vecteurs isotropes, on convient d'identifier deux fonctions f et g dès lors qu'il existe un ensemble négligeable $E_{f,g}$ (dépendant de f et g) tel que f et g ne diffèrent que sur $E_{f,g}$, et coïncident sur $[a,b] \smallsetminus E_{f,g}$. On note $L^p([a,b],\mathbb{C})$ (sans tilda) l'espace obtenu après avoir fait cette identification des fonctions. D'après ce que nous avons expliqué plus haut, les éléments isotropes sont maintenant identifiés à la fonction 0. On en déduit :

Théorème 12.17. *La forme sesquilinéaire ci-dessus est définie positive sur $L^2([a,b],\mathbb{C})$, par conséquent $(L^2([a,b],\mathbb{C}),\langle\ ,\ \rangle)$ est un espace hermitien.*

On introduit de même l'espace hermitien $(L^2(\mathbb{R}/T\mathbb{Z},\mathbb{C}),\langle\ ,\ \rangle)$ des fonctions L^2 périodiques de période T, identifiées modulo les ensemble négligeables, avec

$$\langle f,g\rangle = \frac{1}{T}\int_{x_0}^{x_0+T} \overline{f(x)}g(x)\,dx, \qquad f,g\in L^2(\mathbb{R}/T\mathbb{Z},\mathbb{C}).$$

On calcule comme précédemment les coefficients de Fourier $c_n(f)$ (et de même les coefficients $a_n(f)$, $b_n(f)$) d'une fonction $f\in L^2(\mathbb{R}/T\mathbb{Z},\mathbb{C})$.

Théorème 12.18 (Identité de Parseval, version L^2)**.** *Pour toute fonction $f\in L^2(\mathbb{R}/T\mathbb{Z},\mathbb{C})$, on a*

$$\sum_{n\in\mathbb{Z}}|c_n(f)|^2 = |c_0(f)|^2 + \frac{1}{2}\sum_{n=1}^{+\infty}\big(|a_n(f)|^2+|b_n(f)|^2\big) = \|f\|_2^2 = \frac{1}{T}\int_{x_0}^{x_0+T}|f(x)|^2dx.$$

Démonstration. La preuve est essentiellement la même que celle que nous avons déjà vue : la projection orthogonale f_N de f sur $\mathcal{P}_{N,\omega}$ satisfait l'inégalité

$$\|f_N\|_2^2 = \sum_{|n|\le N}|c_n(f)|^2 \le \|f\|_2^2.$$

En effet, ceci résulte du théorème de Pythagore, qui implique $\|f_N\|_2^2+\|f-f_N\|_2^2=\|f\|_2^2$. On a donc à la limite $\sum_{n\in\mathbb{Z}}|c_n(f)|^2\le\|f\|_2^2$. Pour vérifier l'inégalité opposée $\|f\|_2^2\le\sum_{n\in\mathbb{Z}}|c_n(f)|^2$, on effectue une moyennisation $\mu_\varepsilon f$, et on utilise le résultat de théorie de l'intégrale de Lebesgue qui dit que $\|\mu_\varepsilon f - f\|_2\to 0$. Mais on démontre aussi dans cette théorie que $\mu_\varepsilon f$ est une fonction continue, et l'identité de Parseval déjà connue dans ce cas (Théorème 12.3) implique

$$\|\mu_\varepsilon f\|_2^2 = \sum_{n\in\mathbb{Z}}|c_n(\mu_\varepsilon f)|^2 \le \sum_{n\in\mathbb{Z}}|c_n(f)|^2, \qquad \text{car}\ \ |c_n(\mu_\varepsilon f)| = \Big|\frac{\sin(n\omega\varepsilon)}{n\omega\varepsilon}\Big|\,|c_n(f)|\le|c_n(f)|.$$

À la limite il vient $\|f\|_2^2\le\sum_{n\in\mathbb{Z}}|c_n(f)|^2$. On notera que ce raisonnement fournit dans tous les cas l'inégalité intéressante $\|\mu_\varepsilon f\|_2\le\|f\|_2$. □

Un autre espace hermitien qui intervient naturellement dans la théorie L^2 est l'espace noté $\ell^2(\mathbb{Z},\mathbb{C})$ des **suites de carré sommable**. C'est par définition l'ensemble des suites $s=(s_n)_{n\in\mathbb{Z}}$ à valeurs complexes telles que

$$\|s\|_2^2 := \sum_{n\in\mathbb{Z}}|s_n|^2 < +\infty, \qquad \text{avec}\quad \langle s,t\rangle := \sum_{n\in\mathbb{Z}}\overline{s_n}\,t_n.$$

La convergence du produit scalaire est assurée par le fait que $|\,\overline{s_n}\,t_n\,|\le\frac{1}{2}(|s_n|^2+|t_n|^2)$, et il est clair que l'on obtient ainsi un espace hermitien. Une autre façon de formuler la théorie L^2 des séries de Fourier est de dire que l'on a un isomorphisme isométrique d'espaces hermitiens

$$L^2(\mathbb{R}/T\mathbb{Z},\mathbb{C})\longrightarrow\ell^2(\mathbb{Z},\mathbb{C}), \qquad f\longmapsto(c_n)_{n\in\mathbb{Z}}, \qquad \text{où}\ \ c_n=\widehat{f}(n)=\frac{1}{T}\int_{x_0}^{x_0+T}f(x)\,e^{-in\omega x}dx$$

qui préserve les normes L^2 et donc aussi les produits scalaires hermitiens : c'est le théorème de Riesz-Fischer (1907), qui affirme entre autres la "complétude" de $L^2(\mathbb{R}/T\mathbb{Z},\mathbb{C})$ et la surjectivité de l'isométrie précédente. L'isomorphisme inverse est donné par

$$\ell^2(\mathbb{Z},\mathbb{C}) \longrightarrow L^2(\mathbb{R}/T\mathbb{Z},\mathbb{C}), \qquad (c_n)_{n\in\mathbb{Z}} \longmapsto f, \quad \text{où} \quad f(x) = \sum_{n\in\mathbb{Z}} c_n e^{in\omega x}.$$

En fait, le théorème de Riesz-Fischer est impliqué par le résultat très profond suivant dû à Lennart Carleson (mathématicien suédois né en 1928) – la démonstration est beaucoup trop difficile pour pouvoir être donnée ici.

Théorème 12.19 (Carleson, 1966)**.** *Pour toute suite de carré sommable $(c_n) \in \ell^2(\mathbb{Z},\mathbb{C})$, la série trigonométrique*

$$f(x) = \sum_{n\in\mathbb{Z}} c_n e^{in\omega x}$$

converge presque partout, c'est-à-dire qu'elle converge sur tout intervalle $[x_0, x_0+T]$ de longueur T, sauf éventuellement sur un ensemble négligeable $E_f \subset [x_0, x_0+T]$. La limite f ainsi définie est dans $L^2(\mathbb{R}/T\mathbb{Z},\mathbb{C})$, et ses coefficients de Fourier sont les (c_n).

On notera que les polynômes trigonométriques $f_N(x) = \sum_{|n|\le N} c_n e^{in\omega x}$ convergent vers f pour la norme L^2, puisque le théorème de Parseval implique

$$\|f - f_N\|_2^2 = \sum_{|n|>N} |c_n|^2$$

(ce résultat de convergence L^2 est beaucoup plus simple à justifier que le théorème de convergence ponctuelle de Carleson !) Il s'agit là des premières bribes de la *théorie des espaces de Hilbert*, qui jouent un rôle important en théorie des opérateurs et en mécanique quantique.

Solution de l'équation de la chaleur. Revenons maintenant à l'équation de la chaleur discutée à la section 9 :

$$\frac{\partial\theta}{\partial t} = D\frac{\partial^2\theta}{\partial x^2}, \qquad x\in[0,L], \quad t\in[0,+\infty[.$$

Comme on l'a vu, le flux de chaleur est nul aux extrémités du barreau $[0,L]$, ce qui conduit à imposer les conditions au bord

$$\frac{\partial\theta}{\partial x}(0,t) = \frac{\partial\theta}{\partial x}(L,t) = 0 \qquad \text{pour tous } x\in[0,L] \text{ et } t\in[0,+\infty[.$$

On cherche une solution $\theta(x,t)$ telle que $\theta(x,0)$ soit une distribution de température $f(x)$ donnée a priori au temps $t=0$, définie pour $x\in[0,L]$.

Il est naturel (ne serait-ce que pour pouvoir écrire l'équation de la chaleur) de supposer que $\theta(x,t)$ est deux fois dérivable par rapport à x et une fois dérivable par rapport au temps t. En particulier $f(x) = \theta(x,0)$ doit être supposée deux fois dérivable sur $[0,L]$. Cette hypothèse implique a fortiori que f et f' sont continues (puisque dérivables), donc que f est de classe C^1 sur $[0,L]$; en fait, il nous suffira ici de supposer que f est continue et de classe C^1 par morceaux.

À la section 9, nous avons trouvé des solutions de la forme

$$(*) \qquad \theta(x,t) = \sum_{n=0}^{+\infty} \alpha_n \cos\Big(\frac{n\pi}{L}x\Big)\, e^{-(n^2\pi^2/L^2)Dt}.$$

Pour que cela soit une solution de notre problème, il faut déjà que pour $t = 0$ on ait

$$f(x) = \sum_{n=0}^{+\infty} \alpha_n \cos\left(\frac{n\pi}{L}x\right).$$

Le membre de droite est une fonction paire en x. Si l'on prolonge artificiellement f en une fonction paire $\widetilde{f}$ sur $[-L, 0]$ en posant $\widetilde{f}(-x) = f(x)$ pour $x \in [0, L]$, on obtient une fonction continue et de classe C^1 par morceaux sur $[-L, L]$. Comme $\widetilde{f}(-L) = \widetilde{f}(L)$, on peut même prolonger $\widetilde{f}$ en une fonction périodique de période $T = 2L$ sur $\mathbb{R}$ tout entier de sorte que $\widetilde{f}$ soit continue et de classe C^1 par morceaux. Dans ces conditions, le Théorème 12.2 montre que la fonction paire $\widetilde{f}$ s'écrit comme une série absolument convergente en cosinus, de pulsation $\omega = 2\pi/T = 2\pi/2L = \pi/L$:

$$\widetilde{f}(x) = \sum_{n=0}^{+\infty} a_n \cos\left(\frac{n\pi}{L}x\right), \qquad \sum_{n=0}^{+\infty} |a_n| < +\infty.$$

On est donc amené à choisir $\alpha_n = a_n$ dans $(*)$. Nous allons en déduire le magnifique

Théorème 12.20 (Fourier, 1807 ... et successeurs)**.** *Soit $f : [0, L] \to \mathbb{R}$ une fonction continue et de classe C^1 par morceaux sur $[0, L]$, et soient $a_n = a_n(f)$ les coefficients de Fourier définis par*

$$a_0 = c_0 = \frac{1}{L}\int_0^L f(x)\,dx, \qquad a_n = \frac{2}{L}\int_0^L f(x)\,\cos(n\pi x/L)\,dx, \quad n \geq 1.$$

Alors l'équation de la chaleur

$$\frac{\partial\theta}{\partial t} = D\frac{\partial^2\theta}{\partial x^2}, \qquad x \in [0, L], \quad t \in]0, +\infty[$$

admet une unique solution continue $\theta : [0, L] \times [0, +\infty[\to \mathbb{R}$ telle que $\frac{\partial\theta}{\partial t}, \frac{\partial^2\theta}{\partial x^2}$ existent et soient continues sur $[0, L] \times]0, +\infty[$, vérifiant les conditions au bord

$$\frac{\partial\theta}{\partial x}(0, t) = \frac{\partial\theta}{\partial x}(L, t) = 0, \quad \text{et} \quad \theta(x, 0) = f(x) \quad \text{pour tout } x \in [0, L] \text{ et tout } t \in]0, +\infty[.$$

Elle est donnée par

$$\theta(x, t) = \sum_{n=0}^{+\infty} a_n \cos\left(\frac{n\pi}{L}x\right) e^{-(n^2\pi^2/L^2)Dt}.$$

Démonstration. Posons

$$\psi(x, t) = \sum_{n=0}^{+\infty} a_n \cos\left(\frac{n\pi}{L}x\right) e^{-(n^2\pi^2/L^2)Dt}.$$

Nous savons que $A = \sum_{n=0}^{+\infty} |a_n| < +\infty$, donc la série donnant $\psi(x, t)$ est absolument convergente pour tout $(x, t) \in [0, L] \times [0, +\infty[$. On en déduit comme pour la démonstration du Théorème 11.9 que ψ est bien continue sur $[0, L] \times [0, +\infty[$, avec $\psi(x, 0) = f(x)$ grâce au choix $\alpha_n = a_n(f)$. Comme

$$\left|n^p a_n\, e^{-(n^2\pi^2/L^2)Dt}\right| \leq \frac{A\, n^p}{e^{(n^2\pi^2/L^2)Dt}}$$

et que l'exponentielle l'emporte sur tout polynôme, on voit que

$$\left|n^p a_n\, e^{-(n^2\pi^2/L^2)Dt}\right| \leq \frac{\text{Cte}}{n^2} \qquad \text{pour } t \in [t_0, +\infty[, \ \ t_0 > 0.$$

Ceci entraîne facilement que l'on obtient des séries absolument et uniformémement convergentes sur $[0,L]\times[t_0,+\infty[$, après avoir pris autant de dérivées que l'on veut en x ou en t de chacun des termes $a_n\cos\Big(\frac{n\pi}{L}x\Big)\,e^{-(n^2\pi^2/L^2)Dt}$. Par des théorèmes standards (admis ici) sur les séries de fonctions différentiables, on en déduit que ψ admet en fait des dérivées partielles continues en (x,t) sur $[0,L]\times]0,+\infty[$, à tout ordre. En particulier

$$\frac{\partial\psi}{\partial x}(x,t)=\sum_{n=0}^{+\infty}\Big(-\frac{n\pi}{L}\Big)\,a_n\sin\Big(\frac{n\pi}{L}x\Big)\,e^{-(n^2\pi^2/L^2)Dt}$$

s'annule bien pour $x=0$ et $x=L$, et

$$\frac{\partial^2\psi}{\partial x^2}(x,t)=\sum_{n=0}^{+\infty}-\Big(\frac{n\pi}{L}\Big)^2\,a_n\cos\Big(\frac{n\pi}{L}x\Big)\,e^{-(n^2\pi^2/L^2)Dt}$$

coïncide avec $\dfrac{1}{D}\,\dfrac{\partial\psi}{\partial t}(x,t)$.

Ceci montre que $\psi(x,t)$ est une solution du problème, et il reste à voir que c'est la seule. Si $\theta(x,t)$ est une autre solution, Posons $u(x,t)=\theta(x,t)-\psi(x,t)$. La linéarité de l'équation de la chaleur entraîne

$$(**)\qquad\qquad \frac{\partial u}{\partial t}(x,t)=D\frac{\partial^2 u}{\partial x^2}(x,t)\qquad\text{sur } [0,L]\times]0,+\infty[,$$

tandis que $\frac{\partial u}{\partial x}(0,t)=\frac{\partial u}{\partial x}(L,t)=0$ pour tout $t>0$ et $u(x,0)=f(x)-f(x)=0$. Introduisons maintenant la norme L^2

$$J(t)=\int_0^L u(x,t)^2\,dx,\qquad t\in[0,+\infty[.$$

Nous allons montrer que J est identiquement nulle, ce qui entraînera que $u(x,t)=0$ pour tout $(x,t)\in[0,L]\times[0,+\infty[$ d'après l'hypothèse de continuité de u. Notons déjà que $J(0)=0$ par construction.

Les théorèmes usuels de "continuité et dérivabilité sous le signe somme" (admis ici) impliquent que $t\mapsto J(t)$ est continue pour $t\in[0,+\infty[$ (du fait que $u(x,t)$ est continue sur $[0,L]\times[0,+\infty[$), et que $J(t)$ admet une dérivée $J'(t)$ obtenue en prenant la dérivée partielle de l'intégrande par rapport à t :

$$J'(t)=\int_0^L\frac{\partial}{\partial t}\big(u(x,t)^2\big)\,dx=\int_0^L 2\,u(x,t)\,\frac{\partial u}{\partial t}(x,t)\,dx.$$

Compte tenu de l'équation $(**)$, on peut écrire

$$J'(t)=D\int_0^L 2\,u(x,t)\,\frac{\partial^2 u}{\partial x^2}(x,t)\,dx.$$

En intégrant par parties par rapport à la variable x, il vient, en tenant compte des conditions au bord $\frac{\partial u}{\partial x}(0,t)=\frac{\partial u}{\partial x}(L,t)=0$:

$$J'(t)=-2D\int_0^L\left(\frac{\partial u(x,t)}{\partial x}\right)^2dx.$$

Par conséquent $J'(t)\le 0$ pour tout $t>0$. Ainsi J est décroissante, et on a

$$J(t)\le J(0)=0\quad\text{pour tout } t\ge 0.$$

Mais $J(t)$ est l'intégrale d'une fonction positive ou nulle, et on a donc aussi $J(t)\ge 0$ pour tout $t\ge 0$. On en déduit que $J(t)=0$, par suite $u(x,t)=0$ et $\theta(x,t)=\psi(x,t)$ pour tout $x\in[0,L]$ et tout $t\in[0,+\infty[$. □

Remarque 12.21. Lorsque le temps t tend vers $+\infty$, on peut facilement vérifier que la solution satisfait $\lim_{t\to+\infty}\theta(x,t) = a_0$, car toutes les exponentielles tendent rapidement vers 0 à l'exception du terme constant $n = 0$. On a donc une température qui atteint à la limite une température d'équilibre, laquelle se calcule comme étant la valeur moyenne

$$a_0 = c_0 = \frac{1}{L}\int_0^L f(x)\,dx$$

de la température du barreau au temps $t = 0$. Ce résultat se fonde bien entendu sur l'hypothèse que la déperdition d'énergie du barreau vers l'extérieur est négligeable (d'où en particulier les hypothèses de flux $\frac{\partial\theta}{\partial x}(0,t) = \frac{\partial\theta}{\partial x}(L,t) = 0$ aux extrémités). En première approximation, cette hypothèse est réaliste dans la mesure où la conductivité thermique d'un métal est élevée, bien plus élevée que celle de l'atmosphère. Les exponentielles $e^{-(n^2\pi^2/L^2)Dt}$ décroissent donc suffisamment vite pour que la dissipation de chaleur vers l'atmosphère n'ait pas le temps de jouer un rôle. En pratique, la dispersion de température $\max|\theta(x,t) - a_0|$ au temps t est principalement contrôlée par le terme $n = 1$, on voit qu'elle est majorée par $C\,e^{-(\pi^2/L^2)Dt}$ où $C > 0$ est une constante.