Formes quadratiques, séries de Fourier (Mat404)

Version actuelle : Thierry Bouche, janvier 2024
À partir des cours de Bernard Parisse, 2021
et Jean-Pierre Demailly, 2013

Table des matières

Chapitre 1

Motivations

Les séries de Fourier permettant d'écrire une fonction périodique (par exemple un signal périodique) comme une somme de fonctions périodiques fondamentales (sinus et cosinus, ou exponentielle imaginaire pure). Le but est de simplifier la résolution de problèmes qui vérifient le principe de superposition et faisant intervenir des fonctions périodiques en se ramenant à ces fonctions périodiques fondamentales

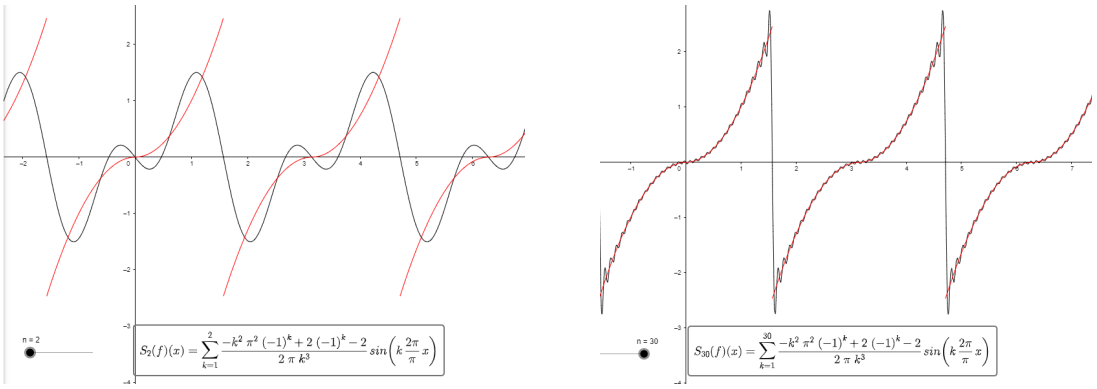


FIGURE 1.1 – Écriture approchée de x^2 (rendue impaire et périodique de période 2π) comme somme de fonctions sinusoïdes fondamentales

Une application immédiate des séries de Fourier est l'analyse d'un son. Si on gratte sur une corde de guitare, on observe un phénomène périodique en temps, qui se décompose en une somme de sinusoïdes dont la fréquence est un multiple entier de la fréquence de base. Pour une même note de musique (par exemple un *la* à 440 Hz), une guitare, un piano, une flûte ne donneront pas le même son parce que les harmoniques sont différents. Voir en fig. 1.2 deux sons purs et un son composé.

On pourrait ainsi numériser le son en stockant les coefficients des sinusoïdes pour la fréquence de base et de ses multiples (les harmoniques) jusqu'à la limite de sensibilité de l'oreille humaine. D'une certaine manière c'est ce que fait une partition de musique en donnant une succession de notes d'une certaine durée à jouer par des instruments de musique (chaque note jouée par un instrument correspondant en quelques sorte à une série de Fourier). Si on représente graphiquement la liste des coefficients des harmoniques en fonction des multiples de la fréquence de base, on obtient le spectre, qui donne une description complète du son (et qu'on peut manipuler avec des logicielspour faire l'analyse spectrale du son, supprimer des harmoniques trop aigües, etc.).

Plus généralement, on parle d'*analyse spectrale*. Cette idée de décomposer en somme de fonctions périodiques « pures » s'applique à diverses généralisations des séries de Fourier : la transformée de Fourier (qui peut servir à comprendre la lumière, les couleurs correspondant à des fréquences, mais vues comme un paramètre continu variant dans $\mathbb{R}^+$ et non discret restreint aux harmoniques d'une fréquence de base), et la transformée de Fourier discrète, adaptée au calcul sur machine.

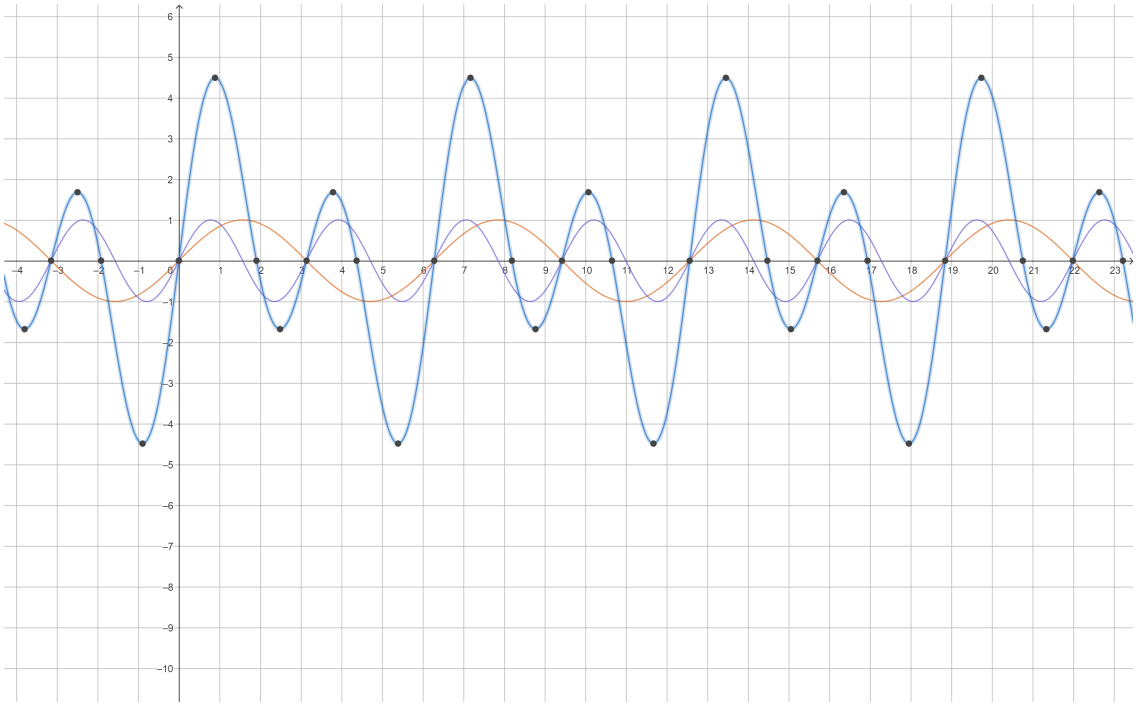


FIGURE 1.2 – Deux sinusoïdes fondamentales et leur somme pondérée.

Un exemple plus mathématique, si on veut résoudre une équation différentielle *linéaire* à coefficients constants avec second membre périodique (ressort soumis à un forçage périodique en temps, circuit RLC soumis à une source périodique en temps, ...), on a des formules simples pour trouver une solution particulière si le second membre est un sinus ou un cosinus (impédance complexe). Comme l'équation est linéaire, le principe de superposition s'applique (pour obtenir la solution particulière correspondant à un second membre somme de deux fonctions, il suffit de faire la somme des solutions particulières correspondant à chacune des deux fonctions). Bien sûr, on sait résoudre ces équations différentielles avec un second membre quelconque, mais la forme de la solution n'est pas toujours explicite et même si elle l'est, elle peut être compliquée et ne pas faire apparaitre certaines propriétés. L'existence de certains phénomènes, par exemple d'une fréquence de résonance ou d'un filtre passe-haut ou passe-bas, et la décomposition en somme de fréquences va permettre de mettre en évidence des propriétés de la solution particulière plus facilement.

Historiquement, les séries de Fourier ont été inventées par Joseph FOURIER (1768-1830) pour résoudre le problème de la diffusion de la chaleur. On ne sait pas résoudre analytiquement l'équation de la chaleur, mais on va voir qu'on sait le faire lorsqu'on décompose la température initiale en somme de cosinus. On va aussi voir que la méthode utilisée pour l'équation de la chaleur est suffisamment générale pour s'appliquer dans d'autres cas, par exemple pour l'équation des ondes (qui elle se résout analytiquement).

Mathématiquement, les concepts qui interviennent sont

1) de l'algèbre linéaire (principe de superposition) ;
2) des sommes (de fonctions sinusoïdes) qui ne sont pas finies (puisqu'il y a une infinité de multiples entiers d'une fréquence de base), on les appelle des *séries*. Ces séries sont plus difficiles à étudier que des sommes de nombres réels, car il s'agit de fonctions. Pour donner un sens à la valeur d'une somme infinie de fonctions, il faut donner un sens à « être petit » pour une fonction. Pour les séries de Fourier, le bon cadre pour cela est
3) une généralisation en dimension infinie de la géométrie euclidienne ou hermitienne (selon qu'on travaille avec des coefficients réels ou complexes). Les formes quadratiques particulières qui interviennent pour les séries de Fourier sont

4) des produits scalaires qui généralisent le produit scalaire usuel dans $\mathbb{R}^2$ et $\mathbb{R}^3$. Pour bien maîtriser cela, nous allons tout d'abord étudier
5) des formes bilinéaires dans un espace vectoriel.

1.1 L'équation de la chaleur

Considérons une tige chauffée de façon inhomogène, par exemple une tige métallique qui vient de servir à remuer les braises d'un feu de bois. Comment se diffuse la chaleur dans cette tige ?

On a donc une tige de longueur finie L dont la température initiale (au temps $t = 0$) en un point d'abscisse x est donnée par une fonction $T_{\text{init}}(x) = T(x,t)_{|t=0}$, $x \in [0,L]$. Dans l'exemple de la tige retirée du feu de bois, si l'extrémité de la tige est en $x = L$, alors $T_{\text{init}}(x)$ est une fonction croissante de x ($T_{\text{init}}(L)$ vaut peut-être 100 degrés, alors que $T_{\text{init}}(0)$ est proche de 20 degrés). On suppose que les échanges de chaleur entre la tige et l'air sont négligeables et que les extrémités de la tige sont au contact d'un parfait isolant, ce qui implique qu'il n'y a pas de flux de chaleur à travers ces extrémités. En particulier le gradient de la température y est nul. On veut comprendre comment la chaleur se diffuse dans la barre avec le temps ; autrement dit, si $T(x,t)$ est la température dans la tige au point x au temps t, alors on veut comprendre l'évolution de la valeur de $T(x,t)$ avec t.

Si la température croit lorsque x augmente, la chaleur va aller vers les x décroissant, d'autant plus vite que $\partial T/\partial x$ est grand. Si on considère un petit élément de tige entre x et $x+dx$, la chaleur entrante en $x+dx$ est proportionnelle à $(\partial T/\partial x)(x+dx)$ et la chaleur sortante en x à $(\partial T/\partial x)(x)$ donc on a un bilan de chaleur entrant de $(\partial T/\partial x)(x+dx) - (\partial T/\partial x)(x)$, qui va réchauffer le morceau de tige entre x et $x+dx$, donc est proportionnel à $(\partial T/\partial t)dx$ Les lois de la physique entrainent donc que T doit satisfaire à l'équation, dite équation de la chaleur :

$$\frac{\partial T}{\partial t} = k\frac{\partial^2 T}{\partial x^2}$$

où k est une constante positive (la *diffusivité*) qui dépend du matériau (proportionnelle à sa conductivité thermique).

Nous avons en plus les conditions au bord

$$\frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(L,t) = 0 \text{ pour tout } t,$$

qui traduisent l'absence de flux de chaleur à travers les extrémités, et la condition initiale

$$T(x,0) = T_{\text{init}}(x).$$

Oublions d'abord la condition $T(x,0) = T_{\text{init}}(x)$. Autrement dit, on cherche les solutions vérifiant seulement les conditions au bord

$$\frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(L,t) = 0 \text{ pour tout } t.$$

L'équation étant beaucoup trop compliquée pour être résolue avec les méthodes dont nous disposons actuellement, nous allons commencer par simplement chercher des *exemples* de fonctions qui la satisfont. Les fonctions *à variables séparées* (c'est-à-dire s'écrivant dans la forme $T(x,t) = f(x)g(t)$) sont une source féconde d'exemples satisfaisant à des équations aux dérivées partielles, puisque de telles équations se simplifient souvent dans ce cas. Nous commencerons donc par chercher des solutions de la forme $T(x,t) = f(x)g(t)$. On a alors :

$$f(x)g'(t) = kf''(x)g(t),$$

soit

$$\frac{f''(x)}{f(x)} = \frac{g'(t)}{kg(t)},$$

au moins sur la région ou ni f ni g ne s'annule. Notons que le membre de gauche est une fonction qui ne dépend que de x et le membre de droite est une fonction qui ne dépend que de t : comme x et t sont

indépendantes, cela implique qu'elles sont constantes. Il existe donc $\alpha \in \mathbb{R}$ tel que

$$\frac{f''(x)}{f(x)} = \frac{g'(t)}{kg(t)} = \alpha.$$

Ainsi, on a

$$f''(x) - \alpha f(x) = 0$$

et

$$g'(t) - k\alpha g(t) = 0.$$

On a donc $g(t) = \lambda e^{k\alpha t}$ pour $\lambda \in \mathbb{R}$, et donc $g(t) \neq 0$ pour tout $t \geqslant 0$ (car on cherche T non identiquement nulle). La contrainte

$$\frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(L,t) = 0$$

entraîne alors $f'(0) = f'(L) = 0$. Pour résoudre l'équation en f il nous faut maintenant distinguer 3 cas.

1) Cas 1 : $\alpha = 0$. On a alors $f''(x) = 0$, et donc $f(x) = b_0 x + a_0$. Les conditions $f'(0) = f'(L) = 0$ imposent alors facilement $f(x) = a_0$ pour tout x. On a donc une première solution de base

$$T_0(x,t) = 1.$$

2) Cas 2 : $\alpha > 0$. On peut exclure ce cas par des considérations physiques, car g serait exponentiellement croissante. D'un point de vue mathématique, on peut alors poser $\alpha = \omega^2$ et f est de la forme $f(x) = ae^{\omega x} + be^{-\omega x}$. Les conditions que $f'(0) = 0$ et $f'(L) = 0$ impliquent alors $a = b = 0$, et f est identiquement nulle, ce qui est exclu.

3) Cas 3 : $\alpha < 0$. On peut alors poser $\alpha = -\omega^2$ et

$$f(x) = a\cos(\omega x) + b\sin(\omega x), a, b, \in \mathbb{R}.$$

Puisque $f'(0) = 0$ on a $b = 0$, et puisque $f'(L) = 0$ on a $a\sin(\omega L) = 0$. Puisque l'on cherche T non nulle, on a $a \neq 0$ et donc $\sin(\omega L) = 0$.
Ainsi $\omega L = \pi n$ pour $n \geqslant 0$ entier (remarque : ceci quantifie les ω possibles qui prennent une suite discrète de valeurs), et donc pour chaque n, on a une solution de la forme

$$T_n(x,t) = \cos\left(\frac{n\pi x}{L}\right) e^{-\frac{\pi^2 n^2}{L^2} kt}.$$

Pour chaque entier positif $n \geqslant 0$ nous avons donc une solution de l'équation de la chaleur

$$T_n(x,t) = \cos\left(\frac{n\pi x}{L}\right) e^{-\frac{\pi^2 n^2}{L^2} kt}.$$

(Nous pouvons intégrer la solution $T_0(x,t) = 1$ dans cette famille de solutions en considérant qu'il s'agit de $T_0(x,t) = \cos(0x)e^{-0t}$.) La condition initiale $T_{\text{init},n}(x)$ correspondant à la solution $T_n(x,t)$ est donnée par $T_{\text{init},n}(x) = T_n(x,0)$, c'est-à-dire

$$T_{\text{init},n}(x) = \cos\left(\frac{n\pi x}{L}\right).$$

Nous avons donc trouvé une solution à l'équation de la chaleur pour certaines conditions initiales bien particulières, c'est-à-dire certains cosinus. Est-ce qu'on peut en construire d'autres solutions pour d'autres conditions initiales ?

Notons tout d'abord que l'équation de la chaleur à une propriété très utile :

Remarque 1.1.1 (Linéarité de l'équation de la chaleur.)**.** Si $T_1(x,t)$ et $T_2(x,t)$ sont deux solutions à l'équation de la chaleur alors pour tous réels $\lambda, \mu \in \mathbb{R}$

$$T(x,t) = \lambda T_1(x,t) + \mu T_2(x,t)$$

est encore une solution de cette équation. (Une telle fonction est appelée une *combinaison linéaire* de T_1 et T_2). On dit alors que l'équation de la chaleur est une *équation linéaire*.

Exercice 1.1.2. Démontrer que l'équation de la chaleur est une équation linéaire.

En particulier, toute fonction qui est une combinaison linéaire finie

$$T(x,t) = \lambda_0 T_0(x,t) + \lambda_1 T_1(x,t) + \lambda_2 T_2(x,t) + \cdots + \lambda_n T_n(x,t)$$

avec des nombres réels $\lambda_0, \ldots, \lambda_n$ est encore une solution de l'équation de la chaleur. Cette solution correspond à la condition initiale

$$T_{\text{init}}(x) = T(x,0)$$

c'est-à-dire

$$T_{\text{init}}(x) = \lambda_0 + \lambda_1 \cos\left(\frac{\pi x}{L}\right) + \lambda_2 \cos\left(\frac{2\pi x}{L}\right) + \cdots + \lambda_n \cos\left(\frac{n\pi x}{L}\right).$$

Nous savons donc trouver une solution pour l'équation de la chaleur pour certaines conditions initiales bien particulières : celles qui s'écrivent comme des sommes finies de cosinus.

Et il vient assez naturellement l'idée : *Peut-on résoudre cette équation de la même façon pour une condition initiale T_{init} quelconque en l'approchant par des sommes avec un très grand nombre de termes ?*

Supposons que, pour chaque n, on choisisse parmi les fonctions de la forme

$$\lambda_0 + \lambda_1 \cos\left(\frac{\pi x}{L}\right) + \lambda_2 \cos\left(\frac{2\pi x}{L}\right) + \cdots + \lambda_n \cos\left(\frac{n\pi x}{L}\right)$$

celle qui est « la plus proche » de T_{init}, appelons-la $S_n(T_{\text{init}})$. On a alors une suite infinie de fonctions $S_1(T_{\text{init}})$, $S_2(T_{\text{init}})$,..., qui approchent « de mieux en mieux » la fonction T_{init}. En effet, le nombre de termes augmente, donc l'approximation ne peut que s'améliorer.

Ainsi, pour chaque condition initiale $S_n(T_{\text{init}})$ « voisine de T_{init} », il existe une solution $T_n(T_{\text{init}})$ de l'équation de la chaleur. L'idée géniale de Fourier est de dire qu'on peut passer à la limite, et résoudre ainsi toutes les équations avec second membre « raisonnable ». Ceci signifie que, quand $n \to \infty$:

1) la suite $S_n(T_{\text{init}})$ converge vers T_{init} et
2) la suite $T_n(T_{\text{init}})$ converge vers une solution de l'équation de la chaleur pour la condition initiale T_{init}.

1.2 L'équation des ondes

Pour illustrer que la méthode utilisée pour l'équation de la chaleur est assez générale, nous allons voir qu'elle peut s'appliquer à une équation que l'on sait résoudre autrement : l'équation des ondes.

Un fil horizontal de longueur L, soumis à une tension T et de densité linéaire μ, est tenu aux deux extrémités Par exemple une corde de guitare de longueur $L = 3$ pincée en un point d'abscisse 1 et d'ordonnée très petite (0.2 sur le dessin) aura le profil suivant (figure 1.3).

Au temps $t = 0$ il est relaché et se met à osciller librement dans un plan vertical.

Soit $y(x,t)$ la fonction égale à la position verticale (par rapport à l'équilibre) à l'instant t de la partie du fil qui se trouve (à l'équilibre) à une distance x d'une des extrémités

Nous avons cette fois les conditions aux bords

$$y(0,t) = y(L,t) = 0,$$

qui traduisent le fait que le fil est attaché aux extrémités. Si le déplacement initial du fil est décrit par la fonction $y_{\text{init}}(x)$ alors nous avons aussi les conditions initiales

$$y(x,0) = y_{\text{init}}(x) \text{ et } \frac{\partial y}{\partial t}(x,0) = 0,$$

cette dernière condition traduisant le fait que le fil est relâché à l'instant $t = 0$ et se trouve donc à ce moment-là au repos.

Si on considère le morceau de fil compris entre les abscisses x et $x + dx$, il est soumis à deux forces :

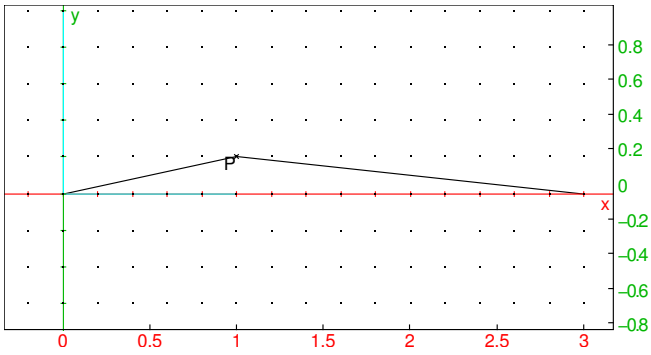


FIGURE 1.3 – Une corde de guitare pincée.

— en $x+dx$, une traction d'intensité T de direction et sens le vecteur directeur de la tangente $(1, y'(x+dx))$
— en x, une traction opposée portée par $(1, y'(x))$

Le principe fondamental de la dynamique donne alors

$$\mu \frac{\partial^2 y}{\partial t^2} dx = T(y'(x+dx) - y'(x))$$

L'évolution de y est décrite (au premier ordre, car on a fait comme si le vecteur $(1, y')$ était normé, et on n'a pas tenu compte de la possible variation locale de tension si y' est non nul) par l'équation des ondes

$$\frac{\partial^2 y}{\partial t^2} = c^2 \frac{\partial^2 y}{\partial x^2}$$

où c est la constante positive $c^2 = T/\mu$.

Si la fonction y_{init} est deux fois dérivables, on sait déterminer la solution de cette équation : on prolonge y_{init} en une fonction f_0 définie sur $\mathbb{R}$ qui est impaire et de période $2L$), on a alors :

$$y(x,t) = \frac{1}{2}(f_0(x+ct) + f_0(x-ct)).$$

Cherchons maintenant comme ci-dessus des solutions de la forme $f(x)g(t)$. On a

$$f(x)g''(t) = c^2 f''(x)g(t),$$

soit

$$\frac{f''(x)}{f(x)} = \frac{g''(t)}{c^2 g(t)}.$$

Notons que le membre de gauche est une fonction qui ne dépend que de x et le membre de droite est une fonction qui ne dépend que de t : comme x et t sont deux variables indépendantes, cela implique à nouveau qu'il existe $\alpha \in \mathbb{R}$ tel que

$$\frac{f''(x)}{f(x)} = \frac{g''(t)}{c^2 g(t)} = \alpha.$$

Ainsi, on a

$$f''(x) - \alpha f(x) = 0 \text{ et } g''(t) - c^2 \alpha g(t) = 0.$$

Le même raisonnement que ci-dessus nous montre que cette équation a une solution telle que $y(0,t) = y(L,t) = 0$ si et seulement si il existe un entier n tel que $\alpha = -\frac{n^2\pi^2}{L^2}$ et dans ce cas on a une solution donnée par

$$y_n(x,t) = \sin\left(\frac{n\pi x}{L}\right) \cos\left(\frac{cn\pi t}{L}\right).$$

Ceci nous donne une solution au problème pour une condition initiale

$$Y_n(x) = \sin\left(\frac{n\pi x}{L}\right).$$

On vérifie bien que $y_n(x,t) = \frac{1}{2}(Y_n(x+ct) + Y_n(x-ct))$. On peut remarquer aussi que Y_n est impaire et $2L$-périodique.

Remarque 1.2.1. L'équation des ondes est encore une équation linéaire,

Exercice 1.2.2. Démontrer que l'équation des ondes est linéaire.

Puisque la fonction $y_n(x,t)$ est une solution pour chaque n, toute combinaison linéaire finie

$$y(x,t) = \lambda_1 y_1(x,t) + \lambda_2 y_2(x,t) + \cdots + \lambda_k y_k(x,t)$$

où les λ_k sont des nombres réels est encore une solution de l'équation de la chaleur. Cette solution correspond à la condition initiale

$$y_{\text{init}}(x) = \lambda_1 \sin\left(\frac{\pi x}{L}\right) + \lambda_2 \sin\left(\frac{2\pi x}{L}\right) + \cdots + \lambda_n \sin\left(\frac{n\pi x}{L}\right).$$

Nous savons donc trouver une solution à cette équation pour des conditions initiales bien particulières : celles qui s'écrivent comme des sommes finies de sinus.

De même que dans le cas de l'équation de la chaleur : *Peut-on résoudre cette équation pour une condition initiale quelconque y_{init} en écrivant y_{init} comme une « somme infinie » de sinus ?*

Cette idée d'écrire y_{init} comme une somme infinie de fonctions trigonométriques par approximations successives est séduisante, mais pose beaucoup de questions :

1) Quel sens donner à une somme infinie de fonctions ?
2) Qu'est-ce que cela signifie quand on dit qu'une suite de fonctions converge vers une autre fonction ? (probablement que le distance tend vers 0).
3) Qu'est-ce que ça veut dire, quand on dit que deux fonctions sont « proches » ? Comment quantifier la « distance » entre deux fonctions ?
4) Comment calculer effectivement cette « meilleure approximation » $S_k(y_{\text{init}})$?
5) Comment calcule-t-on les dérivées d'une somme infinie ? (c'est utile pour vérifier que la somme est solution d'une équation aux dérivées partielles !)

1.3 Le programme du cours

Pour apporter des réponses à ces questions, nous allons nous doter de tous les outils mathématiques nécessaires : Après des révisions d'algèbre linéaire (qui fournissent les bases pour l'étude de solutions d'équations linéaires, y compris dans des espaces de fonctions, qui sont en général de dimension infinie), nous allons étudier de façon générale les formes bilinéaires et les formes quadratiques. Ensuite nous approfondirons la notion de produit scalaire, qui en est un cas particulier. Un produit scalaire permet à la fois de définir une distance entre les éléments d'un espace vectoriel, et de préciser la notion de meilleure approximation. Nous n'aurons pas le temps pour faire toute l'analyse nécessaire des séries de fonctions, mais nous nous en ferons une idée en discutant les séries numériques que l'on peut voir comme le cas des fonctions constantes... Enfin, nous étudierons les séries de Fourier qui combinent tous ces outils et permettront de donner des énoncés précis, en particulier en termes d'hypothèses de régularité nécessaires, pour construire les solutions d'équations comme celles qui viennent d'être évoquées.

Chapitre 2

Rappels d'algèbre linéaire

2.1 Rappels sur les espaces vectoriels : définitions et exemples

Un $\mathbb{R}$-espace vectoriel est un ensemble V tel que la somme de deux éléments de V est encore un élément de V, le produit d'un réel (appelé scalaire réel) par un élément de V est encore un élément de V, et qui vérifie les propriétés habituelles des sommes et produits ($x+y=y+x$, existence d'un élément nul, d'un opposé, distributivité du produit par rapport à la somme...). L'exemple typique est l'ensemble des solutions d'un système homogène d'équations linéaires.

Définition 2.1.1. Plus formellement, un espace vectoriel V est un ensemble non vide qui doit être muni d'une loi interne

$$V \times V \to V, (x,y) \mapsto x+y,$$

et d'une loi externe

$$\mathbb{R} \times V \to V, (\lambda,x) \mapsto \lambda \cdot x,$$

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

1) Il existe un élément $0_V \in V$ tel que $0_V + x = x + 0_V = x$ pour tout $x \in V$.
2) $x+(y+z)=(x+y)+z$ pour tout $x,y \in V$
3) $x+y=y+x$ pour tout $x,y \in V$
4) Pour tout $x \in V$, il existe un élément $x' \in V$ tel que $x+x'=x'+x=0_V$. Cet élément x' est alors unique, et est noté $-x$.
5) $1 \cdot x = x$ pour tout $x \in V$
6) $(\lambda\mu)\cdot x = \lambda \cdot (\mu \cdot x)$ pour tout $\lambda,\mu \in \mathbb{R}, x \in V$
7) $\lambda \cdot (x+y) = \lambda \cdot x + \lambda \cdot y$ pour tout $x,y \in V, \lambda \in \mathbb{R}$
8) $(\lambda+\mu)\cdot x = \lambda \cdot x + \mu \cdot x$ pour tout $x \in V, \lambda,\mu \in \mathbb{R}$.

Un $\mathbb{C}$-espace vectoriel est défini de manière analogue en remplaçant $\mathbb{R}$ par $\mathbb{C}$, on peut donc multiplier un élément de V par un complexe (un scalaire complexe).

Remarque 2.1.2. On écrira λx pour $\lambda \cdot x$.

Exemples 2.1.3.

1) $\mathbb{R}^n$, l'espace de vecteurs colonnes $\underline{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ avec $x_i \in \mathbb{R}$, est un espace vectoriel réel. L'espace $\mathbb{C}^n$ de vecteurs colonnes complexes est un espace vectoriel complexe.

2) $\mathbb{R}[X]$, l'espace de polynômes réels en une variable X, est un espace vectoriel réel. De même, $\mathbb{C}[Y]$, l'espace de polynômes complexes en une variable Y est un espace vectoriel complexe.
3) $\mathbb{R}_n[X]$, l'espace de polynômes réels en une variable X de degré $\leqslant n$, est un espace vectoriel réel. De même, $\mathbb{C}_n[Y]$, l'espace de polynômes complexes en une variable Y de degré $\leqslant n$, est un espace vectoriel complexe.
4) $\mathcal{M}_n(\mathbb{R})$, l'espace de matrices $n \times n$ à coefficients réels, est un espace vectoriel réel,
5) Pour tout $a < b \in \mathbb{R}$ l'espace $\mathscr{C}^0([a,b],\mathbb{R})$ de toutes les fonctions continues réelles sur l'intervalle $[a,b]$, est un espace vectoriel réel.
6) Pour tout $a < b \in \mathbb{R}$ et tout entier $i > 0$ l'espace $\mathscr{C}^i([a,b],\mathbb{C})$ de toutes les fonctions i fois continument dérivables à valeurs dans les complexes sur l'intervalle $[a,b]$, est un espace vectoriel complexe.

Vérifier tous ces axiomes est fastidieux. Heureusement dans la pratique, nous travaillerons souvent avec des espaces vectoriels qui sont inclus dans d'autres pour lesquels on a une procédure de vérification simplifiée.

Définition 2.1.4. Soit V un $\mathbb{R}$-espace vectoriel. Un *sous-espace vectoriel* W de V est un sous-ensemble de $W \subset V$ *contenant le vecteur nul de* V, tel que

1) pour tout $w_1, w_2 \in W$ nous avons que $w_1 + w_2 \in W$
2) pour tout $w_1 \in W$ et $\lambda \in \mathbb{R}$ nous avons que $\lambda w_1 \in W$

On montre que l'ensemble W est bien un espace vectoriel avec l'addition et la multiplication héritées de V.

Exercice 2.1.5. Montrer que les sous-ensembles suivants sont tous des sous-espaces vectoriels.

1) L'ensemble de tous les $(x,y) \in \mathbb{C}^2$ tels que $x+y=0$.
2) L'ensemble des solutions d'un système linéaire homogène d'équations.
3) Un plan d'équation $ax+by+cz=0$ ($a,b,c \in \mathbb{R}$ fixés) dans $\mathbb{R}^3$.
4) L'ensemble $\{P \in \mathbb{R}[X] | P(1)=0\}$ des polynômes à coefficients réels qui s'annulent en 1
5) L'ensemble $\{M \in \mathcal{M}_n(\mathbb{C}) | {}^tM = M\}$ des matrices symétriques dans $\mathcal{M}_n(\mathbb{C})$.
6) L'ensemble de toutes les fonctions deux fois dérivables $f \in \mathscr{C}^2(\mathbb{R},\mathbb{R})$ telles que $f'' = -2f$ dans $\mathscr{C}^2(\mathbb{R},\mathbb{R})$.
7) L'ensemble P des fonctions de $\mathbb{R}$ dans $\mathbb{R}$ de période 2π (i.e. $f \in P$ lorsque $f(x+2\pi) = f(x)$ pour tout réel x). Qu'en est-il des fonctions périodiques ?

2.2 Familles libres, génératrices, bases et coordonnées

On vérifie aisément que l'ensemble E des combinaisons linéaires d'une famille de vecteurs $\{v_1,\dots,v_n\}$ d'un espace vectoriel V est un sous-espace vectoriel de V que l'on notera $E = \mathrm{Vect}(v_1,\dots,v_n)$. On dit alors que $\{v_1,\dots,v_n\}$ engendre E.

Définition 2.2.1. On dit que $\{v_1,\dots,v_n\}$ est une *famille génératrice* de V si $V = \mathrm{Vect}(v_1,\dots,v_n)$ (tout élément de E s'écrit comme combinaison linéaire des éléments de la famille).

Si v_n est une combinaison linéaire de $v_1,\dots,v_{n-1}$

$$v_n = \lambda_1 v_1 + \cdots + \lambda_{n-1} v_{n-1}$$

alors $\mathrm{Vect}(v_1,\dots,v_{n-1}) = \mathrm{Vect}(v_1,\dots,v_n)$, on peut donc enlever v_n de la famille génératrice sans changer l'espace vectoriel engendré.

Définition 2.2.2. On dit qu'une famille de vecteurs $\{v_1,\dots,v_n\}$ est *libre* si aucun vecteur n'est combinaison linéaire des autres ou, de manière équivalente, si l'équation $\sum \lambda_i v_i = 0_V$ d'inconnues $\lambda_1,\dots,\lambda_n$ a pour unique solution $\lambda_1 = \cdots = \lambda_n = 0$

Une *base* d'un espace vectoriel E est une famille qui est à la fois génératrice et libre. On peut obtenir une base en enlevant tous les éléments superflus d'une famille génératrice : on commence par enlever les vecteurs nuls, qui n'engendrent rien, puis les vecteurs qui peuvent s'écrire comme combinaison linéaire des autres, jusqu'à obtenir une famille libre.

Une base permet de *représenter* (de manière unique) un élément d'un espace vectoriel par un vecteur colonne.

Définition 2.2.3. Soit V un espace vectoriel réel. Une famille ordonnée d'éléments de V, $\mathbf{e} = \{e_1, \dots, e_n\}$ est une base (finie) pour V si pour tout élément $v \in V$ il existe un *unique* n-uplet de scalaires $(\lambda_1, \lambda_2, \dots, \lambda_n)$ tel que

$$v = \lambda_1 e_1 + \lambda_2 e_2 + \cdots + \lambda_n e_n.$$

L'écriture est unique car sinon la famille $\{e_1, \dots, e_n\}$ ne serait pas libre.

Définition 2.2.4. Avec les notations de la définition 2.2.3, nous dirons que le vecteur colonne

$$\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}$$

est le vecteur des coordonnées de v dans la base $\mathbf{e}$.

Remarque 2.2.5 (**Attention !**)**.** Le vecteur de coordonnées de v dans une base $\mathbf{e}$ dépend autant de la base $\mathbf{e}$ que du vecteur v.

Remarque 2.2.6 (**Notation**)**.** Dans ce qui suit il sera très important de distinguer l'élément v dans un espace vectoriel V de dimension finie n (qui peut être un vecteur colonne, ou une matrice, ou une fonction, ou un polynôme, ou plein d'autres choses) et le vecteur colonne $\underline{V} \in \mathbb{R}^n$ qui le représente dans une base donnée. Pour bien distinguer ces deux objets, nous soulignerons systématiquement les noms des variables qui sont des vecteurs colonnes, et ne soulignerons pas ceux qui ne le sont pas.

Exemples 2.2.7.

1) Les vecteurs

$$\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \cdots, \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

forment une base de $\mathbb{R}^n$, appelée la *base canonique*.

Si $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ est un élément de $\mathbb{R}^n$ alors on peut écrire

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \cdots + x_n \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix};$$

autrement dit, le vecteur de coordonnées de $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ dans la base canonique est $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$. Ceci est une source importante de confusion.

2) Montrons que $B = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix} \right\}$ est une base de $\mathbb{C}^2$. Nous considérons pour un vecteur arbitraire $\begin{pmatrix} x \\ y \end{pmatrix}$ l'équation

$$\begin{pmatrix} x \\ y \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

c'est-à-dire

$$x = \lambda_1 + \lambda_2$$

$$y = \lambda_1 + 2\lambda_2$$

ce qui (après pivot de Gauss) nous donne l'unique solution

$$\lambda_1 = 2x - y,$$

$$\lambda_2 = y - x.$$

Cette famille est donc une base et le vecteur de coordonnées de $\begin{pmatrix} x \\ y \end{pmatrix}$ dans la base B est

$$\begin{pmatrix} 2x-y \\ y-x \end{pmatrix}.$$

3) La famille $B = (1, X, \ldots, X^n)$ forme une base de l'espace vectoriel $\mathbb{R}_n[X]$ des polynômes à coefficients dans $\mathbb{R}$ de degré au plus n. Si $P = a_0 + a_1 X + \cdots a_n X^n$ est un élément de $\mathbb{R}_n[X]$ alors son vecteur de coefficients dans la base B est

$$\begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix}.$$

4) On considère $\mathcal{M}_2(\mathbb{C})$, l'espace de matrices carrées complexes 2×2. Elle a une base

$$B = \left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \right)$$

et dans cette base la matrice $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ a pour vecteur de coefficients $\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}$.

5) On considère l'espace des fonctions réelles deux fois dérivables sur $\mathbb{R}$ qui satisfont l'équation différentielle $f'' = -2f$. Vous avez vu en L1 que cet espace est de dimension 2 et que la famille

$$(\cos(\sqrt{2}x), \sin(\sqrt{2}x))$$

en est une base. Le vecteur de coordonnées de la fonction $f = a\cos(\sqrt{2}x) + b\sin(\sqrt{2}x)$ dans cette base est $\begin{pmatrix} a \\ b \end{pmatrix}$.

Définition 2.2.8. Lorsqu'un espace vectoriel V possède une base finie on dit que V est de dimension finie.

Soit n le nombre d'éléments de cette base B de V. Alors une famille libre de V a au plus n éléments. En effet, considérons une famille de $n+1$ éléments $\{v_1, \ldots, v_{n+1}\}$. On pose le système

$$\lambda_1 v_1 + \cdots + \lambda_{n+1} v_{n+1} = 0$$

en écrivant les coordonnées des vecteurs dans la base B. Ce système a plus d'inconnues $(n+1)$ que d'équations (n) donc il admet une solution non identiquement nulle, ce qui prouve que la famille n'est pas libre. (En faisant le pivot de Gauss on peut écrire le système sous forme échelonnée. Si on trouve un pivot dans les colonnes de 1 à n, on peut exprimer λ_n en fonction de λ_{n+1} avec la dernière équation, puis λ_{n-1} en fonction de λ_{n+1}, etc. On trouve ainsi une solution non identiquement nulle. S'il y a une colonne sans pivot, par exemple la troisième, alors on prend $\lambda_4 = \cdots = \lambda_{n+1} = 0$, la deuxième équation donne λ_2 en fonction de λ_3 et la première équation λ_1 en fonction de λ_2.)

On en déduit que :

Proposition 2.2.9. *Si V est de dimension finie, toutes les bases de V ont le même nombre d'éléments : ce nombre s'appelle la dimension de V.*

Exemples 2.2.10.

1) L'espace $\mathbb{R}^n$ est de dimension n.
2) L'espace $\mathbb{R}_n[X]$ est de dimension $n+1$.
3) L'espace $\mathcal{M}_2(\mathbb{R})$ est de dimension 4.
4) L'espace $\mathbb{R}[X]$ n'est pas de dimension finie (sinon on aurait une base, on regarde le plus grand degré des éléments de la base, un polynôme de degré plus grand ne peut pas être combinaison linéaire des éléments de la base).
5) On peut aussi montrer que l'espace des fonctions 2π-périodiques n'est pas de dimension finie. Un des objectif des séries de Fourier, c'est en quelque sorte d'en donner une « base » mais ayant un nombre infini d'éléments.

Le résultat suivant sera souvent utilisé pour vérifier qu'une famille de vecteurs est une base.

Lemme 2.2.11. *Soit V un espace vectoriel de dimension n et soit $\{e_1,\dots,e_n\}$ une famille de n vecteurs dans V. Si la famille $\{e_1,\dots,e_n\}$ est libre alors c'est une base.*

En effet, si $v \in V$, alors la famille $\{e_1,\dots,e_n,v\}$ n'est pas libre puisqu'elle a $n+1$ éléments, donc on a une combinaison linéaire non identiquement nulle

$$\lambda_1 e_1 + \cdots + \lambda_n e_n + \lambda v = 0$$

On a $\lambda \neq 0$ car $\{e_1,\dots,e_n\}$ est libre, donc v est combinaison linéaire de $\{e_1,\dots,e_n\}$.

Proposition 2.2.12. *Tout sous-espace W d'un espace V de dimension finie n est de dimension finie $m \leqslant n$ (avec égalité si et seulement si $W = V$).*

En effet, une famille libre de W est une famille libre de V donc a au plus n éléments. On crée ensuite une famille libre de W ayant un nombre maximal d'éléments, c'est une base de W.

Les coordonnées d'un élément $v \in V$ dans une base seront essentielles dans la suite, car elles nous permettront de ramener tous nos calculs à de simples multiplications de matrices. Il nous sera, d'ailleurs, souvent utile de simplifier nos calculs au maximum en choisissant une base bien adaptée. Pour faire cela, il nous faut comprendre comment le vecteur $\underline{V}$ des coordonnées d'un élément $v \in V$ dans une base $\mathbf{e}$ se transforme lorsqu'on change de base.

Définition 2.2.13. Soit V un espace vectoriel de dimension n et soient $\mathbf{E} = \{e_1,\dots,e_n\}$ et $\mathbf{F} = \{f_1,\dots,f_n\}$ des bases de V. On appelle matrice de passage de $\mathbf{E}$ vers $\mathbf{F}$ la matrice obtenue en écrivant en colonnes les coordonnées des f_i dans la base $\mathbf{E}$:

$$P = (\underline{V}_1,\dots,\underline{V}_n)$$

où $\underline{V}_i$ est le vecteur de coordonnés de f_i dans la base $\mathbf{E} = \{e_1,\dots,e_d\}$.

Remarque 2.2.14. Cas particulier
Si $\mathbf{E}$ est la base canonique de $\mathbb{R}^n$, la matrice de passage P est donnée par

$$P = (\underline{f}_1,\dots,\underline{f}_n).$$

C'est-à-dire que la première colonne de P est formée par les composantes de f_1, la deuxième colonne de P par les composantes de f_2, etc.

Soit $\{e_1,\dots,e_n\}$ une base de V. Soit $\{f_1,\dots,f_n\}$ une autre base de V, et $v \in V$ tel que

$$v_1 f_1 + \cdots + v_n f_n = v$$

Cette équation devient un système si on remplace par les coordonnées des f_i et de v dans la base $\{e_1,\dots,e_n\}$. Ce système a pour inconnues les coordonnées de v dans la base $\{f_1,\dots,f_n\}$, il a comme matrice P la matrice de passage de $\{e_1,\dots,e_n\}$ vers $\{f_1,\dots,f_n\}$ et comme second membre les coordonnées de v dans la base $\{e_1,\dots,e_n\}$. D'où le

Théorème 2.2.15. *Soient* $\mathbf{B_1}$ *et* $\mathbf{B_2}$ *des bases de V et soit v un élément de V. Soient* $\underline{V}_1$ *et* $\underline{V}_2$ *les vecteurs de coordonnés de v dans les bases* $\mathbf{B}_1$ *et* $\mathbf{B}_2$*. Soit P la matrice de passage de B_1 vers B_2. Alors*

$$\underline{V}_1 = P\underline{V}_2$$

ou, de façon équivalente

$$\underline{V}_2 = P^{-1}\underline{V}_1$$

Remarque 2.2.16. Attention il faut multiplier par P^{-1} (et pas P) le vecteur colonne des composantes de v dans la base $\mathbf{B_1}$ pour obtenir le vecteur colonnes des composantes de v dans la base $\mathbf{B_2}$.

Il y a une généralisation de la notion de base qui sera utile dans la démonstration d'un théorème ultérieur.

Définition 2.2.17. Soient $V_1,\dots,V_m$ des sous-espaces vectoriels de V. On dit que V est la *somme directe* des sous-espaces $V_1,\dots,V_m$, et on écrit $V = V_1 \oplus V_2 \oplus \dots \oplus V_m$, si et seulement si pour tout $v \in V$ il existe m uniques éléments $v_1 \in V_1,\dots,v_m \in V_m$ tels que

$$v = v_1 + \cdots + v_m.$$

On montre aussi la

Proposition 2.2.18. *Si $V = V_1 \oplus V_2 \oplus \dots \oplus V_m$ et pour chaque i nous avons que* $\mathbf{E}_i$ *est une base de V_i alors la concaténation* $(\mathbf{E}_1,\mathbf{E}_2,\dots,\mathbf{E}_m)$ *est une base de V.*

2.3 Applications linéaires

Considérons maintenant la classe des applications qui préservent la structure d'espace vectoriel.

Définition 2.3.1. Soient V et V' deux $\mathbb{R}$-espaces vectoriels.

Une *application linéaire* de V dans V' est une application $\varphi : V \to V'$ qui préserve l'addition et la multiplication par un réel :

1) $\varphi(v_1+v_2) = \varphi(v_1)+\varphi(v_2)$ pour tous $v_1, v_2 \in V$ (l'image de la somme est la somme des images)
2) $\varphi(\lambda v) = \lambda\varphi(v)$ pour tous $\lambda \in \mathbb{R}, v \in V$ (l'image du produit par λ est le produit par λ de l'image)

Dans le cas où l'espace d'arrivée est $\mathbb{R}$ on dira que φ est une *forme linéaire*

Remarque 2.3.2. Pour toute application linéaire φ on a nécessairement $\varphi(0) = 0$.

Remarque 2.3.3. On peut ramener les deux conditions à une seule vérification :
$\varphi(\lambda v_1 + v_2) = \lambda\varphi(v_1)+\varphi(v_2)$ pour tous $\lambda \in \mathbb{R}$, $v_1, v_2 \in V$.

Pour définir une application linéaire entre deux espaces vectoriels sur $\mathbb{C}$, on remplace ci-dessus $\mathbb{R}$ par $\mathbb{C}$.

Exemples 2.3.4.

1) L'application $\mathbb{R}^3 \to \mathbb{R}^2$ donnée par $\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \end{pmatrix}$ est linéaire. Elle l'est aussi de $\mathbb{C}^3 \to \mathbb{C}^2$.

2) L'application $\mathbb{C}^3 \to \mathbb{C}^2$ donnée par $\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{pmatrix} x \\ y+1 \end{pmatrix}$ n'est pas linéaire.

3) L'application de $\mathbb{C} \to \mathbb{C}$ définie par $\varphi(z) = \bar{z}$ n'est pas linéaire. Mais si on considère $\mathbb{C}$ comme un $\mathbb{R}$-espace vectoriel (de dimension 2) elle le devient.

4) L'application des fonctions continument dérivables dans les fonctions continues ($\mathscr{C}^1(\mathbb{R},\mathbb{R}) \mapsto \mathscr{C}^0(\mathbb{R},\mathbb{R})$), définie par $f \mapsto f' - 2f$ est linéaire.

5) L'application de transposition dans l'espace vectoriel des matrices carrées $\mathscr{M}_n(\mathbb{C}) \mapsto \mathscr{M}_n(\mathbb{C})$ donnée par $M \mapsto {}^tM$ est linéaire.

6) L'application de l'espace des polynômes de degré inférieur ou égal à 3 dans l'espace des polynômes de degré inférieur ou égal à 1 $\mathbb{R}_3[X] \mapsto \mathbb{R}_1[X]$, $P \mapsto P''$, est une application linéaire.

Exercice 2.3.5. Démontrer que les applications 1, 3, 4, 5 sont bien linéaires et que 2 ne l'est pas.

Définition 2.3.6. Le *noyau* de φ, noté $\mathrm{Ker}(\varphi)$, est l'ensemble

$$\mathrm{Ker}(\varphi) = \{v \in V \mid \varphi(v) = 0\}.$$

C'est un sous-espace vectoriel de V.

Définition 2.3.7. *L'image* de φ, notée $\mathrm{Im}(\varphi)$, est l'ensemble

$$\mathrm{Im}(\varphi) = \{\varphi(v), v \in V\}.$$

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

Exercice 2.3.8.

1) Montrer que le noyau et l'image d'une application linéaire sont des sous-espaces vectoriels.

2) Calculer l'image et le noyau des applications linéaires données en exemple.

Définition 2.3.9. On appelle *rang* d'une application linéaire φ la dimension de son image $\mathrm{Im}(\varphi)$.

On rappelle le théorème du rang, qui est très utile.

Théorème 2.3.10. *Soit $\varphi : V \to W$ une application linéaire. On suppose que V est de dimension finie. Alors* $\mathrm{Im}(\varphi)$ *est de dimension finie et*

$$\dim(V) = \dim(\mathrm{Ker}(\varphi)) + \dim(\mathrm{Im}(\varphi)).$$

Preuve. On prend une base $\{v_1, \ldots, v_n\}$ de V, les images $\{\varphi(v_1), \ldots, \varphi(v_n)\}$ forment une partie génératrice de $\mathrm{Im}\,\varphi$ qui est donc de dimension finie. De cette partie, on peut extraire une base. Pour simplifier, on pose r le rang de φ ($0 \leqslant r \leqslant n$), et on renumérote les v_i de telle sorte que $\{\varphi(v_1), \ldots, \varphi(v_r)\}$ soit une base de $\mathrm{Im}\,\varphi$. Alors, chaque colonne supplémentaire de la matrice de φ est liée avec les r précédentes : pour $r < k \leqslant n$, $\varphi(v_k) = \sum_{i=1}^{r} \lambda_i^k \varphi(v_i)$. Par linéarité de φ on a donc $n-r$ vecteurs du noyau : $v'_k = v_k - \sum_{i=1}^{r} \lambda_i^k v_i$. Il est facile de voir que le système $\{v_1, \ldots, v_r, v'_{r+1}, \ldots, v'_n\}$ forme une base de V telle que $\{v'_{r+1}, \ldots, v'_n\}$ est une base de $\mathrm{Ker}\,\varphi$.

2.4 Calcul Matriciel

Dans cette section nous ferons des rappels sur les matrices et leurs manipulations. Celles-ci seront un élément clé de notre travail ce semestre.

Définition 2.4.1. Étant donnés deux entiers m et n strictement positifs, une *matrice à m lignes et n colonnes* est un tableau rectangulaire de réels $A = (a_{i,j})$. L'indice de ligne i va de 1 à m, l'indice de colonne j va de 1 à n.

$$A = (a_{i,j}) = \begin{pmatrix} a_{1,1} & \cdots & a_{1,j} & \cdots & a_{1,n} \\ \vdots & & \vdots & & \vdots \\ a_{i,1} & \cdots & a_{i,j} & \cdots & a_{i,n} \\ \vdots & & \vdots & & \vdots \\ a_{m,1} & \cdots & a_{m,j} & \cdots & a_{m,n} \end{pmatrix}.$$

Les entiers m et n sont les *dimensions* de la matrice, $a_{i,j}$ est son *coefficient d'ordre* (i, j).

Notons qu'une matrice A peut être définie en donnant une expression pour ses coefficients $a_{i,j}$ Par exemple, la matrice A de taille 2×2 donnée par le formule $a_{i,j} = i + j$ est la matrice

$$A = \begin{pmatrix} 1+1 & 1+2 \\ 2+1 & 2+2 \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 3 & 4 \end{pmatrix}.$$

L'ensemble des matrices à m lignes et n colonnes et à coefficients réels est noté $\mathcal{M}_{m,n}(\mathbb{R})$. Ce qui suit s'applique aussi, si on remplace $\mathbb{R}$ par $\mathbb{C}$, à l'ensemble des matrices à coefficients complexes.

Notons trois cas spéciaux :

1) Un vecteur de n éléments peut s'écrire comme une vecteur colonne $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ (matrice $n \times 1$).
2) Un vecteur de n éléments peut s'écrire comme un vecteur ligne $\begin{pmatrix} x_1, & x_2, & \ldots, & x_n \end{pmatrix}$ (matrice $1 \times n$).
3) Un nombre réel x peut être vu comme une matrice 1×1.

Notation. Si $\underline{X}$ est un vecteur colonne à n éléments, on notera le coefficient $\underline{X}_{1,i}$ par $\underline{X}_i$.

L'ensemble $\mathcal{M}_{m,n}(\mathbb{R})$ est naturellement muni d'une addition (on peut ajouter deux matrices de mêmes dimensions terme à terme) et de multiplication par des scalaires (on peut multiplier une matrice par un réel terme à terme).

- *Addition :* Si $A = (a_{i,j})$ et $B = (b_{i,j})$ sont deux matrices de $\mathcal{M}_{m,n}(\mathbb{R})$, leur somme $A + B$ est la matrice $(a_{i,j} + b_{i,j})$. Par exemple :

$$\begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 1 & -1 \end{pmatrix} + \begin{pmatrix} -3 & 1 \\ 5 & -3 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} -2 & 2 \\ 7 & 0 \\ 1 & 1 \end{pmatrix}$$

- *Multiplication par un scalaire :* Si $A = (a_{i,j})$ est une matrice de $\mathcal{M}_{m,n}(\mathbb{R})$, et λ est un réel, le produit λA est la matrice $(\lambda a_{i,j})$. Par exemple :

$$-2 \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} -2 & -2 \\ -4 & -6 \\ -2 & 2 \end{pmatrix}$$

Observons que ces opérations auraient le même effet si les matrices étaient disposées comme des mn-uplets de réels (toutes les lignes étant concaténées, par exemple).

Définition 2.4.2 (Matrice d'une application linéaire)**.** Soit φ une application linéaire d'un espace vectoriel V_1 de base $B_1 = (e_1, \ldots, e_n)$ dans un espace vectoriel V_2 de base $B_2 = (f_1, \ldots, f_n)$. On appelle matrice de φ dans les bases B_1 et B_2 la matrice dont les colonnes sont les composantes dans la base B_2 des images $\varphi(e_1), \ldots, \varphi(e_n)$ des vecteurs $e_1, \ldots, e_n$ de la base B_1.

Si $V_1 = V_2$ on choisit (presque toujours) $B_1 = B_2$.

Exemple 2.4.3. Soit l'application linéaire de $\mathbb{R}^3$ dans $\mathbb{R}^2$ qui à un vecteur $X=(x,y,z)$ associe le vecteur $Y=(x+2y-z,3x-2z)$. Sa matrice dans les bases canoniques de $\mathbb{R}^3$ et $\mathbb{R}^2$ a pour première colonne les composantes de $\varphi(e_1)=\varphi((1,0,0))=(1,3)$, pour deuxième colonne les composantes de $\varphi(e_2)=\varphi((0,1,0))=(2,0)$ et pour troisième colonne les composantes de $\varphi(e_3)=\varphi((0,0,1))=(-1,-2)$ donc

$$\begin{array}{ccc} \varphi(e_1) & \varphi(e_2) & \varphi(e_3) & \\ 1 & 2 & -1 & f_1 \\ 3 & 0 & -2 & f_2 \end{array} \quad \Rightarrow \quad M=\begin{pmatrix} 1 & 2 & -1 \\ 3 & 0 & -2 \end{pmatrix}.$$

On observe qu'on a en ligne les coefficients en x,y,z des coordonnées du vecteur image.

Calcul du noyau et de l'image de φ. Soit $\varphi : V \mapsto V'$ une application linéaire qui a pour matrice relativement à des bases B et B' de V et V' la matrice M ci-dessus. Pour calculer le noyau de φ, il faut résoudre le système linéaire

$$\begin{cases} x+2y-z &= 0 \\ 3x-2z &= =0 \end{cases}$$

dont la matrice est M. On réduit donc M (en lignes) par l'algorithme du pivot de Gauss pour se ramener à une matrice triangulaire. Dans l'exemple ci-dessus, on remplace la ligne L_2 par L_2-3L_1 ce qui donne la matrice

$$M=\begin{pmatrix} 1 & 2 & -1 \\ 0 & -6 & 1 \end{pmatrix}$$

La deuxième équation donne $-6y+z=0$ soit $y=z/6$. Ensuite la première équation donne $x+2y-z=0$ soit $x=-2y+z=2z/3$. Donc $(x,y,z)=z(2/3,1/6,1)$ et $\mathrm{Ker}(\varphi)$ est de dimension 1, engendré par le vecteur $(2/3,1/6,1)$. Le théorème du rang donne alors que $\mathrm{Im}(\varphi)$ est de dimension $3-1=2$, c'est donc $\mathbb{R}^2$ tout entier.

Dans le cas général, les vecteurs colonnes de M forment une famille génératrice de $\mathrm{Im}(\varphi)$. Il suffit de réduire M en colonnes par l'algorithme du pivot de Gauss, une fois la réduction terminée les colonnes non nulles forment une base de $\mathrm{Im}(\varphi)$. Comme nous l'avons vu pour la démonstration du théorème du rang, les relations sur les colonnes fournissent des vecteurs du noyau.

Proposition 2.4.4. *Soit* φ *une application linéaire de* V_1 *muni de la base* $B_1=\{e_1,\ldots,e_n\}$ *vers* V_2 *muni de la base* $B_2=\{f_1,\ldots,f_m\}$ *et* M *la matrice de* φ *dans les bases* B_1 *et* B_2. *Soit* $v\in V_1$ *un vecteur de composantes* $\underline{X}$ *dans la base* B_1.

Alors les composantes de $\varphi(v)$ *dans la base* B_2 *sont données par le vecteur* $M\underline{X}$ *de composantes :*

$$(M\underline{X})_i := \sum_{j=1}^{n} M_{i,j}\underline{X}_j.$$

En effet :

$$\varphi(v)=\varphi\left(\sum_{j=1}^{n} X_j e_j\right)=\sum_{j=1}^{n} X_j\varphi(e_j)=\sum_{j=1}^{n} X_j\sum_{i=1}^{m} M_{i,j}f_i=\sum_{i=1}^{m}\left(\sum_{j=1}^{n} M_{i,j}X_j\right)f_i$$

Soit φ une application linéaire de V_1 de base B_1 dans V_2 de base B_2 et ψ une autre application linéaire de V_2 dans V_3 de base B_3. On peut montrer que la composée $\psi(\varphi(.))$ est une application linéaire de V_1 dans V_3. Que se passe-t-il pour les matrices représentant ψ, φ et la matrice de la composée ? On vérifie que la matrice de la composée s'obtient en faisant le *produit matriciel* des matrices de ψ et φ (cela peut même être une façon de définir le produit de matrices).

Définition 2.4.5. Soient m,n,p trois entiers strictement positifs. Soit $A=(a_{i,j})$ une matrice de $\mathcal{M}_{m,n}(\mathbb{R})$ et soit $B=(b_{j,k})$ une matrice de $\mathcal{M}_{n,p}(\mathbb{R})$. On appelle *produit matriciel* de A par B la matrice $C\in\mathcal{M}_{m,p}(\mathbb{R})$ dont le terme général $c_{i,k}$ est défini, pour tout $i=1,\ldots,m$ et pour tout $k\in 1,\ldots,p$ par :

$$c_{i,k}=\sum_{j=1}^{n} a_{i,j}\,b_{j,k}\ .$$

Nous insistons sur le fait que le produit AB de deux matrices n'est défini que si le nombre de colonnes de A et le nombre de lignes de B sont les mêmes (pour la composition des applications linéaires, ceci correspond au fait que l'espace vectoriel de départ de la deuxième application ψ est le même que l'espace vectoriel d'arrivée de la première application φ, ils ont donc même dimension). Dans le cas particulier où B est un vecteur colonne de taille $n \times 1$ cette opération nous fournit un vecteur colonne de taille $m \times 1$.

$$\begin{pmatrix} b_{1,1} & \cdots & b_{1,k} & \cdots & b_{1,n} \\ \vdots & & \vdots & & \vdots \\ & \cdots & b_{j,k} & \cdots & \\ \vdots & & \vdots & & \vdots \\ b_{n,1} & \cdots & b_{n,k} & \cdots & b_{n,p} \end{pmatrix}$$

$$\begin{pmatrix} a_{1,1} & \cdots & & \cdots & a_{1,n} \\ \vdots & & \vdots & & \vdots \\ a_{i,1} & \cdots & a_{i,j} & \cdots & a_{i,n} \\ \vdots & & \vdots & & \vdots \\ a_{m,1} & \cdots & & \cdots & a_{m,n} \end{pmatrix} \quad \begin{pmatrix} c_{1,1} & & \vdots & & c_{1,p} \\ & & \vdots & & \\ \cdots & \cdots & c_{i,k} & & \\ & & & & \\ c_{m,1} & & & & c_{m,p} \end{pmatrix}$$

Posons par exemple :

$$A = \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 1 & -1 \end{pmatrix} \quad \text{et} \quad B = \begin{pmatrix} 0 & 1 & -1 & -2 \\ -3 & -2 & 0 & 1 \end{pmatrix}.$$

La matrice A a 3 lignes et 2 colonnes, la matrice B a 2 lignes et 4 colonnes. Le produit AB a donc un sens : c'est une matrice à 3 lignes et 4 colonnes.

$$\begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 0 & 1 & -1 & -2 \\ -3 & -2 & 0 & 1 \end{pmatrix} = \begin{pmatrix} -3 & -1 & -1 & -1 \\ -9 & -4 & -2 & -1 \\ 3 & 3 & -1 & -3 \end{pmatrix}$$

Le produit matriciel a les propriétés habituelles d'un produit, à une exception notable près : il n'est pas commutatif

Proposition 2.4.6. *Le produit matriciel possède les propriétés suivantes.*

1) Associativité : *Si les produits AB et BC sont définis, alors les produits $A(BC)$ et $(AB)C$ le sont aussi et ils sont égaux.*

$$A(BC) = (AB)C\,.$$

2) Linéarité à droite : *Si B et C sont deux matrices de mêmes dimensions, si λ et μ sont deux réels et si A a autant de colonnes que B et C ont de lignes, alors*

$$A(\lambda B + \mu C) = \lambda AB + \mu AC\,.$$

3) Linéarité à gauche : *Si A et B sont deux matrices de mêmes dimensions, si λ et μ sont deux réels et si C a autant de lignes que A et B ont de colonnes, alors*

$$(\lambda A + \mu B)C = \lambda AC + \mu BC\,.$$

Ces propriétés se démontrent par le calcul à partir de la définition 2.4.5 ou en interprétant le produit comme une composition d'applications linéaires.

La *transposition* est une opération qui va intervenir plus loin dans le calcul matriciel avec les formes bilinéaires (d'un point de vue théorique cela provient de la dualité, qui dépasse le cadre de ce cours).

Définition 2.4.7. Étant donnée une matrice $A = (a_{i,j})$ de $\mathcal{M}_{m,n}(\mathbb{R})$, sa *transposée* est la matrice de $\mathcal{M}_{n,m}(\mathbb{R})$ dont le coefficient d'ordre (j,i) est $a_{i,j}$.

Pour écrire la transposée d'une matrice, il suffit de transformer ses lignes en colonnes. Par exemple :

$$A = \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 1 & -1 \end{pmatrix} \quad , \quad {}^tA = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 3 & -1 \end{pmatrix} .$$

Observons que la transposée de la transposée est la matrice initiale.

$${}^t({}^tA) = A .$$

La transposée d'un produit est le produit des transposées, mais il faut inverser l'ordre des facteurs.

Proposition 2.4.8. *Soient m,n,p trois entiers strictement positifs. Soient $A = (a_{i,j})$ une matrice de $\mathcal{M}_{m,n}(\mathbb{R})$ et $B = (b_{j,k})$ une matrice de $\mathcal{M}_{n,p}(\mathbb{R})$. La transposée du produit de A par B est le produit de la transposée de B par la transposée de A.*

$${}^t(AB) = {}^tB\,{}^tA .$$

Par exemple, en reprenant les matrices A et B définies ci-dessus :

$$\begin{pmatrix} 0 & -3 \\ 1 & -2 \\ -1 & 0 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 1 & 3 & -1 \end{pmatrix} = \begin{pmatrix} -3 & -9 & 3 \\ -1 & -4 & 3 \\ -1 & -2 & -1 \\ -1 & -1 & -3 \end{pmatrix}$$

Définition 2.4.9. Soit n un entier strictement positif et A une matrice carrée à n lignes et n colonnes. On dit que A est symétrique si pour tous $i,j = 1,\ldots,n$, ses coefficients d'ordre $a_{i,j}$ et $a_{j,i}$ sont égaux, ce qui est équivalent à dire que A est égale à sa transposée.

Le produit d'une matrice par sa transposée est toujours une matrice symétrique. En effet :

$${}^t(A\,{}^tA) = {}^t({}^tA)\,{}^tA = A\,{}^tA .$$

2.5 Matrices carrées

En général si le produit AB est défini, le produit BA n'a aucune raison de l'être. Le produit d'une matrice par sa transposée est une exception, les matrices carrées en sont une autre : si A et B sont deux matrices à n lignes et n colonnes, les produits AB et BA sont tous deux définis et ils ont les mêmes dimensions que A et B. En général ils ne sont pas égaux. Par exemple,

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

Nous noterons simplement $\mathcal{M}_n(\mathbb{R})$ l'ensemble $\mathcal{M}_{n,n}(\mathbb{R})$ des matrices carrées à n lignes et n colonnes, à coefficients réels. Parmi elles la *matrice identité*, notée I_n, joue un rôle particulier.

$$I_n = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

En effet, elle est l'élément neutre du produit matriciel : pour toute matrice $A \in \mathcal{M}_{m,n}(\mathbb{R})$,

$$A\,I_m = I_n\,A = A .$$

On le vérifie facilement à partir de la définition 2.4.5.

Définition 2.5.1. Soit A une matrice de $\mathcal{M}_n$. On dit que A est inversible s'il existe une matrice A' dans $\mathcal{M}_n$ telle que

$$AA' = A'A = I_n \, .$$

On appelle A' la *matrice inverse* de A, et on la note A^{-1}.

Par exemple :

$$\begin{pmatrix} 1 & 0 & -1 \\ 1 & -1 & 0 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 & 1 \\ 1 & -2 & 1 \\ 0 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 \\ 1 & -2 & 1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1 \\ 1 & -1 & 0 \\ 1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Observons que l'inverse, s'il existe, est nécessairement unique. En effet, soient B_1 et B_2 deux matrices telles que $AB_1 = B_1A = I_n$ et $AB_2 = B_2A = I_n$. En utilisant l'associativité, le produit B_1AB_2 vaut $B_1(AB_2) = B_1I_n = B_1$, mais aussi $(B_1A)B_2 = I_nB_2 = B_2$. Donc $B_1 = B_2$.

Nous rappelons la proposition suivante, qui nous dit qu'il suffit de trouver une matrice B telle que $AB = I_n$ pour être sûr que A est inversible et que son inverse est B.

Proposition 2.5.2. *Soit A une matrice de $\mathcal{M}_n$. Supposons qu'il existe une matrice B telle que $AB = I_n$* ou bien *$BA = I_n$. Alors A est inversible et $B = A^{-1}$.*

Si A et B sont deux matrices inversibles de $\mathcal{M}_n$, leur produit est inversible.

Proposition 2.5.3. *Soient A et B deux matrices inversibles de $\mathcal{M}_n(\mathbb{R})$. Le produit AB est inversible et son inverse est $B^{-1}A^{-1}$.*

Preuve. Nous utilisons le théorème 2.5.2, ainsi que l'associativité du produit :

$$(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}I_nB = B^{-1}B = I_n \, .$$

L'inverse d'une matrice et la proposition 3.3.6 permettent de donner une formule de changement de base pour une application linéaire.

Proposition 2.5.4. *Soit φ une application linéaire d'un espace vectoriel V_1 de base B_1 vers un espace vectoriel V_2 de base B_2, de matrice M relativement à ces bases B_1 et B_2. Soit B_1' une autre base de V_1 de matrice de passage P_1 dans la base B_1, et B_2' une autre base de V_2 de matrice de passage P_2 dans la base B_2. Alors la matrice M' de φ relativement aux bases B_1' et B_2' est donnée par*

$$M' = P_2^{-1}MP_1$$

Si $V_1 = V_2$ on prend $B_1 = B_2$ et $B_1' = B_2'$ donc $P_1 = P_2 = P$ et on a

$$M' = P^{-1}MP.$$

Exemple 2.5.5. Dans $\mathbb{R}^2$ vu comme le plan complexe, on considère l'application linéaire $f : z \to \bar{z}$. On vérifie qu'il s'agit bien d'une application linéaire (c'est une symétrie par rapport à l'axe Ox).

Dans la base canonique B, sa matrice est

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

Prenons la base B' dont les vecteurs ont pour affixe $1+i$ et $1-i$, la matrice de passage de B à B' est

$$P = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

donc la matrice de f dans B' est

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

ce qu'on vérifie directement puisque les deux vecteurs de base sont conjugués l'un de l'autre.

Exemple 2.5.6. Dans $\mathbb{R}^2$, on considère la projection orthogonale sur la droite vectorielle engendrée par le vecteur $v(1,1)$. On prend pour $B_1 = B_2$ la base canonique (e_1, e_2) et pour $B'_1 = B'_2$ la base formée par v et un vecteur orthogonal $w(1,-1)$.

L'image de v est lui-même i.e. $1v+0w$, donc la première colonne de M' est $(1,0)$. L'image de w est le vecteur nul, donc

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

L'image du vecteur $(1,0)$ par la projection est $\frac{1}{2}v = (\frac{1}{2}, \frac{1}{2})$; de même pour $(0,1)$ donc les 2 colonnes de M ont pour coordonnées $(\frac{1}{2}, \frac{1}{2})$

$$M = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

La matrice de passage de B'_1 est (coordonnées de v et w en colonnes)

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

On calcule

$$P^{-1} = -\frac{1}{2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}.$$

Vérifions que $M' = P^{-1}MP$:

$$-\frac{1}{2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix} \cdot \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = -\frac{1}{4}\begin{pmatrix} -2 & -2 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Définition 2.5.7. On définit le *rang* d'une matrice M comme étant la dimension du sous-espace vectoriel engendré par ses vecteurs colonnes. Il s'agit donc du rang de toute application linéaire ayant M comme matrice.

Proposition 2.5.8. *Multiplier une matrice à droite ou à gauche par une matrice inversible ne change pas son rang.*

Cela résulte du fait que le produit de matrices correspond à la composition de deux applications linéaires et que composer avec une application linéaire inversible ne change pas le rang. En effet

— pour la composition à droite, si φ est inversible, alors $\mathrm{Im}(\psi \circ \varphi)$=$\mathrm{Im}(\psi)$,

— pour la composition à gauche, les images par φ d'une base de $\mathrm{Im}(\psi)$ forment une base de $\mathrm{Im}(\varphi \circ \psi)$.

Enfin, nous aurons parfois besoin du lemme suivant :

Lemme 2.5.9. *Soit $M \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée $n \times n$. Si pour tout $\underline{X}, \underline{Y} \in \mathbb{R}^n$ nous avons que ${}^t\underline{X}M\underline{Y} = 0$ alors $M = 0$.*

Preuve. Soit pour tout i le vecteur colonne $\underline{e}_i \in \mathbb{R}^n$ défini par

$$(\underline{e}_i)_j = 1 \text{ si } i = j,\ 0 \text{ si } i \neq j.$$

Alors pour tout $1 \leqslant i, j \leqslant n$ on a que

$${}^t\underline{e}_i M \underline{e}_j = M_{i,j} = 0$$

et donc $M = 0$.

Réécrivons maintenant notre problème initial dans le langage des espaces vectoriels. Nous considérons une fonction réelle continue f, définie sur une intervalle $[0,L]$ ($f \in V = \mathscr{C}^0([0,L],\mathbb{R})$). Nous voulons chercher une fonction g_n qui est de la forme

$$g_n(x) = a_0 + \sum_{k=1}^{n} a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \sin\left(\frac{k\pi x}{L}\right)$$

et qui doit être « aussi proche que possible » de f.

Dans le langage des espaces vectoriels on pourrait écrire la chose suivante : Soit W le sous-espace de tous les éléments $g \in V$ qui peuvent s'écrire sous la forme

$$g_n(x) = a_0 + \sum_{k=1}^{n} a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \sin\left(\frac{k\pi x}{L}\right).$$

W est alors un sous-espace vectoriel de V (exercice : le démontrer !) : de plus, W est de dimension finie et admet pour base finie la famille

$$\mathbf{e} = (1, \cos\frac{\pi x}{L}, \sin\frac{\pi x}{L}, \ldots, \cos\frac{n\pi x}{L}, \sin\frac{n\pi x}{L}).$$

Nous cherchons à identifier un élément $g \in W$ qui est « le plus proche que possible » de $f \in V$.

Notre problème initial est donc un exemple particulier du problème suivant :

Question. J'ai un espace vectoriel V et un élément $v \in V$. Il y a dans V un sous-espace spécial de dimension finie $W \subset V$. Je veux approcher au mieux v par un élément $w \in W$. Comment faire ? Et tout d'abord, qu'est-ce que ça veut dire « approcher au mieux » ?

Dans les deux prochains chapitres, nous aborderons surtout la question : qu'est-ce que ça veut dire « approcher au mieux » ?

Chapitre 3

Formes bilinéaires

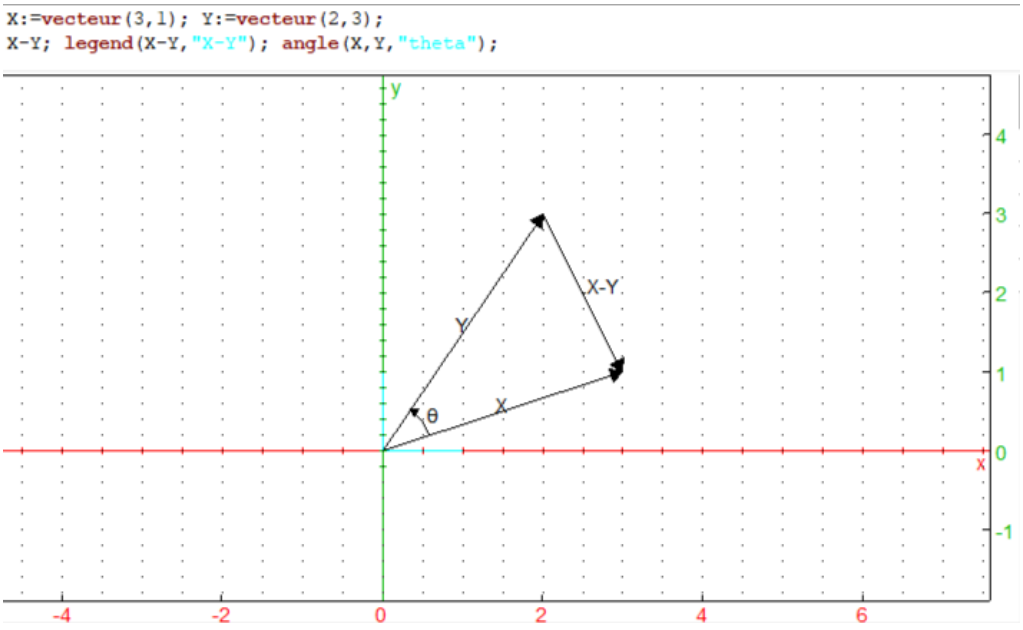
3.1 Le produit scalaire canonique sur $\mathbb{R}^3$

Dans le chapitre précédent, nous avons étudié la notion d'espace vectoriel. Cette notion est utile parce qu'elle englobe à la fois des espaces géométriques tels que $\mathbb{R}^2$ et $\mathbb{R}^3$ et des espaces de fonctions tels que $\mathbb{R}_n[X]$ et $\mathscr{C}^0([0,1],\mathbb{R})$. Notre but est maintenant d'utiliser cette notion pour étendre des idées géométriques (distance et angle, par exemple) à des espaces de fonctions. Pour faire cela, il nous sera nécessaire d'identifier une formule purement algébrique qui permet de calculer distances et angles dans $\mathbb{R}^3$, faisant intervenir le produit scalaire canonique sur $\mathbb{R}^3$.

Définition 3.1.1. Le produit scalaire canonique sur $\mathbb{R}^3$ est une fonction prenant en argument deux vecteurs $\underline{X} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ et $\underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$ définie par

$$\langle \underline{X} | \underline{Y} \rangle = x_1 y_1 + x_2 y_2 + x_3 y_3.$$

Le produit scalaire canonique tire son intérêt du fait qu'il encode la géométrie de l'espace $\mathbb{R}^3$.



Théorème 3.1.2. *Soient $\underline{X}$ et $\underline{Y}$ deux vecteurs dans $\mathbb{R}^3$, soit d la longueur de la différence $\underline{X}-\underline{Y}$ et soit θ l'angle entre ces deux vecteurs. On a :*

$$d = \sqrt{\langle \underline{X}-\underline{Y} | \underline{X}-\underline{Y} \rangle}, \quad \theta = \arccos\left(\frac{\langle \underline{X}|\underline{Y}\rangle}{\sqrt{\langle \underline{X}|\underline{X}\rangle \langle \underline{Y}|\underline{Y}\rangle}} \right).$$

Il existe donc une formule qui permet de calculer la distance et l'angle entre deux vecteurs en utilisant seulement le produit scalaire. Nous allons donc essayer de définir des classes de fonctions sur des espaces vectoriels qui ressemblent au produit scalaire sur $\mathbb{R}^3$ dans l'espoir qu'elles nous livreront une bonne notion de « distance ».

Une des propriétés clés du produit scalaire est qu'il se comporte effectivement comme un produit sous les opérations algébriques de base sur les vecteurs, c'est-à-dire qu'on a, pour tout $\underline{X},\underline{Y},\underline{Z} \in \mathbb{R}^3$ et pour tout $\lambda \in \mathbb{R}$

1) $\langle \underline{X}+\underline{Y}, \underline{Z}\rangle = \langle \underline{X}|\underline{Z}\rangle + \langle \underline{Y}|\underline{Z}\rangle$
2) $\langle \underline{X}|\underline{Y}+\underline{Z}\rangle = \langle \underline{X}|\underline{Y}\rangle + \langle \underline{X}|\underline{Z}\rangle$
3) $\langle \underline{X}|\lambda\underline{Y}\rangle = \langle \lambda\underline{X}|\underline{Y}\rangle = \lambda\langle \underline{X}|\underline{Y}\rangle$

Nous allons donc commencer par étudier les fonctions de deux vecteurs qui respectent ces conditions.

3.2 Formes bilinéaires : définitions et exemples

Dans cette section, de nouveau, nous présenterons la théorie des formes bilinéaires réelles, mais tous nos résultats seront valables pour des formes complexes.

Définition 3.2.1. Soient V un $\mathbb{R}$-espace vectoriel, et soit φ une fonction qui envoie deux éléments de V sur un réel $\varphi : V \times V \to \mathbb{R}$. On dit que φ est *une forme bilinéaire* si elle vérfie :

1) pour tout $v_1, v_2 \in V$ et $v \in V$ nous avons que $\varphi(v_1+v_2, v) = \varphi(v_1, v) + \varphi(v_2, v)$
2) pour tout $v \in V$, $v' \in V$ et $\lambda \in \mathbb{R}$ nous avons que $\varphi(\lambda v, v') = \lambda\varphi(v, v')$.
3) pour tout $v \in V$ et $v_1, v_2 \in V$ nous avons que $\varphi(v, v_1+v_2) = \varphi(v, v_1) + \varphi(v, v_2)$
4) pour tout $v \in V$, $v' \in V$ et $\lambda \in \mathbb{R}$ nous avons que $\varphi(v, \lambda v') = \lambda\varphi(v, v')$.

On dit que φ est *linéaire à gauche* si elle vérifie les conditions 1-2.
On dit que φ est *linéaire à droite* si elle vérifie les conditions 3-4.
On dit que φ est *symétrique* si $\varphi(v,u) = \varphi(u,v)$ pour tout $u, v \in V$.
On dit que φ est *antisymétrique* si $\varphi(v,u) = -\varphi(u,v)$ pour tout $u, v \in V$.

Remarque 3.2.2. On utilise le terme *forme* parce que la valeur de φ est un réel. Le terme *bilinéaire* vient du fait que si on fixe un des arguments, on a une application linéaire par rapport à l'autre argument (linéaire *à gauche* et *à xsdroite*).

Exemples 3.2.3.

1) L'application
$$\varphi : \begin{cases} \mathbb{R}\times\mathbb{R} & \to & \mathbb{R} \\ (x,y) & \mapsto & xy \end{cases}$$
est une forme bilinéaire symétrique.

2) Le produit scalaire
$$\varphi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \quad \left(\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right) \mapsto x \cdot y = \sum_{i=1}^{n} x_i y_i$$
est une forme bilinéaire symétrique. Lorsque $n = 2$ ou 3, on retrouve le produit scalaire étudié ci-dessus. Nous appelons cette forme le *produit scalaire canonique* sur $\mathbb{R}^n$.

3) L'application qui à deux polynômes P et Q associe le produit $P(0)Q(1)$

$$\varphi : \left\{ \begin{array}{ccc} \mathbb{C}[X]\times\mathbb{C}[X] & \to & \mathbb{C} \\ (P,Q) & \mapsto & P(0)Q(1) \end{array} \right.$$

est une forme bilinéaire. Elle n'est ni symétrique ni antisymétrique.

4) L'application qui à deux matrices carrées M et N associe la trace du produit des deux matrices

$$\varphi : \left\{ \begin{array}{ccc} mathcalM_n(\mathbb{R})\times & & \\ mathcalM_n(\mathbb{R}) & \to & \mathbb{R} \\ (M,N) & \mapsto & \mathrm{tr}(MN) \end{array} \right.$$

est une forme bilinéaire symétrique.

5) En dimension 2, l'application déterminant

$$\varphi : \mathbb{R}^2\times\mathbb{R}^2 \to \mathbb{R}, \quad \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right) \mapsto x_1y_2 - x_2y_1$$

est bilinéaire et antisymétrique.
Attention, ça n'est plus vrai en dimension supérieure !

6) L'application

$$\varphi : \mathbb{C}^2\times\mathbb{C}^2 \to \mathbb{C}, \quad \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right) \mapsto x_1x_2 + 2x_1y_2$$

n'est pas bilinéaire.
En effet, posons $\underline{U} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \underline{V} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$. On a

$$\varphi(5\underline{U},\underline{V}) = (5\times 1)\times(5\times 2) + 2\times(5\times 1)\times 4 = 90 \neq 5\times\varphi(\underline{U},\underline{V}) = 50.$$

7) L'application qui associe à deux fonctions continues f et g l'intégrale de leur produit sur $[0,1]$

$$\varphi : \left\{ \begin{array}{ccc} \mathscr{C}^0([0,1],\mathbb{R})\times\mathscr{C}^0([0,1],\mathbb{R}) & \to & \mathbb{R} \\ (f,g) & \to & \int_0^1 f(x)g(x)dx \end{array} \right.$$

est une forme bilinéaire symétrique.

8) Pour toute fonction continue $p : [0,1] \to \mathbb{R}$, l'application

$$\varphi : \left\{ \begin{array}{ccc} \mathscr{C}^0([0,1],\mathbb{R})\times\mathscr{C}^0([0,1],\mathbb{R}) & \to & \mathbb{R} \\ (f,g) & \to & \int_0^1 p(x)f(x)g(x)dx \end{array} \right.$$

est une forme bilinéaire symétrique.

Un cas particulier intéressant est celui où on applique une forme bilinéaire à deux vecteurs identiques.

Définition 3.2.4. Soit V un espace vectoriel sur $\mathbb{R}$ et soit φ une forme bilinéaire symétrique sur V. Alors la forme quadratique associée à φ, notée q_φ, est la fonction définie sur V par

$$q_\varphi(v) = \varphi(v,v)$$

La forme quadratique associée à une forme bilinéaire est un analogue de la fonction carré d'un nombre réel, ou de la norme de v au carré ($\|v\|^2$) quand v est un vecteur dans $\mathbb{R}^2$ ou $\mathbb{R}^3$. Les formules suivantes (dites « formule de polarisation » et « formule du parallélogramme ») permettent de retrouver une forme bilinéaire symétrique à partir de la forme quadratique associée.

Lemme 3.2.5. *Soit V un espace vectoriel, φ une forme bilinéaire symétrique sur $V \times V$ et q_φ la forme quadratique associée. Alors pour tout $v, w \in V$ on a*

$$\begin{aligned}\varphi(v,w) &= \frac{1}{2}(q_\varphi(v+w) - q_\varphi(v) - q_\varphi(w)) \\ &= \frac{1}{4}(q_\varphi(v+w) - q_\varphi(v-w))\end{aligned}$$

$$q_\varphi(v+w) + q_\varphi(v-w) = 2(q_\varphi(v) + q_\varphi(w)).$$

La démonstration de ce lemme est laissée en exercice.

Remarque 3.2.6. Ces formules sont les généralisations des relations suivantes sur $\mathbb{R}$:

$$\begin{aligned}xy &= \frac{1}{2}((x+y)^2 - x^2 - y^2) \\ &= \frac{1}{4}\big((x+y)^2 - (x-y)^2\big),\end{aligned}$$

$$(x+y)^2 + (x-y)^2 = 2(x^2+y^2).$$

3.3 Formes bilinéaires : représentation matricielle

Nous allons maintenant définir la matrice d'une forme bilinéaire dans une base, qui va nous permettre, modulo le choix d'une base, de réduire les calculs faisant intervenir des formes bilinéaires sur des espaces de dimension finie à des multiplications de matrices.

Définition 3.3.1. Soit V un $\mathbb{R}$-espace vectoriel de dimension finie n, soit $\mathbf{e} = (e_1, \ldots, e_n)$ une base de V, et soit $\varphi : V \times V \to \mathbb{R}$ une forme bilinéaire. La matrice de φ dans la base $\mathbf{e}$ est la matrice M $n \times n$ dont les coefficients sont donnés par

$$M_{i,j} = (\varphi(e_i, e_j))_{1 \leqslant i,j \leqslant n}.$$

Lemme 3.3.2. *Soit V un espace vectoriel de dimension finie n, soient $x, y \in V$, soit $\mathbf{e} = (e_1, \ldots, e_n)$ une base de V, notons $\underline{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ et $\underline{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ les vecteurs coordonnées de x et y dans la base $\mathbf{e}$ (autrement dit $x = \sum_{i=1}^n x_i e_i, y = \sum_{i=1}^n y_i e_i$). Soit $\varphi : V \times V \to \mathbb{R}$ une forme bilinéaire, et soit M la matrice de φ dans la base $\mathbf{e}$. Alors on a*

$$\varphi(x,y) = {}^t\underline{X} M \underline{Y} = \sum_{i,j} \varphi(e_i, e_j) x_i y_j.$$

Preuve. On a

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

puisque φ est linéaire en y. Or on a aussi

$$\varphi(\sum_{i=1}^n x_i e_i, e_j) = \sum_{i=1}^n \varphi(x_i e_i, e_j) = \sum_{i=1}^n x_i \varphi(e_i, e_j).$$

Ainsi, on obtient

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

On a aussi

$$M\underline{Y} = \begin{pmatrix} \vdots \\ \sum_{j=1}^{n} \varphi(e_i, e_j) y_j \\ \vdots \end{pmatrix},$$

et donc

$${}^t\underline{X} M \underline{Y} = \begin{pmatrix} \cdots & x_i & \cdots \end{pmatrix} \begin{pmatrix} \vdots \\ \sum_{j=1}^{n} \varphi(e_i, e_j) y_j \\ \vdots \end{pmatrix} = \sum_{i,j} x_i \varphi(e_i, e_j) y_j = \sum_{i,j} \varphi(e_i, e_j) x_i y_j.$$

Corollaire 3.3.3. *Soit V un espace vectoriel de dimension finie n. Soit $\varphi : V \times V \to \mathbb{R}$ une forme bilinéaire. Les propositions suivantes sont équivalentes.*

1) *φ est symétrique*
2) *Pour tout base $\mathbf{e}$ de V, la matrice M de φ dans la base $\mathbf{e}$ est symétrique.*
3) *Il existe une base $\mathbf{e}$ de V telle que la matrice M de φ dans la base $\mathbf{e}$ est symétrique.*

Preuve. Soit $\varphi : V \times V \to \mathbb{R}$ une forme bilinéaire, et soit $\mathbf{e}$ une base de V.

Si φ est symétrique, alors on a

$$\varphi(e_i, e_j) = \varphi(e_j, e_i) \text{ pour tout } i, j,$$

et ceci s'écrit matriciellement ${}^tM = M$, par définition de la matrice de φ. On a donc $(1) \Rightarrow (2)$. L'implication $(2) \Rightarrow (3)$ étant claire, il reste à montrer $(3) \Rightarrow (1)$.

Supposons qu'il existe une base $\mathbf{e}$ de V telle que M est symétrique. Soient $x, y \in V$, et soient $\underline{X}, \underline{Y}$ leurs vecteurs de coordonnées dans la base $\mathbf{e}$. On a alors que

$$\varphi(x, y) = {}^t\underline{X} M \underline{Y}$$

Le membre de droite est une matrice 1×1 : elle est donc égale à sa propre transposée et on a

$$\varphi(x, y) = {}^t\underline{X} M \underline{Y} = {}^t({}^t\underline{X} M \underline{Y}) = {}^t\underline{Y}\, {}^tM \underline{X} = {}^t\underline{Y} M \underline{X} = \varphi(y, x).$$

Le lemme précédent admet une réciproque, bien utile pour démontrer qu'une application est bilinéaire et donner sa matrice représentative dans une base fixée.

Lemme 3.3.4. *Soit V un $\mathbb{R}$-espace vectoriel de dimension finie, et soit $\mathbf{e} = (e_1, \ldots, e_n)$ une base de V. Pour tout $a_{ij} \in \mathbb{R}, 1 \leqslant i, j \leqslant n$, l'application*

$$\varphi : \begin{cases} V \times V & \to & \mathbb{R} \\ (\sum_{i=1}^{n} x_i e_i, \sum_{j=1}^{n} y_j e_j) & \mapsto & \sum_{1 \leqslant i,j \leqslant n} a_{ij} x_i y_j \end{cases}$$

est une forme bilinéaire, dont la matrice A dans la base $\mathbf{e}$ est donnée par $A_{ij} = (a_{ij})$.

Exemples 3.3.5.

1) L'application

$$\varphi : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}, \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right) \mapsto x_1 y_1 + x_2 y_2 + 3 x_1 y_2 - x_2 y_1$$

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

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

2) Considérons l'application qui à deux polynômes de degrés inférieurs ou égaux à 2 associe le produit de leur valeur en 1 et 0

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

On peut vérifier directement que φ est bilinéaire, mais on peut aussi utiliser la remarque précédente. Pour cela, considérons la base $1, X, X^2$ de $\mathbb{R}_2[X]$. On écrit

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

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

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

Regardons maintenant ce qui se passe lorsque l'on effectue un changement de base.

Proposition 3.3.6. *Soit V un $\mathbb{R}$-espace vectoriel de dimension finie n, soient $\mathbf{e}$ et $\mathbf{e}'$ deux bases de V, et soit P la matrice de passage de la base $\mathbf{e}$ à la base $\mathbf{e}'$ (c'est-à-dire colonne par colonne la matrice des coordonnées des vecteurs de $\mathbf{e}'$ dans la base $\mathbf{e}$). Soit $\varphi : V \times V \to \mathbb{R}$ une forme bilinéaire, soit M sa matrice dans la base $\mathbf{e}$ et soit N sa matrice dans la base $\mathbf{e}'$. Alors on a $N = {}^tPMP$.*

Preuve. Soient $x, y \in V$, soient $\underline{X}, \underline{Y}$ leur vecteurs de coordonnées dans la base $\mathbf{e}$ et soient $\underline{X}', \underline{Y}'$ leurs coordonnées dans la base $\mathbf{e}'$. On a alors $\underline{X} = P\underline{X}'$ et $\underline{Y} = P\underline{Y}'$ pour tout x, y et donc

$$\varphi(x,y) = {}^t\underline{X}M\underline{Y} = {}^t(P\underline{X}')MP\underline{Y}' = {}^t\underline{X}'{}^tPMP\underline{Y}' = {}^t\underline{X}'N\underline{Y}'.$$

c'est-à-dire que $N = {}^tPMP$ par 2.5.9.

Nous sommes prêts à définir la notion de rang.

Définition 3.3.7. Soit $\varphi : V \times V \to \mathbb{R}$ une forme bilinéaire. Le *rang* de φ est le rang de n'importe quelle matrice représentative de φ dans une base de V.

Le rang est bien défini et ne dépend pas de la base choisie d'après la proposition précédente et la proposition 2.5.8.

3.4 Orthogonalité

Les expressions permettant de calculer $\varphi(x,y)$ peuvent se simplifier grandement lorsque la base $\mathbf{e}$ est adaptée. Par exemple, il est souvent utile de se débarrasser des termes croisés lorsque c'est possible. On introduit pour cela la notion d'orthogonalité.

Définition 3.4.1. Soit V un espace vectoriel de dimension n sur $\mathbb{R}$, et soit $\varphi : V \times V \to \mathbb{R}$ une forme bilinéaire *symétrique*.

On dit que deux vecteurs $x, y \in V$ sont φ-*orthogonaux* si $\varphi(x,y) = 0$.

On le note $x \underset{\varphi}{\perp} y$, ou $x \perp y$ s'il n'y a pas de confusion possible.

On dit que la base $\mathbf{e} = (e_1, \ldots, e_n)$ est φ-*orthogonale* si les vecteurs de la base sont φ-orthogonaux deux à deux, c'est-à-dire si on a

$$\varphi(e_i, e_j) = 0 \text{ pour tout } i \neq j.$$

Lemme 3.4.2. *La base $\mathbf{e}$ est φ-orthogonale si et seulement si M, la matrice de φ dans la base $\mathbf{e}$, est diagonale.*

Preuve. La base $\mathbf{e}$ est φ-orthogonale $\Leftrightarrow \varphi(e_i, e_j) = 0$ si $i \neq j \Leftrightarrow M_{i,j} = 0$ si $i \neq j \Leftrightarrow M$ est diagonale.

Définition 3.4.3. On dit que **e** est φ-*orthonormée* si on a

$$\varphi(e_i, e_j) = \begin{cases} 0 & \text{si } i \neq j, \\ 1 & \text{si } i = j. \end{cases}$$

Lemme 3.4.4. *La base* **e** *est* φ*-orthonormée si et seulement si* $\mathrm{Mat}(\varphi, \mathbf{e})$ *est la matrice identité.*

Preuve. Laissée en exercice.

Définition 3.4.5. On dit que deux sous-espaces W, W' de V sont *orthogonaux* si on a

$$\varphi(w, w') = 0 \text{ pour tout } w \in W, w' \in W'.$$

On dit que V est la *somme directe orthogonale* des sous-espaces $V_1, \ldots, V_m$ si $V = V_1 \oplus \ldots \oplus V_m$ et les sous-espaces $V_1, \ldots, V_m$ sont orthogonaux deux à deux. On note alors

$$V = V_1 \overset{\perp}{\oplus} \ldots \overset{\perp}{\oplus} V_m.$$

On a le

Lemme 3.4.6. *Soit V un espace vectoriel et soit* φ *une forme bilinéaire sur V. Soient $V_1, \ldots V_k$ des sous-espaces de V tels que $V = V_1 \overset{\perp}{\oplus} \ldots \overset{\perp}{\oplus} V_k$. Si pour chaque i, $\mathbf{v_i}$ est une base orthonormée de V_i alors la concaténation $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k)$ est une base orthonormée de V.*

En effet tout vecteur w de cette base de V est dans une des $\mathbf{v}_i$ donc $\varphi(w, w) = 1$, et il est orthogonal à tout autre vecteur w' de cette base de V, soit parce que $\mathbf{v}_i$ est orthonormée si $w' \in \mathbf{v}_i$, soit parce que tous les vecteurs de V_j sont orthogonaux à ceux de V_i quand $i \neq j$.

Exemples 3.4.7.

1) L'application qui a une paire de polynômes de degré au plus 2 associe

$$\varphi(P, Q) = \int_{-1}^{1} P(t)Q(t)\,\mathrm{d}t$$

est bilinéaire symétrique. De plus, $1 \underset{\varphi}{\perp} X$ et $X \underset{\varphi}{\perp} X^2$. Par contre, 1 et X^2 ne sont pas φ-orthogonaux, puisque l'on a $\varphi(1, X^2) = \frac{2}{3}$. La base $1, X, X^2$ n'est donc pas φ-orthogonale.
On peut vérifier que la base

$$1, X, X^2 - \frac{1}{3}$$

est φ-orthogonale. Elle n'est pas φ-orthonormée puisque

$$\varphi(1,1) = 2,\ \varphi(X, X) = 2/3,\ \varphi(X^2 - \frac{1}{3}, X^2 - \frac{1}{3}) = 8/45.$$

On peut la rendre φ-orthonormée en multipliant chaque élément de la base par une constante bien choisie. Plus précisément, la base :

$$\frac{1}{\sqrt{2}}, \sqrt{\frac{3}{2}}X, \sqrt{\frac{45}{8}}(X^2 - \frac{1}{3})$$

est une base φ-orthonormée.

2) La base canonique de $\mathbb{R}^n$ est φ-orthonormée pour la forme bilinéaire symétrique

$$\varphi(x, y) = x \cdot y = \sum_{i=1}^{n} x_i y_i$$

3) Soit $V = \mathscr{C}^0([-1,1],\mathbb{R})$, et soient $\mathcal{P}$ et I le sous-espace des fonctions paires et impaires respectivement. On sait que l'on a
$$V = \mathcal{P} \oplus I.$$
Considérons sur $V \times V$ l'application
$$\varphi(f,g) = \int_{-1}^{1} f(t)g(t)\,\mathrm{d}t$$
Alors, on a
$$\varphi(f,g) = 0 \text{ pour tout } f \in \mathcal{P}, g \in I.$$
On a donc
$$V = \mathcal{P} \overset{\perp}{\oplus} I.$$

4) Soit φ la forme bilinéaire symétrique sur $\mathbb{R}^3$ de matrice
$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix}$$
Alors $(1,0,1)$ est orthogonal à tout vecteur, $(1,0,0)$ est orthogonal à lui-même. La base $\{(1,0,1),(1,1,0),(1,-1,0)\}$ est φ-orthogonale. Il n'existe pas de base φ-orthonormée.

Le lemme 3.3.2 entraîne immédiatement :

Lemme 3.4.8. *Soit V un espace vectoriel de dimension finie n, soit $\mathbf{e} = (e_1,\dots,e_n)$ une base de V, et soient*
$$x = \sum_{i=1}^{n} x_i e_i, y = \sum_{i=1}^{n} y_i e_i$$
deux vecteurs de V. Soit $\varphi : V \times V \to \mathbb{R}$ une forme bilinéaire symétrique. Si $\mathbf{e}$ est φ-orthogonale, on a
$$\varphi(x,y) = \sum_{i=1}^{n} \varphi(e_i,e_i)x_i y_i.$$
En particulier, si $\mathbf{e}$ est φ-orthonormée, on a
$$\varphi(x,y) = \sum_{i=1}^{n} x_i y_i.$$

Il n'existe pas toujours une base φ-orthonormée. En effet, si $\varphi : V \times V \to \mathbb{R}$ est bilinéaire symétrique et s'il existe une base φ-orthonormée alors le lemme précédent montre que $\varphi(x,x) > 0$ pour tout $x \neq 0$.

Par exemple, la forme bilinéaire symétrique sur $\mathbb{R}^2 \times \mathbb{R}^2$ définie par
$$\varphi((x_1,x_2),(y_1,y_2)) = x_1y_1 - x_2y_2.$$
n'admet pas de base φ-orthonormée, puisque $\varphi((0,1),(0,1)) = -1 < 0$.

En revanche, on a le théorème suivant :

Théorème 3.4.9. *Soit V un espace vectoriel de dimension finie sur $\mathbb{R}$, et soit $\varphi : V \times V \to \mathbb{R}$ une forme bilinéaire symétrique. Alors il existe une base de V qui est φ-orthogonale.*

Preuve. On démontre l'existence d'une base φ-orthogonale par récurrence sur $n = \dim(V)$.

Idée de la preuve : prenons un vecteur e_0, et regardons l'ensemble des vecteurs φ-orthogonaux à e_0, c'est un sous-espace de dimension n ou $n-1$. Si la dimension vaut n, e_0 est orthogonal à tout le monde, on peut prendre un sous-espace de dimension $n-1$ qui ne contient pas e_0, une base φ-orthogonale de ce sous-espace auquel on ajoute e_0 convient. Si la dimension vaut $n-1$, on prend une base φ-orthogonale de ce sous-espace, si e_0 n'appartient pas au sous-espace, on ajoute e_0 à la base. On a donc intérêt à choisir

e_0 tel que $\varphi(e_0,e_0)\neq 0$ (dans l'exemple sur $\mathbb{R}^3$, on ne peut pas par exemple prendre $e_0=(1,0,0)$ qui est orthogonal à lui-même).

Soit donc (P_n) la propriété : « Pour tout $\mathbb{R}$-espace vectoriel de dimension n et tout $\varphi : V\times V\to\mathbb{R}$, il existe une base φ-orthogonale. »

Si $n=1$, il n'y a rien à démontrer.

Supposons que (P_n) soit vraie, et soit $\varphi : V\times V\to\mathbb{R}$ une forme bilinéaire symétrique avec $\dim(V)=n+1$.

Si $\varphi=0$, toute base est φ-orthogonale, et on a fini. On suppose donc que $\varphi\neq 0$. Soit q la forme quadratique associée. Par la formule de polarisation, si $q=0$ alors $\varphi=0$, ce qui n'est pas le cas. Il existe donc un e_0 tel que $q(e_0)\neq 0$, c'est-à-dire, $\varphi(e_0,e_0)\neq 0$.

L'application

$$f:\left\{\begin{array}{ccc} V & \to & \mathbb{R}\\ y & \mapsto & \varphi(e_0,y)\end{array}\right.$$

est alors une application linéaire non nulle, puisque

$$f(e_0)=\varphi(e_0,e_0)\neq 0$$

et son image est donc $=\mathbb{R}$. Par le théorème du rang,

$$\dim\mathrm{Ker}(f)=n+1-1=n.$$

Par hypothèse de récurrence, il existe une base $(e_1,\dots,e_n)$ de $\mathrm{Ker}(f)$ qui est orthogonale pour la forme

$$\varphi':\left\{\begin{array}{ccc} \mathrm{Ker}(f)\times\mathrm{Ker}(f) & \to & \mathbb{R}\\ (x,y) & \mapsto & \varphi(x,y)\end{array}\right.$$

Montrons que $\mathbf{e}=(e_0,e_1,\dots,e_n)$ est une base de V. Puisque $\dim(V)=n+1$, il suffit de montrer que la famille $(e_0,\dots,e_n)$ est libre. Soient $\lambda_0,\dots,\lambda_n\in\mathbb{R}$ tels que

$$\lambda_0e_0+\lambda_1e_1+\cdots+\lambda_ne_n=0.$$

En appliquant f à cette égalité et en utilisant la linéarité, on obtient

$$\lambda_0f(e_0)+\lambda_1f(e_1)+\cdots+\lambda_nf(e_n)=0.$$

Puisque $e_1,\dots,e_n\in\mathrm{Ker}(f)$, on obtient $\lambda_0f(e_0)=0$. Comme $f(e_0)\neq 0$, on obtient $\lambda_0=0$. On a donc

$$\lambda_1e_1+\cdots+\lambda_ne_n=0.$$

Comme $(e_1,\dots,e_n)$ est une base de $\mathrm{Ker}(f)$, c'est une famille libre, et on obtient donc

$$\lambda_1=\cdots=\lambda_n=0.$$

Ceci prouve que $\mathbf{e}$ est une base de V. Il reste à vérifier que cette base est φ-orthogonale.

Par choix des e_i, on a

$$\varphi(e_i,e_j)=\varphi'(e_i,e_j)=0 \text{ pour tout } i\neq j, 1\leqslant i,j\leqslant n$$

et aussi

$$\varphi(e_0,e_j)=f(e_j)=0 \text{ pour tout } j>0$$

parce que $e_j\in\mathrm{Ker}(f)$. On a donc que

$$\varphi(e_i,e_j)=0 \text{ pour tout } 0\leqslant i\neq j\leqslant n.$$

Ainsi, $(e_0,e_1,\dots,e_n)$ est une base φ-orthogonale. Ceci achève la récurrence.

Remarque 3.4.10. Le résultat précédent peut être faux si φ n'est pas bilinéaire symétrique. Par exemple, si $\varphi : V \times V \to \mathbb{R}$ est antisymétrique, c'est-à-dire si on a

$$\varphi(y,x) = -\varphi(x,y) \text{ pour tout } x,y \in V,$$

et si φ est *non nulle*, alors il n'existe pas de base φ-orthogonale de V.

En effet, si φ est une telle forme, alors on a

$$\varphi(x,x) = -\varphi(x,x) \text{ pour tout } x \in V.$$

On a donc

$$\varphi(x,x) = 0 \text{ pour tout } x \in V.$$

Supposons maintenant que $\mathbf{e} = (e_1,\dots,e_n)$ est une base φ-orthogonale. On a donc

$$\varphi(e_i,e_i) = 0 \text{ pour tout } i = 1,\dots,n.$$

Comme $\varphi(e_i,e_j) = 0$ pour tout $i \neq j$ puisque $\mathbf{e}$ est φ-orthogonale, on en déduit que si M est la matrice de φ dans $\mathbf{e}$ alors $M = 0$.

Le Lemme 3.3.2 entraîne alors que l'on a

$$\varphi(x,y) = 0 \text{ pour tout } x,y \in V,$$

ce qui contredit le fait que φ est non nulle.

Un exemple d'une telle forme bilinéaire φ est donné par le déterminant de deux vecteurs de $\mathbb{R}^2$,

$$\det\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) = x_1y_2 - x_2y_1.$$

Proposition 3.4.11. *Soit E un sous-ensemble d'un espace vectoriel V, et φ une forme bilinéaire symétrique sur V. L'ensemble W des vecteurs φ-orthogonaux à tous les éléments de E est un sous-espace vectoriel de V, on le note $E^\perp$. On a $E^\perp = \mathrm{Vect}(E)^\perp$ et si F est une famille génératrice de $\mathrm{Vect}(E)$ alors $E^\perp = F^\perp$.*

Preuve. Utiliser la linéarité de φ par rapport à un de ses arguments.

Pour chercher l'orthogonal d'un ensemble E (en dimension finie), il suffit donc de trouver une base $\{e_1,\dots,e_n\}$ de $\mathrm{Vect}(E)$ et de résoudre le système linéaire $\varphi(v,e_j) = 0, j = 1,\dots,n$.

Définition 3.4.12. Soit V un espace vectoriel et φ une forme bilinéaire symétrique sur V. On appelle noyau de φ l'orthogonal de l'espace V tout entier :

$$\mathrm{Ker}(\varphi) = V^\perp.$$

En dimension finie, si on a une base B de V, et si M est la matrice de φ, le noyau de φ est le noyau de l'endomorphisme de matrice M

$$\mathrm{Ker}(\varphi) = \mathrm{Ker}(M)$$

En effet, si v et w ont pour coordonnées les vecteurs colonnes X et Y, on a $\varphi(v,w) = {}^tXMY$, donc si w est dans le noyau de l'endomorphisme de matrice M, alors $MY = 0$ et $\varphi(v,w) = 0$. Réciproquement, on prend $X = MY$ et on remarque que ${}^tXX = \|X\|^2$.

Exemple 3.4.13. Calculer les noyaux des formes des exemples ci-dessus.

Si B est une base φ-orthogonale, on voit que le noyau de φ a pour base l'ensemble des vecteurs e_j de B tels que $\varphi(e_j,e_j) = 0$, la dimension du noyau de φ est le nombre de coefficients nuls sur la diagonale de M (qui est diagonale). Ce nombre ne change donc pas si on prend une autre base φ-orthogonale.

Définition 3.4.14. Soit V un espace vectoriel de dimension finie et φ une forme bilinéaire symétrique sur V. On définit le rang de φ par

$$\mathrm{rg}(\varphi) = \dim(V) - \dim(\mathrm{Ker}(\varphi))$$

SI B est une base de V, c'est aussi le rang de la matrice M de φ dans la base B.

Le calcul du rang se fait donc comme si M était une matrice d'application linéaire. Si B est une base φ-orthogonale, le rang de M est le nombre de coefficients non nuls sur la diagonale de M. Ce nombre ne change donc pas si on prend une autre base φ-orthogonale.

En fait on a un résultat un peu plus général, qui dit que le nombre de coefficients strictement positifs et le nombre de coefficients strictement négatifs ne dépend pas de la base φ-orthogonale, c'est le théorème de Sylvester (et la définition de la signature) que nous verrons plus bas.

3.5 Calcul effectif d'une base φ-orthogonale

3.5.1 Lien avec la forme quadratique correspondante

Nous allons calculer une base φ-orthogonale en exploitant la forme quadratique q qui lui est associée. Rappelons que la forme bilinéaire symétrique φ peut être reconstruite à partir de la forme quadratique q via la formule de polarisation

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

Nous disons alors que φ est la forme polaire de q, que nous noterons parfois φ_q.

Exemples 3.5.1.

1) La norme euclidienne de $\mathbb{R}^n$ définie pour x de coordonnées $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ par

$$q(x) = x_1^2 + \cdots + x_n^2$$

est une forme quadratique, de forme polaire le produit scalaire usuel

$$\varphi_q\left(\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}\right) = x_1y_1 + \cdots + x_ny_n.$$

En effet, l'application φ est bilinéaire symétrique et on a clairement $\varphi(x,x) = q(x)$.
Vérifions la formule de polarisation :

$$q(x+y) = \sum_{i=1}^{n}(x_i+y_i)^2 = \sum_{i=1}^{n}(x_i^2 + 2x_iy_i + y_i^2) = q(x) + q(y) + 2\varphi(x,y).$$

2) L'application qui à une fonction continue sur $[0,1]$ à valeurs réelles associe

$$q(f) = \int_0^1 f(t)^2 \,\mathrm{d}t$$

est une forme quadratique, de forme polaire

$$\varphi_q(f,g) = \int_0^1 f(t)g(t) \,\mathrm{d}t.$$

Vérifions la formule de polarisation.

$$\begin{aligned} q(f+g) &= \int_0^1 (f(t)+g(t))^2 \,\mathrm{d}t \\ &= \int_0^1 \left(f(t)^2 + 2f(t)g(t) + g(t)^2\right) \mathrm{d}t \\ &= q(f) + q(g) + 2\int_0^1 f(t)g(t) \,\mathrm{d}t. \end{aligned}$$

Définition 3.5.2. Soit V un $\mathbb{R}$-espace vectoriel de dimension finie n, et soit $q : V \to \mathbb{R}$ une forme quadratique. Soit $\mathbf{e}$ une base de V. La matrice M de q dans la base $\mathbf{e}$ est la matrice de la forme polaire φ_q dans la base $\mathbf{e}$. C'est une matrice symétrique par le corollaire 3.3.3.

Le *rang* de q, noté $\mathrm{rg}(q)$, est le rang de sa forme polaire.

On dit que $\mathbf{e}$ est q-orthogonale (resp. q-orthonormée) si elle est φ_q-orthogonale (resp. φ_q-orthonormée).

L'égalité $q(x) = \varphi_q(x,x)$ et le lemme 3.3.2 donnent immédiatement :

Lemme 3.5.3. *Soit V un espace vectoriel de dimension finie n et $\mathbf{e}$ une base pour V. Soit $x \in V$, et soit $\underline{X}$ le vecteur coordonnées de x dans la base $\mathbf{e}$.*

Soit $q : V \to \mathbb{R}$ une forme quadratique, et soit M sa matrice dans la base $\mathbf{e}$. Alors on a

$$q(x) = {}^t\underline{X} M \underline{X}.$$

En particulier, si $\mathbf{e}$ est q-orthogonale, c'est-à-dire si M est diagonale, alors on a

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

Le lemme suivant nous permet de passer directement de la forme quadratique q a sa matrice M sans calculer le forme polaire φ.

Lemme 3.5.4. *Soit V un espace vectoriel de dimension finie n. Soient $x, y \in V$, et soit $\mathbf{e} = (e_1, \ldots, e_n)$ une base de V. Alors pour tout $a_{ij} \in \mathbb{R}, 1 \leqslant i \leqslant j \leqslant n$, l'application définie sur V par*

$$q(\sum_{i=1}^{n} x_i e_i) = \sum_{i=1}^{n} a_{ii} x_i^2 + 2 \sum_{1 \leqslant i < j \leqslant n} a_{ij} x_i x_j$$

est une forme quadratique, et sa matrice A dans la base $\mathbf{e}$ est donnée par

$$A = (a_{ij}).$$

La démonstration est laissée en exercice au lecteur. *Attention au facteur* 2 !

Exemple 3.5.5. L'application définie sur $\mathbb{R}^2$ par

$$q(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}) = 3x_1^2 + 4x_1x_2 + 5x_2^2$$

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

$$\begin{pmatrix} 3 & 2 \\ 2 & 5 \end{pmatrix}.$$

Soient maintenant φ une forme bilinéaire sur un espace V, q sa forme polaire, $\mathbf{e}$ une base pour V. Soit $x \in V$ un élément arbitraire et $\underline{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ son vecteur de coordonnées dans la base $\mathbf{e}$. Alors les propositions suivantes sont équivalentes :

1) $\mathbf{e}$ est φ-orthogonale ;
2) la matrice de φ dans la base $\mathbf{e}$ est diagonale ;
3) la matrice de q dans la base $\mathbf{e}$ est diagonale ;
4) $\exists\, a_i \in \mathbb{R}$, $i = 1, \ldots, n$ tels que $q(x) = \sum_{i=1}^{n} a_i x_i^2$.

3.5.2 Algorithme de Gauss, signature

Nous allons maintenant décrire un algorithme, dit algorithme de Gauss, qui permet de trouver une base q-orthogonale.

Soit B' une base φ-orthogonale et B une base quelconque, P la matrice de passage de B' à B. Si un vecteur v a pour coordonnées ${}^tX = (x_1,..,x_n)$ dans la base B et ${}^tX' = (x'_1,\ldots,x'_n)$ dans la base B', on a $PX = X'$ donc :

$$q(v) = \sum_{i=1}^{n} a_i x_i'^2$$

$$q(v) = \sum_{i=1}^{n} a_i \left(\sum_{j=1}^{n} P_{ij} x_j \right)^2 . \tag{3.1}$$

C'est-à-dire que la forme q s'écrit comme une somme (pondérée) de carrés de formes linéaires. Cette écriture est déjà en soi très utile, elle permet par exemple de déterminer le cas échéant le signe de q.

Pour trouver une base q-orthogonale, nous allons effectuer le processus inverse : partir de l'expression de $q(v)$ en fonction des x_j et essayer de l'écrire sous la forme (3.1) de somme (pondérée) de carrés de combinaisons linéaires *indépendantes* des coordonnées de v. La matrice de passage de B à B' s'obtient alors en inversant P. La i-ième colonne de cette matrice P^{-1}, qui est le vecteur colonne des coordonnées du i-ième vecteur de la base q-orthogonale, s'obtient en résolvant le système

$$\begin{cases} x'_1 = 0 = \sum_{j=1}^{n} P_{1j} x_j \\ \cdots \\ x'_i = 1 = \sum_{j=1}^{n} P_{ij} x_j \\ \cdots \\ x'_n = 0 = \sum_{j=1}^{n} P_{nj} x_j \end{cases}$$

ou par toute autre méthode d'inversion de matrice, comme en particulier résoudre le système à n inconnues x_i et n paramètres x'_i au second membre : $X' = PX \Leftrightarrow X = P^{-1}X'$.

Algorithme de Gauss

Soit V un $\mathbb{R}$-espace vectoriel de dimension finie n, et soit $\mathbf{e}$ une base de V. Soit $q : V \to \mathbb{R}$ une forme quadratique, et soit $M = (a_{ij})_{1 \leqslant i,j \leqslant n}$ sa matrice représentative dans la base $\mathbf{e}$. Si $x = \sum_{i=1}^{n} x_i e_i$, on a donc

$$q(x) = \sum_{i=1}^{n} a_{ii} x_i^2 + 2 \sum_{1 \leqslant i < j \leqslant n} a_{ij} x_i x_j = P(x_1,\ldots,x_n).$$

On procède par récurrence sur le nombre de variables. À chaque étape, il y a deux cas.

1) S'il existe un indice k tel que $a_{kk} \neq 0$, on regroupe tous les termes faisant intervenir la variable x_k, et on complète le carré. On écrit

$$P(x_1,\ldots,x_n) = a_{kk} x_k^2 + 2 f_k x_k + P_0,$$

où f_k est une forme linéaire en les variables $x_i, i \neq k$, et P_0 est une forme quadratique en les variables $x_i, i \neq k$.
On a alors

$$\begin{aligned} P(x_1,\ldots,x_n) &= a_{kk}(x_k^2 + \tfrac{2}{a_{kk}} f_k x_k) + P_0 \\ &= a_{kk}((x_k + \tfrac{f_k}{a_{kk}})^2 - \tfrac{f_k^2}{a_{kk}^2}) + P_0 \end{aligned} .$$

On peut donc écrire

$$P(x_1,\ldots,x_n) = a_{kk}(x_k + \frac{f_k}{a_{kk}})^2 + P_1,$$

où P_1 est une forme quadratique en les variables $x_i, i \neq k$.

2) Si $a_{kk} = 0$ pour tout k, mais qu'il existe k et ℓ tels que $k < \ell$ et $a_{k\ell} \neq 0$. C'est le cas délicat.
On écrit
$$P(x_1, \ldots, x_n) = 2a_{k\ell}x_k x_\ell + 2f_k x_k + 2f_\ell x_\ell + P_0,$$
où f_k et f_ℓ sont des formes linéaires en les variables $x_i, (i \neq k, \ell)$, et P_0 est une forme quadratique en les variables $x_i, (i \neq k, \ell)$.
On a ainsi
$$P(x_1, \ldots, x_n) = 2a_{k\ell}(x_k + \frac{1}{a_{k\ell}} f_\ell)(x_\ell + \frac{1}{a_{k\ell}} f_k) - \frac{2}{a_{k\ell}} f_k f_\ell + P_0.$$
On a donc
$$P(x_1, \ldots, x_n) = 2a_{k\ell}AB + P_1,$$
avec $A = x_k + \frac{1}{a_{k\ell}} f_\ell, B = x_\ell + \frac{1}{a_{k\ell}} f_k$, et P_1 est une forme quadratique en les variables $x_i, i \neq k, \ell$.
On a alors
$$P(x_1, \ldots, x_n) = \frac{a_{k\ell}}{2}((A+B)^2 - (A-B)^2) + P_1.$$

Si $P_1 = 0$, on arrête. Sinon, on recommence le procédé avec P_1.

Dans chacun des cas, on remarque qu'on supprime autant de variables dans P_1 qu'on fait apparaître de carrés dans l'expression de P. L'algorithme repose sur cette propriété : il est impératif de vérifier que *tous* les termes de P qui contiennent la ou les variables que l'on va faire disparaître sont pris en compte dans les carrés. De ce fait, on supprime une ou deux variables à chaque étape, il est donc évident que le processus se termine par l'écriture de la forme quadratique P comme somme (pondérée) de carrés de la forme

$$q(x) = \alpha_1(L_1(x))^2 + \cdots + \alpha_r(L_r(x))^2,$$

où :

1) chaque $\alpha_i \in \mathbb{R}^*$;
2) chaque L_i est une forme linéaire sur V ;
3) la famille de formes $(L_1, \ldots, L_r)$ est libre.

Si q n'est pas de rang maximal ($r < n$), on complète par des formes linéaires $L_{r+1}, L_{r+2}, \ldots, L_n$ (on les choisit par exemple parmi les formes coordonnées $x_1, \ldots, x_n$) de telle sorte que la famille $(L_1, \ldots, L_n)$ soit libre et on écrit

$$q(x) = \alpha_1(L_1(x))^2 + \cdots + \alpha_r(L_r(x))^2 + 0(L_{r+1})^2 + \ldots + 0(L_n(x)))^2.$$

Calcul de la base q-orthogonale

Pour trouver une base orthogonale pour la forme q, on est ramené à chercher une base dans laquelle les coordonnées d'un vecteur v de coordonnées $(x_1, \ldots, x_n)$ dans la base initiale sont les $(L_1(v), \ldots, L_n(v))$. En effet, dans une telle base l'expression de la forme quadratique sera bien la somme (pondérée) de carrés des coordonnées.

On cherche donc $\mathbf{e}' = (e'_1, e'_2, \ldots, e'_n)$ telle que pour tout v on ait $v = \sum_i L_i(v)e'_i$.
Cela revient à

$$L_j(e'_i) = 0 \text{ si } i \neq j \text{ et } 1 \text{ si } i = j.$$

Les coordonnées de e'_i vérifient donc un système dont la matrice M est obtenue en écrivant en ligne les coefficients des L_j, et de second membre la i-ème colonne de la matrice identité. Il s'agit donc du i-ième vecteur colonne de M^{-1}. Dit autrement, la matrice M constituée par les coefficients des L_i en ligne est la matrice qui fournit le changement de coordonnées de $\mathbf{e}$ vers $\mathbf{e}'$, et la matrice de passage de $\mathbf{e}$ vers $\mathbf{e}'$ est précisément l'inverse de cette matrice (rappel : $X = PX' \Leftrightarrow X' = P^{-1}X$!).

Il résulte du lemme 3.5.4 que la matrice de q dans la base $\mathbf{e}'$ est la matrice

$$M = \text{diag}(\alpha_1, \alpha_2, \ldots, \alpha_r, 0, \ldots, 0)$$

Exemples 3.5.6.

1) On considère la forme quadratique q définie sur $\mathbb{R}^2$ par

$$q(x,y) = x^2 + 4xy.$$

On élimine la variable x en formant un carré contenant tous les termes dépendant de x (forme canonique d'un polynôme du second degré en x dépendant de y vu comme paramètre)

$$q(x,y) = (x+2y)^2 - 4y^2 = x'^2 - 4y'^2, \text{ en posant } x' = x+2y, y' = y.$$

Avec les notations ci-dessus, on a donc $M = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$. Pour trouver la base q-orthogonale, il suffit de chercher l'inverse de cette matrice. La matrice de passage de la base canonique à la base q-orthogonale est donc

$$P = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}.$$

On peut vérifier, en notant Q la matrice de q :

$${}^tPQP = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -4 \end{pmatrix}.$$

2) On considère la forme quadratique q définie sur $\mathbb{R}^3$ par

$$q(x,y,z) = x^2 + 2xy + 4xz + 2yz.$$

On élimine la variable x

$$q(x,y,z) = (x+y+2z)^2 - (y+2z)^2 + 2yz = (x+y+2z)^2 - y^2 - 4z^2 - 2yz.$$

Puis on élimine y dans ce qui reste

$$q(x,y,z) = (x+y+2z)^2 - (y+z)^2 - 3z^2 = x'^2 - y'^2 - 3z'^2.$$

Pour trouver la base q-orthogonale correspondante, on résout le système :

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

en fonction des paramètres (x',y',z'). On remarque que ce système est échelonné, ce qui est une conséquence de l'élimination d'une variable à chaque étape, et implique l'indépendance des formes linéaires apparues dans l'algorithme.
La matrice du système résolu

$$\begin{cases} x &= x'-y'-z' \\ y &= y'-z' \\ z &= z' \end{cases}$$

est la matrice de passage vers la nouvelle base, qui est donc

$$e'_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad e'_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad e'_3 = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}.$$

3) On considère la forme quadratique q définie sur $\mathbb{R}^3$ par

$$q(x,y,z) = x^2 + 3y^2 + 5z^2 + 4xy + 6xz + 8yz.$$

On élimine la variable x en absorbant les termes en xy et xz dans un carré :

$$\begin{aligned} q(x,y,z) &= (x+2y+3z)^2-(2y+3z)^2+3y^2+5z^2+8yz \\ &= (x+2y+3z)^2-y^2-4z^2-4yz \\ &= (x+2y+3z)^2-(y+2z)^2. \end{aligned}$$

Le processus s'arrête alors qu'on n'a trouvé que deux carrés indépendants : la forme q n'est pas de rang maximal. Pour trouver une base q-orthogonale, on résout le système

$$\left\{\begin{array}{rcl} x+2y+3z &=& x' \\ y+2z &=& y' \\ z &=& z' \end{array}\right.$$

en rajoutant une dernière ligne indépendante des précédentes de façon totalement arbitraire (ici choisie pour rendre l'inversion de la matrice particulièrement facile !).
La matrice du système résolu

$$\left\{\begin{array}{rcl} x &=& x'-2y'+z' \\ y &=& y'-2z' \\ z &=& z' \end{array}\right.$$

est la matrice de passage vers la nouvelle base, qui est donc

$$e_1' = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad e_2' = \begin{pmatrix} -2 \\ 1 \\ 0 \end{pmatrix}, \quad e_3' = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

4) Soit $q : \mathbb{R}^4 \to \mathbb{R}$ l'application qui à $\mathbf{u} = \begin{pmatrix} x \\ y \\ z \\ t \end{pmatrix}$ associe

$$q(\mathbf{u}) = x^2+2xy+2xz+2xt+y^2+6yz-2yt+z^2+10zt+t^2.$$

L'application q est bien une forme quadratique car c'est un polynôme de degré 2 homogène. Sa matrice est

$$Q = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 3 & -1 \\ 1 & 3 & 1 & 5 \\ 1 & -1 & 5 & 1 \end{pmatrix}.$$

Appliquons l'algorithme de Gauss à q pour trouver une base q-orthogonale. On a

$$\begin{aligned} q(\mathbf{u}) &= x^2+2(y+z+t)x+y^2+6yz-2yt+z^2+10zt+t^2 \\ &= (x+y+z+t)^2-(y+z+t)^2+y^2+6yz-2yt+z^2+10zt+t^2 \\ &= (x+y+z+t)^2+4yz-4yt+8zt. \end{aligned}$$

On est dans le 2e cas où il ne reste aucun carré. On va donc supprimer deux variables en interprétant un produit comme une différence de deux carrés :

$$\begin{aligned} 4yz-4yt+8zt &= 4(yz+(-t)y+(2t)z) \\ &= 4((y+2t)(z-t)+2t^2) \\ &= 4(y+2t)(z-t)+8t^2 \\ &= (y+z+t)^2-(y-z+3t)^2+8t^2 \end{aligned}$$

Finalement, on obtient

$$q(\mathbf{u}) = (x+y+z+t)^2 + (y+z+t)^2 - (y-z+3t)^2 + 8t^2.$$

On a donc $\mathrm{rg}(q) = 4$. On a

$$\begin{cases} L_1(u) &= x+y+z+t \\ L_2(u) &= y+z+t \\ L_3(u) &= y-z+3t \\ L_4(u) &= t \end{cases}$$

La matrice du système est donnée par

$$M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Cette matrice est presque triangulaire supérieure, il y a donc assez peu de manipulations à faire pour résoudre le système. On calcule M^{-1} et on lit e'_1 dans la 1re colonne de M^{-1}, e'_2 dans la 2e colonne, etc.

$$e'_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \; e'_2 = \begin{pmatrix} -1 \\ 1/2 \\ 1/2 \\ 0 \end{pmatrix}, \; e'_3 = \begin{pmatrix} 0 \\ 1/2 \\ -1/2 \\ 0 \end{pmatrix}, \; e'_4 = \begin{pmatrix} 0 \\ -2 \\ 1 \\ 1 \end{pmatrix}$$

Ces vecteurs (e'_1, e'_2, e'_3, e'_4) forment donc une base q-orthogonale. On vérifie en appliquant la formule de changement de base de la base (e'_1, e'_2, e'_3, e'_4) où q est diagonale (de coefficients 1, 1, -1 et 8) vers la base canonique.

Remarque 3.5.7. Si $\phi : V \times V \to \mathbb{R}$ est bilinéaire symétrique, alors en appliquant l'algorithme de Gauss à la forme quadratique

$$q_b : V \to \mathbb{R}, x \mapsto \varphi(x,x),$$

on trouve une base $\mathbf{v}$ qui est q_φ-orthogonale. Mais par définition, $\mathbf{v}$ est donc orthogonale pour la forme polaire de q_φ, qui est φ.

En particulier, le nombre r de carrés qui apparaissent dans l'écriture $q(x) = \sum_{i=1}^{r} a_i L_i(x)^2$ est le rang de la forme bilinéaire.

Cet algorithme permet donc de trouver une base φ-orthogonale pour n'importe quelle forme bilinéaire symétrique φ, ainsi que son rang. On peut programmer l'algorithme de Gauss sur machine, mais à condition que les coefficients de la forme quadratique soient représentables exactement sur machine, sinon le résultat obtenu peut être invalide en raison des erreurs d'arrondis (toutefois Gauss fonctionne avec des coefficients approchés si $r_+ = n$ ou si $r_- = n$, cela correspond à la factorisation de Cholesky d'une matrice).

Le théorème qui suit affirme que r_+, le nombre de coefficients strictement positifs, et r_-, le nombre de coefficients strictement négatifs des carrés $L_i(x)^2$, ne dépendent pas des choix faits au cours de l'algorithme de Gauss de réduction de la forme quadratique.

Théorème 3.5.8 (Théorème d'inertie de Sylvester)**.** *Soit V un $\mathbb{R}$-espace vectoriel de dimension finie n, et soit $q : V \to \mathbb{R}$ une forme quadratique. Soit* **e** *une base q-orthogonale. Soit*

$$r_+ = \mathrm{card}\{i | q(e_i) > 0\}, \quad r_- = \mathrm{card}\{i | q(e_i) < 0\}.$$

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

Ce théorème n'est valable que pour des formes réelles.

Preuve. Soit $\mathbf{e} = (e_1, \dots, e_n)$ une base q-orthogonale. Posons $\alpha_i = q(e_i) = \varphi_q(e_i, e_i)$ et $r = r_+ + r_-$. Changer l'ordre des vecteurs de $\mathbf{e}$ ne change pas r_+ et r_-, ni le fait que la base soit q-orthogonale. On peut donc supposer sans perte de généralité que l'on a

$$q(e_i) > 0, i = 1, \dots, r_+, \quad q(e_i) < 0, i = r_+ + 1, \dots, r, \quad q(e_i) = 0, i = r+1, \dots, n.$$

Puisque $\mathbf{e}$ est q-orthogonale (c'est-à-dire φ_q-orthogonale), on obtient que M, la matrice de q dans la base $\mathbf{e}$, s'écrit

$$M = \begin{pmatrix} q(e_1) & \dots & 0 \\ & \ddots & \\ 0 & \dots & q(e_n) \end{pmatrix}.$$

Or, seuls les réels $q(e_1), \dots, q(e_r)$ sont non nuls. Le rang d'une matrice diagonale étant le nombre de termes diagonaux non nuls, on a bien $rg(q) = r = r_+ + r_-$.

Soit maintenant $\mathbf{e}'$ une autre base q-orthogonale. Soient (r'_+, r'_-) le couple d'entiers correspondant. Remarquons que l'on a $r'_+ + r'_- = \mathrm{rg}(q) = r$ par le point précédent. Comme précédemment, quitte à changer l'ordre des vecteurs, on peut supposer que

$$q(e'_i) > 0, i = 1, \dots, r'_+, \quad q(e'_i) < 0, i = r'_+ + 1, \dots, r, \quad q(e'_i) = 0, i = r+1, \dots, n.$$

Montrons que $e_1, \dots, e_{r_+}, e'_{r'_+ + 1}, \dots, e'_n$ sont linéairement indépendants. Supposons que l'on ait une relation

$$\lambda_1 e_1 + \dots + \lambda_{r_+} e_{r_+} + \lambda_{r'_+ + 1} e'_{r'_+ + 1} + \dots + \lambda_n e'_n = 0.$$

On a donc

$$\lambda_1 e_1 + \dots + \lambda_{r_+} e_{r_+} = -(\lambda_{r'_+ + 1} e'_{r'_+ + 1} + \dots \lambda_n e'_n).$$

En appliquant q des deux côtés, et en utilisant le fait que les bases $\mathbf{e}$ et $\mathbf{e}'$ sont q-orthogonales, on obtient

$$\sum_{i=1}^{r_+} q(e_i)\lambda_i^2 = \sum_{i=r'_+ + 1}^{n} q(e'_i)\lambda_i^2.$$

Par choix de $\mathbf{e}$ et de $\mathbf{e}'$, le membre de gauche est $\geqslant 0$ et le membre de droite est $\leqslant 0$.

On en déduit que l'on a

$$\sum_{i=1}^{r_+} q(e_i)\lambda_i^2 = 0,$$

et puisque $q(e_i) > 0$ pour $i = 1, \dots, r_+$, on en déduit

$$\lambda_1 = \dots = \lambda_{r_+} = 0.$$

Mais alors, on a

$$\lambda_{r'_+ + 1} e'_{r'_+ + 1} + \dots \lambda_n e'_n = 0,$$

et comme $\mathbf{e}'$ est une base, on en déduit

$$\lambda_{r'_+ + 1} = \dots = \lambda_n = 0.$$

Ainsi, $e_1, \dots, e_{r_+}, e'_{r'_+ + 1}, \dots, e'_n$ sont $r_+ + (n - r'_+)$ vecteurs linéairement indépendants dans un espace vectoriel de dimension n. On a donc

$$r_+ + (n - r'_+) \leqslant n,$$

et donc $r_+ \leqslant r'_+$. En échangeant les rôles de $\mathbf{e}$ et $\mathbf{e}'$, on a de même $r'_+ \leqslant r_+$.

On a donc $r_+ = r'_+$, et comme on a $\mathrm{rg}(q) = r_+ + r_- = r'_+ + r'_-$, on en déduit $r_- = r'_-$. Ceci achève la démonstration.

Cela conduit à la définition suivante.

Définition 3.5.9. Soit V un $\mathbb{R}$-espace vectoriel de dimension finie n, et soit $q : V \to \mathbb{R}$ une forme quadratique. Le couple (r_+, r_-) est appelé la *signature* de q.

Remarque 3.5.10. Pour calculer la signature d'une forme quadratique q, il suffit d'utiliser l'algorithme de Gauss pour écrire $q(x)$ sous la forme

$$\alpha_1(u_{11}x_1 + \cdots + u_{1n}x_n)^2 + \cdots + \alpha_r(u_{r1}x_1 + \cdots + u_{rn}x_n)^2,$$

et de compter le nombre de coefficients α_i qui sont strictement plus grand que 0 et strictement plus petit que 0.

En effet, on a vu que si $\mathbf{v} = (v_1, \ldots, v_n)$ est la base q-orthogonale obtenue à la fin de l'algorithme de Gauss, et M est la matrice de q dans cette base, alors

$$M = \mathrm{diag}(\alpha_1, \ldots, \alpha_r, 0, \ldots, 0).$$

Mais les coefficients diagonaux de M sont exactement les réels $q(v_i)$, et on conclut en utilisant la définition de r_+ et r_-.

Exemple 3.5.11. Les signatures des formes quadratiques des exemples 3.5.6 sont : 1) (1,1) ; 2) (1,2) ; 3) (1,1) ; 4) (3,1).

Chapitre 4

Produits scalaires

4.1 Rappels dans le plan et l'espace

4.1.1 Dans le plan

Soient $u_1(x_1,y_1)$ et $u_2(x_2,y_2)$ deux vecteurs du plan. On définit le produit scalaire de u_1 et u_2 par

$$\langle u_1|u_2\rangle = x_1x_2 + y_1y_2.$$

Propriétés

— Le produit scalaire se comporte comme un produit, il est linéaire par rapport à chacun de ses arguments (les vecteurs u_1 et u_2)

— Le produit scalaire de $u(x,y)$ avec lui-même est $\langle u|u\rangle = x^2+y^2$, il est donc toujours positif ou nul. Il n'est nul que si u est nul.

— On définit la norme de u par $\|u\| = \sqrt{\langle u|u\rangle}$. La distance entre deux points ou vecteurs est la norme de leur différence.

Si z_1 est l'affixe de u_1 (le complexe correspondant à u_1) et z_2 celui de u_2, alors en notant $\mathrm{Re}(z)$ la partie réelle de z :

$$\langle u_1|u_2\rangle = x_1x_2 + y_1y_2 = \mathrm{Re}((x_1 - iy_1)(x_2+iy_2)) = \mathrm{Re}(\overline{z_1}z_2)$$

Donc le produit scalaire est invariant par rotation [1], puisque

$$\mathrm{Re}(\overline{e^{i\theta}z_1}e^{i\theta}z_2) = \mathrm{Re}(e^{-i\theta}\overline{z_1}e^{i\theta}z_2) = \mathrm{Re}(\overline{z_1}z_2).$$

On peut aussi le vérifier avec la matrice P de la rotation d'angle θ :

$$P = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$

qui vérifie $P^tP = I_2$.

Soit φ l'angle entre les vecteurs u_1 et u_2. Effectuons la rotation qui met u_1 selon l'axe des x dans le bon sens, on a alors $x_1 = \|u_1\|, y_1 = 0$ donc $\langle u_1|u_2\rangle = x_1x_2 = \|u_1\|\|u_2\|\cos(\varphi)$ En particulier, on a *l'inégalité de Cauchy-Schwarz* :

$$|\langle u_1|u_2\rangle| \leqslant \|u_1\|\|u_2\|.$$

Si $\langle u|v\rangle = 0$, on dit que les vecteurs u et v sont orthogonaux, on a alors le *théorème de Pythagore*

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2.$$

1. Le produit scalaire est aussi invariant par symétrie

Lorsqu'une base est composée de vecteurs de norme 1 orthogonaux entre eux, on parle de base orthonormée. Si $\{u_1, u_2\}$ est une telle base, alors on a

$$u = \langle u_1|u\rangle u_1 + \langle u_2|u\rangle u_2.$$

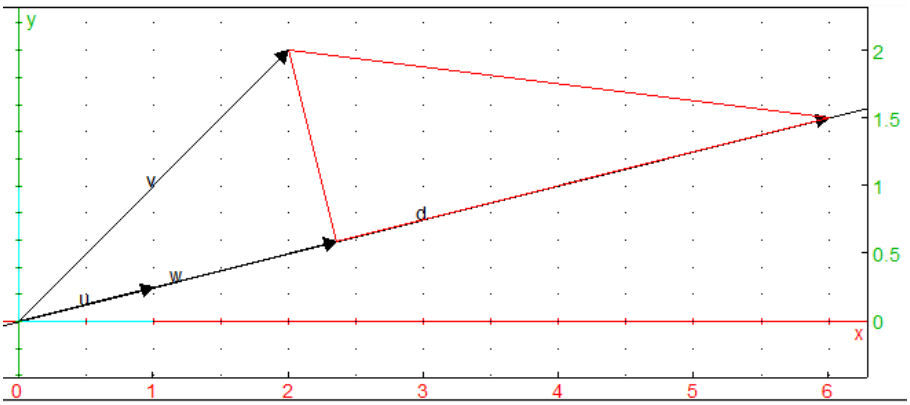
Si on se donne un vecteur $u \neq 0$, on peut construire une base orthonormée dont le premier vecteur est $u_1 = \frac{u}{\|u\|}$. On définit la projection orthogonale sur la droite vectorielle D engendrée par u par

$$p(v) = \langle u_1|v\rangle u_1.$$

On vérifie que $v - p(v)$ est orthogonal à u_1 :

$$\langle u_1|v - p(v)\rangle = \langle u_1|v\rangle - \langle u_1|p(v)\rangle = \langle u_1|v\rangle - \langle u_1|\langle u_1|v\rangle u_1\rangle = \langle u_1|v\rangle - \langle u_1|v\rangle\langle u_1|u_1\rangle = 0.$$

Le vecteur de D le plus proche de v est $w = p(v)$. En effet si d est un vecteur de D, on applique le théorème de Pythagore dans le triangle de sommets les extrémités de d, $w = p(v)$ et v qui est rectangle (en $w = p(v)$).



4.1.2 Dans l'espace

Si $u_1(x_1, y_1, z_1)$ et $u_2(x_2, y_2, z_2)$ sont deux vecteurs de $\mathbb{R}^3$, on définit leur produit scalaire par :

$$\langle u_1|u_2\rangle = x_1x_2 + y_1y_2 + z_1z_2$$

On vérifie les mêmes propriétés que dans le plan : le produit scalaire se comporte comme un produit (linéarité par rapport à chaque argument), $\langle u|u\rangle$ est positif et ne s'annule que si $u = 0$. Comme c'est le produit scalaire du plan si on se restreint aux plans de coordonnées Oxy, Oxz, Oyz, il est invariant par rotation d'axe les vecteurs de base. On a donc toujours

$$\langle u_1|u_2\rangle = \|u_1\|\|u_2\|\cos(u_1, u_2)$$

(en utilisant les angles d'Euler : faire une rotation d'axe Oz pour que le plan u_1, u_2 contienne Ox, puis une rotation selon Ox pour que le plan u_1, u_2 soit le plan de coordonnées Oxy). Donc l'inégalité de Cauchy-Schwarz est toujours valide. De même que le théorème de Pythagore.

On parle toujours de base orthonormée pour une base de 3 vecteurs de norme 1 orthogonaux entre eux 2 à 2. Les coordonnées d'un vecteur u dans une base orthonormée $\{u_1, u_2, u_3\}$ se calculent par la formule :

$$u = \langle u_1|u\rangle u_1 + \langle u_2|u\rangle u_2 + \langle u_3|u\rangle u_3$$

Si on se donne une droite vectorielle D de vecteur directeur u, on peut créer une base orthonormale de premier vecteur $u_1 = \frac{u}{\|u\|}$. La projection orthogonale d'un vecteur v sur la droite D est toujours obtenue par

$$p(v) = \langle u_1|v\rangle u_1$$

et c'est le vecteur de D le plus proche de v.

Si on se donne un plan vectoriel P engendré par deux vecteurs u et v on peut créer une base orthonormale de premier vecteur $u_1 = \frac{u}{\|u\|}$ et dont le deuxième vecteur est dans le plan u, v. Pour cela, on modifie v en un vecteur $\tilde{v}$ orthogonal à u en retirant à v la projection orthogonale de v sur u :

$$\tilde{v} = v - \langle u_1 | v \rangle u_1$$

(ce qui revient à projeter v sur la droite orthogonale à celle engendrée par u) puis on normalise ce qui donne un vecteur u_2 de norme 1 orthogonal à u_1

$$u_2 = \frac{\tilde{v}}{\|\tilde{v}\|}.$$

À ce stade, on peut définir la projection orthogonale sur P par

$$p(w) = \langle u_1 | w \rangle u_1 + \langle u_2 | w \rangle u_2.$$

On peut compléter la famille orthonormée $\{u_1, u_2\}$ avec le produit vectoriel des deux vecteurs u_1 et u_2, mais cette construction est spécifique à la dimension 3. Pour pouvoir généraliser en dimension plus grande, on peut aussi prendre un troisième vecteur w qui n'appartient pas au plan P, on le modifie en un vecteur orthogonal à P en lui retirant sa projection orthogonale sur P et on le normalise en un vecteur u_3. Le vecteur de P le plus proche de w est $p(w)$, toujours à cause du théorème de Pythagore.

Exemple 4.1.1. Soit P le plan engendré par les vecteurs $u = (1,1,0)$ et $v = (1,0,-1)$. On a $u_1 = u/\sqrt{2}$. Donc

$$p_u(v) = v - \langle u_1 | v \rangle u_1 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} - \langle \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} | \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \rangle \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix}$$

puis

$$u_2 = \frac{p_u(v)}{\|p_u(v)\|} = \frac{\begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix}}{\left\| \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix} \right\|} = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix}.$$

Pour compléter la base avec un vecteur u_3, en dimension 3 on peut utiliser le produit vectoriel de u_1 et u_2

$$u_3 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \wedge \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix} = \frac{1}{2\sqrt{3}} \begin{pmatrix} -2 \\ 2 \\ -2 \end{pmatrix} = \frac{1}{\sqrt{3}} \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix}$$

ou prendre un vecteur w, par exemple $w = (1,0,0)$ et retrancher la projection orthogonale de w sur P

$$\tilde{u_3} = w - \langle u_1 | w \rangle u_1 - \langle u_2 | w \rangle u_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \langle \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} | \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \rangle \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - \langle \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix} | \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \rangle \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix}$$

donc

$$\tilde{u_3} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - \frac{1}{6} \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix} = \frac{1}{6} \begin{pmatrix} 6-3-1 \\ 0-3+1 \\ 2 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

on retrouve bien un multiple du u_3 précédent. En normalisant, on trouverait $-u_3$, ce qui illustre qu'il y a deux choix à chaque étape.

4.2 Définitions et exemples

Nous voulons maintenant généraliser la notion de produit scalaire — et donc de longueur, de distance et d'angle — à un espace vectoriel réel arbitraire. Soient

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

deux vecteurs de $\mathbb{R}^n$, le produit scalaire canonique est défini par :

$$x \cdot y = {}^t xy = \sum_{i=1}^{n} x_i y_i.$$

L'application $(x,y) \mapsto x \cdot y$ est une forme bilinéaire symétrique. La longueur d'un vecteur $x \in \mathbb{R}^n$ pour $n = 2$ et $n = 3$ peut être calculée par la formule

$$\|x\| = \sqrt{x \cdot x}.$$

De même, nous souhaiterions associer une notion de longueur (on parle plutôt de norme pour un vecteur) à une forme bilinéaire φ en posant $\|x\| = \sqrt{\varphi(x,x)}$. Malheureusement, il n'est pas sûr que cette quantité soit définie : en effet si $\varphi(x,x) < 0$, la racine carrée n'est pas définie. De plus, on souhaite que la norme d'un vecteur soit strictement positive pour un x non nul (or nous ne voulons pas une distance 0 entre deux vecteurs distincts).

Ces considérations amènent les définitions suivantes :

Définition 4.2.1. Soit V un espace vectoriel réel. On dit qu'une forme bilinéaire symétrique $\varphi : V \times V \to \mathbb{R}$ est *positive* si $\varphi(x,x) \geqslant 0$ pour tout $x \in V$, et *définie positive* si $\varphi(x,x) > 0$ pour tout $x \in V, x \neq 0$.

Remarquons que φ est définie positive si et seulement si

— $\forall x \in V$, $\varphi(x,x) \geqslant 0$, et

— $\varphi(x,x) = 0 \Rightarrow x = 0_V$.

C'est en général cette reformulation de la définition que l'on utilise en pratique pour vérifier si oui ou non une forme bilinéaire donnée est définie positive.

Définition 4.2.2. Soit V un $\mathbb{R}$-espace vectoriel (non nécessairement de dimension finie). Un *produit scalaire* sur V est une forme bilinéaire symétrique et définie positive sur V :

$$\langle \cdot | \cdot \rangle : \begin{cases} V \times V & \to & \mathbb{R} \\ (x,y) & \mapsto & \langle x|y \rangle \end{cases}$$

On dit que V muni du produit scalaire $\langle \cdot | \cdot \rangle$ est un espace *préhilbertien* réel.

Remarque 4.2.3. On expliquera brièvement plus loin l'utilisation du préfixe « pré »-hilbertien, voir la remarque 4.3.13. On utilise aussi le terme d'espace *euclidien* si V est un $\mathbb{R}$-espace vectoriel de dimension finie muni d'un produit scalaire. Le terme préhilbertien s'emploie aussi dans le cas de produits scalaires hermitiens définis sur un $\mathbb{C}$-espace vectoriel, cf. l'appendice C. Dans la suite de ce chapitre, on donne des résultats pour des espaces préhilbertiens dans le cas réel, la plupart des résultats se généralisent aux préhilbertiens complexes.

Exemples 4.2.4.

1) Le produit scalaire usuel sur $\mathbb{R}^n$ $x \cdot y = \sum_{i=1}^{n} x_i y_i$.
2) La forme bilinéaire qui à deux fonctions f et g continues de $[a,b]$ à valeurs dans $\mathbb{R}$ associe l'intégrale entre a et b de leur produit :

$$\langle \cdot | \cdot \rangle : \begin{cases} \mathscr{C}^0([a,b],\mathbb{R}) \times \mathscr{C}^0([a,b],\mathbb{R}) & \to & \mathbb{R} \\ (f,g) & \mapsto & \langle f|g \rangle = \int_a^b f(t)g(t)\,\mathrm{d}t \end{cases}$$

Montrons que c'est un produit scalaire.

(a) Montrons que $\langle \cdot | \cdot \rangle$ est symétrique. En effet, pour tout $f, g \in \mathscr{C}^0([a,b],\mathbb{R})$, on a

$$\langle g|f\rangle = \int_a^b g(t)f(t)\,\mathrm{d}t = \int_a^b f(t)g(t)\,\mathrm{d}t = \langle f|g\rangle.$$

(b) Montrons que $\langle \cdot | \cdot \rangle$ est bilinéaire. Pour tout $f_1, f_2, f, g \in \mathscr{C}^0([a,b],\mathbb{R}), \lambda \in \mathbb{R}$, on a

$$\begin{array}{rcl} \langle f_1+f_2|g\rangle & = & \int_a^b (f_1+f_2)(t)g(t)\,\mathrm{d}t \\ & = & \int_a^b (f_1(t)+f_2(t))g(t)\,\mathrm{d}t \\ & = & \int_a^b f_1(t)g(t)\,\mathrm{d}t + \int_a^b f_2(t)g(t)\,\mathrm{d}t \\ & = & \langle f_1|g\rangle + \langle f_2|g\rangle \end{array}$$

et :

$$\begin{array}{rcl} \langle \lambda f|g\rangle & = & \int_a^b (\lambda f)(t)g(t)\,\mathrm{d}t \\ & = & \int_a^b \lambda f(t)g(t)\,\mathrm{d}t \\ & = & \lambda \int_a^b f(t)g(t)\,\mathrm{d}t \\ & = & \lambda\langle f|g\rangle. \end{array}$$

Par symétrie, il découle que

$$\langle f|g_1+g_2\rangle = \langle f|g_1\rangle + \langle f|g_2\rangle \text{ et } \langle f|\lambda g\rangle = \lambda\langle f|g\rangle$$

pour tout $f, g, g_1, g_2 \in \mathbb{R}[X], \lambda \in \mathbb{R}$

Ainsi, $\langle \cdot | \cdot \rangle$ est bilinéaire.

(c) Montrons enfin que $\langle \cdot | \cdot \rangle$ est définie positive. On va utiliser pour cela la reformulation de la définition 4.2.1.

Pour tout $f \in \mathscr{C}^0([a,b],\mathbb{R})$, on a

$$\langle f|f\rangle = \int_a^b f(t)^2\,\mathrm{d}t.$$

Or, l'intégrale d'une fonction positive est positive. Comme la fonction $f^2(t)$ est positive, on en déduit que

$$\langle f|f\rangle \geqslant 0 \text{ pour tout } f \in \mathscr{C}^0([a,b],\mathbb{R}).$$

Supposons maintenant que l'on a $\langle f|f\rangle = 0$, c'est-à-dire que

$$\int_a^b f(t)^2\,\mathrm{d}t = 0.$$

Or l'intégrale d'une fonction *positive* et *continue* $f : [a,b] \to \mathbb{R}$ est nulle si et seulement si f est identiquement nulle. Comme la fonction

$$[a,b] \to \mathbb{R},\ t \mapsto f(t)^2$$

est positive et continue, on en déduit

$$f(t)^2 = 0 \text{ pour tout } t \in [a,b],$$

c'est-à-dire $f = 0$.

3) Pour toute fonction p continue et strictement positive sur $[a,b]$, la forme bilinéaire :

$$\langle \cdot | \cdot \rangle : \left\{ \begin{array}{ccc} \mathscr{C}^0([a,b],\mathbb{R}) \times \mathscr{C}^0([a,b],\mathbb{R}) & \to & \mathbb{R} \\ (f,g) & \mapsto & \langle f|g\rangle = \int_a^b p(t)f(t)g(t)\,\mathrm{d}t \end{array} \right.$$

est un produit scalaire (exercice).

4) L'application définie sur les matrices carrées réelles $\mathcal{M}_n(\mathbb{R})$ par

$$(M,N) \mapsto \mathrm{Tr}({}^tMN)$$

est un produit scalaire.

5) La forme bilinéaire définie sur $\mathbb{R}^2$ par :

$$\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right) \mapsto x_1y_1 - x_2y_2$$

n'est pas un produit scalaire. C'est bien une forme bilinéaire symétrique, mais elle n'est pas positive.

6) L'application qui associe à deux polynômes le produit de leur valeur en 0 :

$$\varphi : \begin{cases} \mathbb{R}[X] \times \mathbb{R}[X] & \to & \mathbb{R} \\ (P,Q) & \mapsto & P(0)Q(0) \end{cases}$$

n'est pas un produit scalaire. Elle est bien bilinéaire, symétrique, positive, mais pas définie positive. Par exemple, on a $\varphi(X,X) = 0$, mais X n'est pas le polynôme nul.

4.3 Géométrie

Les propriétés du produit scalaire permettent, comme dans le cas classique, de définir la « longueur », ou *norme* d'un vecteur de V.

Définition 4.3.1. Soit $(V, \langle \cdot | \cdot \rangle)$ un espace préhilbertien. Pour tout $x \in V$, on définit la *norme* de x, notée $\|x\|$, par

$$\|x\| = \sqrt{\langle x|x \rangle}.$$

Notons que par définition d'un produit scalaire, $\|x\| \geqslant 0$, et $\|x\| = 0$ si et seulement si $x = 0$.

Définition 4.3.2. Soit $(V, \langle \cdot | \cdot \rangle)$ un espace préhilbertien. Soient $v, w \in V$. On définit la distance entre v et w par

$$d(v,w) = \|v - w\|.$$

Encore une fois, la distance entre v et w est positive et n'est nulle que si $v = w$.

Par analogie avec la géométrie euclidienne en dimension 2 ou trois, il est tentant de poser la définition suivante :

Définition 4.3.3. Soit $(V, \langle \cdot | \cdot \rangle)$ un espace préhilbertien. Soient $v, w \in V$ avec $v, w \neq 0$. On définit l'angle entre v et w par

$$\theta = \arccos\left(\frac{\langle v|w \rangle}{\|v\| \times \|w\|} \right).$$

Remarque 4.3.4. Avec cette définition de θ, l'angle entre v et w, nous avons automatiquement $\theta \in [0, \pi]$. Par ailleurs, il s'agit d'une angle non orienté : θ ne dépend pas de l'ordre de v et w.

Malheureusement, ce n'est pas évident que cette définition soit bien posée. En effet, la fonction arccos n'est définie que pour des nombres réels x satisfaisant la condition $-1 \leqslant x \leqslant 1$ ou autrement dit $|x| \leqslant 1$. Nous devons donc vérifier la proposition suivante :

Proposition 4.3.5 (Inégalité de Cauchy-Schwarz)**.** *Soit $(V, \langle \cdot | \cdot \rangle)$ un espace préhilbertien. Pour tout $x, y \in V$, on a*

$$|\langle x|y \rangle| \leqslant \|x\| \times \|y\|,$$

et on a égalité dans cette expression si et seulement si la famille $\{x, y\}$ est liée dans V, c'est-à-dire s'il existe $\lambda, \mu \in \mathbb{R}, (\lambda, \mu) \neq (0,0)$, tels que $\lambda x + \mu y = 0$.

Exemples 4.3.6.

1) Avec le produit scalaire usuel sur $\mathbb{R}^2$ et les vecteurs $x=(1,1)$ et $y=(a,b)$, on obtient l'inégalité :

$$|\langle x|y\rangle| = |a+b| \leqslant \|x\|\|y\| = \sqrt{2}\sqrt{a^2+b^2}$$

avec égalité si et seulement si x et y sont colinéaires, donc si $a=b$.

2) Avec le produit scalaire usuel sur $\mathbb{R}^3$ et les vecteurs $x=(1,1,1)$ et $y=(a,b,c)$, on obtient l'inégalité :

$$|\langle x|y\rangle| = |a+b+c| \leqslant \|x\|\|y\| = \sqrt{3}\sqrt{a^2+b^2+c^2}$$

avec égalité si et seulement si x et y sont colinéaires, donc si $a=b=c$.

3) Avec le produit scalaire

$$\langle f|g\rangle = \int_a^b f(t)g(t)\,dt, \quad \|f\| = \sqrt{\int_a^b f(t)^2\,dt}$$

sur les fonctions continues sur $[a,b]$,

(a) pour $f(t)=1$ on obtient l'inégalité

$$\left|\int_a^b g(t)\,dt\right| \leqslant \sqrt{\int_a^b dt}\sqrt{\int_a^b g(t)^2\,dt} = \sqrt{b-a}\sqrt{\int_a^b g(t)^2\,dt}\,;$$

(b) pour $f(t)=t$ on obtient l'inégalité

$$\left|\int_a^b tg(t)\,dt\right| \leqslant \sqrt{\int_a^b t^2\,dt}\sqrt{\int_a^b g(t)^2\,dt} = \sqrt{\frac{b^3-a^3}{3}}\sqrt{\int_a^b g(t)^2\,dt}.$$

Preuve. Le résultat étant immédiat si x ou y est égal à 0, on peut supposer $x,y\neq 0$: si $x,y\neq 0$ nous avons qu'il existe $\lambda,\mu\in\mathbb{R},(\lambda,\mu)\neq(0,0)$ tels que $\lambda x+\mu y=0$ si et seulement s'il existe $t\in\mathbb{R}$ tel que $x+ty=0$. Considérons la fonction de t

$$f(t) = \langle x+ty|x+ty\rangle = t^2\|y\|^2+2t\langle x|y\rangle+\|x\|^2.$$

Ceci est une fonction quadratique de t qui ne prend pas de valeurs négatives : elle a donc un discriminant $\Delta\leqslant 0$, c'est-à-dire

$$\Delta = 4(\langle x|y\rangle)^2 - 4\|x\|^2\|y\|^2 \leqslant 0.$$

On a donc que

$$(\langle x|y\rangle)^2 \leqslant \|x\|^2\|y\|^2$$

d'où

$$|\langle x|y\rangle| \leqslant \|x\|\|y\|.$$

De plus, on a égalité dans cette expression si et seulement si $\Delta=0$, c'est-à-dire si et seulement si il existe t tel que $f(t)=0$. Par définition de $f(t)$, nous avons égalité dans cette expression si et seulement s'il existe t tel que $x+ty=0$.

L'inégalité de Cauchy-Schwarz est donc valable et notre définition de θ est bien posée.

Un certain nombre de formules de géométrie dans l'espace sont toujours valables dans ce contexte :

Lemme 4.3.7 (Théorème de Pythagore)**.** *Soit $(V,\langle\cdot|\cdot\rangle)$ un espace préhilbertien et soient $v,w\in V$ avec $v,w\neq 0_V$. Soit θ l'angle entre v et w. Alors on a*

$$\|v-w\|^2 = \|v\|^2+\|w\|^2 \Leftrightarrow \theta = \pi/2.$$

Preuve. On note tout d'abord que par définition $\theta = \pi/2$ si et seulement si $\langle v|w\rangle = 0$. Par définition,

$$\begin{aligned}\|v-w\|^2 &= \langle v-w|v-w\rangle \\ &= \langle v|v\rangle + \langle w|w\rangle - 2\langle v|w\rangle \\ &= \|v\|^2 + \|w\|^2 - 2\langle v|w\rangle\end{aligned}$$

et donc

$$\|v-w\|^2 = \|v\|^2 + \|w\|^2 \Leftrightarrow \langle v|w\rangle = 0 \Leftrightarrow \theta = \pi/2.$$

Lemme 4.3.8 (Identité du parallélogramme)**.** *Soit $(V, \langle\cdot|\cdot\rangle)$ un espace préhilbertien et soient $v, w \in V$. On a alors*

$$\|v+w\|^2 + \|v-w\|^2 = 2(\|v\|^2 + \|w\|^2).$$

Preuve. Exercice pour le lecteur.

Lemme 4.3.9 (Inégalité triangulaire)**.** *Soit $(V, \langle\cdot|\cdot\rangle)$ un espace préhilbertien et soient $v, w \in V$. On a alors*

$$\|v+w\| \leqslant \|v\| + \|w\|.$$

Preuve. On a que

$$\|v+w\|^2 = \|v\|^2 + \|w\|^2 + 2\langle v|w\rangle.$$

Par l'inégalité de Cauchy-Schwarz on a que

$$\|v+w\|^2 \leqslant \|v\|^2 + \|w\|^2 + 2\|v\| \times \|w\| = (\|v\| + \|w\|)^2.$$

Puisque $\|v+w\|$ et $\|v\| + \|w\|$ sont positifs, on peut prend la racine carrée des deux membres pour déduire que

$$\|v+w\| \leqslant \|v\| + \|w\|.$$

Les deux lemmes suivants sont souvent très utiles.

Lemme 4.3.10. *Soit $(V, \langle\cdot|\cdot\rangle)$ un espace préhilbertien, et soient $x_1, \ldots, x_k \in V$ une famille de vecteurs deux à deux orthogonaux. Alors on a*

$$\|x_1 + \cdots + x_k\|^2 = \|x_1\|^2 + \cdots + \|x_k\|^2.$$

Preuve. Supposons $x_1, \ldots, x_k \in V$ deux à deux orthogonaux. On a donc

$$\langle x_i|x_j\rangle = 0 \text{ pour tout } i \neq j.$$

Par ailleurs, on a que

$$\|x_1 + \cdots + x_k\|^2 = \langle x_1 + \cdots + x_k | x_1 + \cdots + x_k\rangle = \sum_{i,j=1}^{k} \langle x_i|x_j\rangle.$$

Mais puisque $\langle x_i|x_j\rangle = 0$ pour tout $i \neq j$, on obtient

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

ce que l'on voulait démontrer. (On peut aussi faire une récurrence.)

Lemme 4.3.11. *Soit $(V, \langle\cdot|\cdot\rangle)$ un espace préhilbertien, et soient $x_1, \ldots, x_k \in V$ des vecteurs non nuls deux à deux orthogonaux. Alors $(x_1, \ldots, x_k)$ est une famille libre.*

Preuve. Soient $\lambda_1, \dots, \lambda_k \in \mathbb{R}$ tels que

$$\lambda_1 x_1 + \cdots + \lambda_k x_k = 0_V.$$

Soit $j \in \{1, \dots, k\}$. On a

$$\langle x_j | \lambda_1 x_1 + \cdots + \lambda_k x_k \rangle = \langle x_j | 0_V \rangle = 0,$$

et donc

$$\sum_{i=1}^{k} \lambda_i \langle x_j | x_i \rangle = 0.$$

Puisque les x_i sont deux à deux orthogonaux, cela s'écrit

$$\lambda_j \langle x_j | x_j \rangle = 0.$$

Puisque par hypothèse $x_j \neq 0$, on a $\langle x_j | x_j \rangle > 0$, et donc $\lambda_j = 0$. Ceci achève la démonstration.

Revenons maintenant à l'existence de bases orthonormées.

Proposition 4.3.12. *Soit $(V, \langle \cdot | \cdot \rangle)$ un espace préhilbertien de dimension finie. Alors V possède une base $(v_1, \dots, v_n)$ orthonormée pour le produit scalaire.*

De plus, si $(v_1, \dots, v_n)$ est une base orthonormée, alors pour tout $x \in V$, on a

$$x = \langle v_1 | x \rangle v_1 + \cdots + \langle v_n | x \rangle v_n.$$

Remarque 4.3.13. En dimension infinie, on parle d'espace de Hilbert lorsque les propriétés des bases orthonormées vues ici en dimension finie se généralisent (existence, décomposition de tout vecteur comme une somme infinie (ou série) par rapport à ce qui ressemble le plus dans ce contexte à une base orthonormée…). L'étude générale des espaces de Hilbert en dimension infinie dépasse le cadre de ce cours. La série de Fourier d'une fonction périodique de période T peut être vue comme l'écriture selon une base orthonormée infinie composée par les harmoniques des sinus et cosinus de période T.

Preuve. Pour montrer l'existence d'une base orthonormée, on peut au choix

- appliquer l'algorithme de Gram-Schmidt présenté à la section 4.4 à une base $(e_1, \dots, e_n)$ de V. D'un point de vue plus géométrique, si $n = 1$, on normalise le vecteur e_1. Sinon, on construit une base orthonormée de $V' = \mathrm{Vect}(e_1, \dots, e_{n-1})$, on retranche de e_n son projeté orthogonal sur V' et on normalise pour compléter la base orthonormée de V'. Par exemple, si on a une base (e_1, e_2, e_3) de $\mathbb{R}^3$,
  - on normalise e_1 en v_1
  - on se place dans le plan (e_1, e_2), on retranche à e_2 son projeté sur la droite vectorielle engendrée par e_1, on normalise, c'est v_2,
  - on projette e_3 sur le plan (e_1, e_2), on le retranche à e_3 et on normalise, c'est v_3.
- utiliser le théorème 3.4.9 sur les formes bilinéaires, il existe une base $\mathbf{e} = (e_1, \dots, e_n)$ de V qui est orthogonale pour le produit scalaire. Comme $\mathbf{e}$ est une base, $e_i \neq 0$ pour tout i, et on a donc $\|e_i\| \neq 0$. Pour tout $i = 1, \dots, n$, on pose
$$v_i = \frac{1}{\|e_i\|} e_i.$$
Il est clair que $(v_1, \dots, v_n)$ est une base de V.
De plus, on a :
$$\langle v_i | v_j \rangle = \langle \frac{1}{\|e_i\|} e_i | \frac{1}{\|e_j\|} e_j \rangle = \frac{1}{\|e_i\| \times \|e_j\|} \langle e_i | e_j \rangle \text{ pour tout } i, j.$$
Comme $\mathbf{e}$ est une base orthogonale, on obtient
$$\langle v_i | v_j \rangle = 0 \text{ pour tout } i \neq j.$$
De plus, pour tout i, on a
$$\langle v_i | v_i \rangle = \frac{1}{\|e_i\|^2} \langle e_i | e_i \rangle = \frac{\langle e_i | e_i \rangle}{\langle e_i | e_i \rangle} = 1.$$
Ainsi, $(v_1, \dots, v_n)$ est une base orthonormée.

Soit maintenant $(v_1,\dots,v_n)$ une base orthonormée, et soit $x \in V$. Comme $v_1,\dots,v_n$ est une base, on peut écrire

$$x = \lambda_1 v_1 + \cdots + \lambda_n v_n.$$

Pour tout j, on a alors

$$\langle v_j | x \rangle = \sum_{i=1}^{n} \lambda_i \langle v_j | v_i \rangle = \lambda_j,$$

la dernière égalité provenant du fait que $v_1,\dots,v_n$ est une base orthonormée. On a donc bien l'égalité annoncée.

Nous avons donc maintenant une notion satisfaisante de distance entre deux éléments d'un espace vectoriel muni d'un produit scalaire. Rappelons que la question qui a motivé ce travail est la suivante : je veux construire dans un espace vectoriel V un « bon approximant » w pour un élément v sous la contrainte que w doit être contenu dans un sous-espace W. Ceci est fourni par la projection orthogonale w de v sur W, qui est l'analogue des projections dans le plan ou dans l'espace.

On peut clarifier maintenant ce qu'on veut dire exactement par un « bon approximant » : on veut que la distance $d(v,w)$ entre v et w soit la plus petite possible. Le lemme suivant nous donne un critère numérique pour que $w \in W$ soit le « meilleur approximant » pour v.

Lemme 4.3.14. *Soit V un espace préhilbertien, W un sous espace de V et v un élément de V. Si $w \in W$ a la propriété que $\langle v-w | w' \rangle = 0$ pour tout $w' \in W$ alors pour tout $w' \in W$ on a que $d(v,w) \leqslant d(v,w')$, avec égalité si et seulement si $w' = w$.*

Autrement dit, si la droite qui relie v à $w \in W$ est perpendiculaire à W alors w est le point de W le plus proche de v. Ce résultat vous est familier lorsque $v \in \mathbb{R}^2$ et W est une droite dans $\mathbb{R}^2$, ou lorsque $v \in \mathbb{R}^3$ et W est un plan dans $\mathbb{R}^3$.

Preuve. On a que

$$d(v,w') = \|v-w'\| = \|(v-w)+(w-w')\|.$$

Maintenant, $w-w' \in W$ donc par hypothèse $(v-w) \perp (w-w')$ et par le théorème de Pythagore

$$d(v,w')^2 = \|(v-w)\|^2 + \|(w-w')\|^2 \geqslant d(v,w)^2$$

avec égalité si et seulement si $\|w-w'\| = 0$, c'est-à-dire $w = w'$.

Notre critère est que $(v-w)$ doit être orthogonal à tous les éléments de W. Étudions donc l'ensemble constitué de tels éléments.

Définition 4.3.15. Soit $(V, \langle\,|\,\rangle)$ un espace préhilbertien et soit S un sous-ensemble de V. L'orthogonal de S, noté $S^\perp$, est le sous-ensemble de V défini par

$$S^\perp = \{x \in V \mid \langle s | x \rangle = 0 \text{ pour tout } s \in S\}.$$

Exercice 4.3.16. Démontrer que $S^\perp$ est toujours un sous-espace vectoriel de W.

Théorème 4.3.17. *Soit $(V, \langle\cdot\,|\,\cdot\rangle)$ un espace préhilbertien et soit W un sous-espace vectoriel de V. Alors :*

1) *Pour tout $w \in W$ et tout $w' \in W^\perp$, on a $w \perp w'$. De plus, $W \cap W^\perp = \{0_V\}$.*
2) *Si W est de dimension finie, on a $V = W \oplus W^\perp$. Autrement dit, tout $x \in V$ s'écrit de manière unique sous la forme*

$$x = w + w', w \in W, w' \in W^\perp.$$

De plus, si $(v_1,\dots,v_k)$ est une base orthonormée pour W alors on a $w = \sum_{i=1}^{k} \langle v_i | x \rangle v_i$.

Preuve.

1) Si $w \in W$ et $w' \in W^\perp$, alors on a $\langle w | w' \rangle = 0$ par définition de $W^\perp$. On a donc $w \perp w'$. Soit maintenant $w \in W \cap W^\perp$. Puisque $w \in W^\perp$ et $w \in W$ on a que $\langle w | w \rangle = 0$ et donc $w = 0$ d'après les propriétés du produit scalaire.
   Ainsi, on a $W \cap W^\perp = \{0\}$, ce qu'il fallait vérifier.

2) D'après (1), il reste à démontrer que $V = W + W^\perp$, c'est-à-dire que tout vecteur $v \in V$ peut s'écrire $v = w + w'$ avec $w \in W$ et $w' \in W^\perp$.
Si $W = \{0\}$, on a $W^\perp = V$, et il n'y a rien à faire. On peut donc supposer que W n'est pas l'espace trivial $\{0_V\}$. La restriction à W du produit scalaire sur V est encore un produit scalaire. Puisque W est de dimension finie, W possède une base orthonormée $(v_1, \dots, v_k)$ d'après la proposition précédente.
Soit $v \in V$. On pose

$$w = \sum_{i=1}^{k} \langle v_i | v \rangle v_i.$$

Alors $w \in W = \mathrm{Vect}(v_1, \dots, v_k)$. D'autre part, on a

$$\begin{aligned} \langle v_j | v - w \rangle &= \langle v_j | v \rangle - \langle v_j | w \rangle \\ &= \langle v_j | v \rangle - \langle v_j | \sum_{i=1}^{n} \langle v_i, v \rangle v_i \rangle \\ &= \langle v_j | v \rangle - \sum_{i=1}^{k} \langle v_i | v \rangle \langle v_j | v_i \rangle. \end{aligned}$$

Puisque $v_1, \dots, v_k$ est orthonormée, on en déduit :

$$\langle v_j | v - w \rangle = \langle v_j | v \rangle - \langle v_j | v \rangle = 0,$$

et ceci pour tout $j = 1, \dots, k$.
Soit $s \in W$. Alors on peut écrire $s = s_1 v_1 + \cdots + s_k v_k$, et donc

$$\langle s | v - w \rangle = \sum_{i=j}^{k} s_j \langle v_j | v - w \rangle = 0.$$

Ainsi, $v - w \in W^\perp$, et donc on a la décomposition voulue en posant $w' = v - w$. Si maintenant on a deux décompositions

$$v = w_1 + w_1' = w_2 + w_2', w_i \in W, w_i' \in W^\perp,$$

on a

$$w_1 - w_2 = w_2' - w_1' \in W \cap W^\perp,$$

car W et $W^\perp$ sont des sous-espaces vectoriels de V. Par le premier point, on en déduit $w_1 - w_2 = w_2' - w_1' = 0_V$, et donc $w_1 = w_2, w_1' = w_2'$.

Remarque 4.3.18. Le point (2) est faux sans hypothèse de finitude de la dimension de W.

D'après le deuxième point du théorème, lorsque W est de dimension finie, tout $x \in V$ se décompose de manière unique sous la forme

$$x = w + w', w \in W, w' \in W^\perp.$$

Cela conduit à la définition suivante :

Définition 4.3.19. Soit $(V, \langle \cdot | \cdot \rangle)$ un espace préhilbertien, et soit W un sous-espace de V de dimension finie. Pour tout $x = w + w' \in V$ avec $w \in W$ et $w' \in W^\perp$ on pose

$$p_W(x) = w.$$

Le vecteur $p_W(x) \in W$ est appelé la *projection orthogonale* de x sur W. Si $(v_1, \dots v_k)$ est une base orthonormée de W alors on a

$$p_W(x) = \sum_i \langle v_i | x \rangle v_i.$$

Le lecteur pourra vérifier à titre d'exercice les propriétés suivantes :

1) L'application $p_W : V \to V$ est linéaire.

2) Pour tout $x \in V$, on a $p_W(x) \in W$ et $(x - p_W(x)) \in W^\perp$ (en fait, $(x - p_W(x))$ est la projection orthogonale sur $W^\perp$).
3) Pour tout $x \in V$, on a $p_W(x) = x \iff x \in W$.
4) Pour tout $x \in V$, on a $p_W(x) = 0_V \iff x \in W^\perp$.
5) Pour tout $x \in V$, on a $p_W(p_W(x)) = p_W(x)$.

La projection orthogonale a la propriété essentielle suivante :

$$p_W(x) \text{ est le point de } W \text{ le plus proche de } x.$$

Si on dispose d'une base orthonormée $(v_1, \ldots, v_k)$ pour W, on a une formule explicite pour calculer une projection orthogonale :

$$p_W(x) = \sum_{i=1}^{k} \langle v_i | x \rangle v_i \tag{4.1}$$

Exemple 4.3.20. On reprend pour W l'exemple du plan P engendré par les vecteurs $u = (1,1,0)$ et $v = (1,0,-1)$. On a vu qu'une base orthonormée de W est donnée par

$$u_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, u_2 = \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix}$$

La projection orthogonale du vecteur v de composantes (x,y,z) est donc

$$\begin{aligned} p_W\left(\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right) &= \langle u_1 | v \rangle u_1 + \langle u_2 | v \rangle u_2 \\ &= \frac{1}{2} \left\langle \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \middle| \begin{pmatrix} x \\ y \\ z \end{pmatrix} \right\rangle \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \frac{1}{6} \left\langle \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix} \middle| \begin{pmatrix} x \\ y \\ z \end{pmatrix} \right\rangle \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix} \\ &= \frac{x+y}{2} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \frac{x-y-2z}{6} \begin{pmatrix} 1 \\ -1 \\ -2 \end{pmatrix} \\ &= \frac{1}{3} \begin{pmatrix} 2x+y-z \\ x+2y+z \\ -x+y+2z \end{pmatrix}. \end{aligned}$$

Reste à construire des bases orthonormées adaptées dans le cas général, c'est l'objet du prochain paragraphe.

4.4 Procédé d'orthonormalisation de Gram-Schmidt

Soit $(V, \langle \cdot | \cdot \rangle)$ un espace préhilbertien de dimension finie. On suppose donnée une base pour V, $\mathbf{e} = (e_1, \ldots, e_n)$. On présente un algorithme de construction d'une famille orthonormée $(v_1, \ldots v_k)$ à partir de $\mathbf{e}$ pour $k = 1$, puis $k = 2 \ldots$ puis $k = n$. Cette famille engendrera le même sous-espace vectoriel que la famille $(e_1, \ldots e_k)$.

1) Initialisation : pour $k = 1$, on pose $v_1 = \frac{e_1}{\|e_1\|}$. v_1 est alors de norme 1 par construction et l'espace engendré par (v_1) est égal à l'espace engendré par (e_1).
2) Début du corps de la boucle
Pour $k > 1$, on suppose $(v_1, \ldots, v_{k-1})$ déjà construits. On va construire v_k, il doit être orthogonal à l'espace W engendré par $(v_1, \ldots, v_{k-1})$.

3) Étape d'orthogonalisation
On a vu que pour tout vecteur z, en lui soustrayant $p_W(z)$ son projeté orthogonal sur un sous-espace vectoriel W, on obtient un vecteur $z - p_W(z)$ qui est orthogonal à W.
On définit donc un vecteur auxiliaire f_k en soustrayant de e_k son projeté orthogonal sur W, donc en appliquant (4.1) :
$$f_k = e_k - \sum_{j=1}^{k-1} \langle v_j | e_k \rangle v_j.$$
Par construction f_k est orthogonal aux vecteurs $v_1, \ldots, v_{k-1}$. Par contre, il n'est pas forcément de longueur 1.

4) Étape de normalisation
On observe que e_k n'est pas combinaison linéaire des v_j pour $j \leqslant k-1$ (en effet la famille $(v_1, \ldots, v_{k-1})$ engendre le même sous-espace que la famille $(e_1, \ldots, e_{k-1})$, or la famille $(e_1, \ldots, e_k)$ est libre). On a donc $f_k \neq 0$, on pose :
$$v_k = \frac{f_k}{\|f_k\|}.$$

5) Nous avons maintenant construit $(v_1, \ldots, v_k)$. On voit que la famille $(v_1, \ldots, v_k)$ engendre bien le même sous-espace vectoriel que $(e_1, \ldots, e_k)$ Si $k < n$, on revient au début de la boucle (étape 2) en incrémentant k de 1.

On a donc :

Proposition 4.4.1. *Les vecteurs de la famille* **v** *construite par le procédé de Gram-Schmidt ci-dessus forment une base orthonormée pour V et le sous-espace vectoriel engendré par* $(v_1, \ldots, v_k)$ *est le même que celui engendré par* $(e_1, \ldots, e_k)$

Exemple 4.4.2. On considère la base de $\mathbb{R}^3$
$$e_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, e_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, e_3 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

Appliquons le procédé de Gram-Schmidt à cette base afin d'obtenir une base orthonormée pour le produit scalaire.
On pose
$$v_1 = \frac{e_1}{\|e_1\|} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{pmatrix}$$
On a
$$f_2 = e_2 - \langle v_1 | e_2 \rangle v_1 = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ 1 \end{pmatrix}.$$
On pose
$$v_2 = \frac{f_2}{\|f_2\|} = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \end{pmatrix}.$$
Enfin
$$f_3 = e_3 - \langle v_1 | e_3 \rangle v_1 - \langle v_2 | e_3 \rangle v_2 = \begin{pmatrix} -2/3 \\ 2/3 \\ 2/3 \end{pmatrix},$$
et donc
$$v_3 = \frac{f_3}{\|f_3\|} = \frac{\sqrt{3}}{2} \begin{pmatrix} -2/3 \\ 2/3 \\ 2/3 \end{pmatrix}.$$

On a donc

$$v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\\0\end{pmatrix}, v_2 = \sqrt{\frac{2}{3}}\begin{pmatrix}1/2\\1/2\\1\end{pmatrix}, v_3 = \frac{\sqrt{3}}{2}\begin{pmatrix}-2/3\\2/3\\2/3\end{pmatrix}.$$

Exemple 4.4.3. Construisons une base orthonormée pour le plan d'équation $x+y+z=0$ dans $\mathbb{R}^3$. Il a une base non orthonormée (e_1, e_2) donnée par

$$e_1 = \begin{pmatrix}1\\-1\\0\end{pmatrix}, e_2 = \begin{pmatrix}1\\0\\-1\end{pmatrix}.$$

On pose $v_1 = \dfrac{e_1}{\|e_1\|} = \begin{pmatrix}1/\sqrt{2}\\-1/\sqrt{2}\\0\end{pmatrix}$. On introduit alors

$$f_2 = e_2 - \langle v_1|e_2\rangle v_1 = e_2 - \frac{1}{\sqrt{2}}v_1 = \begin{pmatrix}1/2\\1/2\\-1\end{pmatrix}$$

et on pose

$$v_2 = \frac{f_2}{\|f_2\|} = \begin{pmatrix}1/\sqrt{6}\\1/\sqrt{6}\\-2/\sqrt{6}\end{pmatrix}.$$

Ceci nous donne la base (v_1, v_2) avec

$$v_1 = \begin{pmatrix}1/\sqrt{2}\\-1/\sqrt{2}\\0\end{pmatrix}, v_2 = \begin{pmatrix}1/\sqrt{6}\\1/\sqrt{6}\\-2/\sqrt{6}\end{pmatrix}.$$

Exemple 4.4.4. Sur les polynômes de degré au plus 2, on définit le produit scalaire

$$\phi(P,Q) = P(-1)Q(-1) + P(0)Q(0) + P(1)Q(1)$$

C'est bien un produit scalaire, car $\phi(P,P) = 0$ entraine $P(-1) = P(0) = P(1) = 0$ donc $P = 0$ (3 racines pour degré au plus 2). On peut orthonormaliser la base canonique $\{1, X, X^2\}$. On normalise le premier vecteur de la base en $v_1 = 1/\sqrt{3}$ car $\phi(1,1) = 3$. Le deuxième vecteur de la base est orthogonal au premier car

$$\phi(1,X) = -1 + 0 + 1 = 0.$$

Il suffit de le normaliser en $v_2 = X/\sqrt{2}$ ($\phi(X,X) = (-1)^2 + 0^2 + 1^2 = 2$). On projette X^2 sur le plan $\{v_1, v_2\}$

$$p(X^2) = \phi(v_1, X^2)v_1 + \phi(v_2, X^2)v_2 = \frac{1}{3}\phi(1, X^2) + \frac{1}{2}\phi(X, X^2)X = \frac{2}{3}.$$

Donc v_3 est $X^2 - 2/3$ normalisé, soit $v_3 = (X^2 - 2/3)/\sqrt{2/3}$ car

$$\phi(X^2 - 2/3, X^2 - 2/3) = (1/3)^2 + (-2/3)^2 + (1/3)^2 = 2/3.$$

Finalement, la base orthonormée obtenue est

$$\left\{\frac{1}{\sqrt{3}}, \frac{X}{\sqrt{2}}, \frac{X^2 - \frac{2}{3}}{\sqrt{\frac{2}{3}}}\right\}.$$

Remarque 4.4.5. En calcul exact ou à la main, il peut être plus simple de ne pas normaliser les vecteurs f_k à chaque étape, donc de construire une base orthogonale :

$$f_k = e_k - \sum_{j=1}^{k-1} \frac{\langle f_j | e_k \rangle}{\|f_j\|^2} f_j$$

et de normaliser la base seulement à la fin.
♠ En calcul approché, cette méthode de calcul n'est pas adaptée en raison des erreurs d'arrondis. On utilise plutôt la factorisation QR d'une matrice, qui est la version matricielle de l'orthonormalisation. L'orthonormalisation se fait en utilisant des matrices de symétries (réflexions de Householder) ou de rotations (méthode de Givens).

Remarque 4.4.6. Le procédé de Gram-Schmidt permet de calculer la projection orthogonale de tout vecteur $x \in V$ sur un sous-espace W de dimension finie, en calculant une base orthonormée $(v_1, \ldots, v_k)$ de W à partir d'une base quelconque $e_1, \ldots, e_k$ de W (pour le produit scalaire sur W obtenu par restriction du produit scalaire sur W). On aura alors

$$p_W(x) = \sum_{j=1}^{k} \langle v_j | x \rangle v_j.$$

Rappelons que $p_W(x)$ est le meilleur approximant de x dans W.

4.5 Exemples de problèmes de minimisation

4.5.1 Projection sur un plan de l'espace

Utilisons cette méthode pour construire pour tout $v \in \mathbb{R}^3$ le point le plus proche de v dans W, le plan d'équation $x+y+z=0$.

Nous avons vu qu'une base orthonormée pour ce plan est donnée par $v_1 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{pmatrix}, v_2 = \begin{pmatrix} 1/\sqrt{6} \\ 1\sqrt{6} \\ -2/\sqrt{6} \end{pmatrix}$.

Soit $v = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$: on a donc

$$\begin{aligned} p_W(v) &= \langle v|v_1\rangle v_1 + \langle v|v_2\rangle \\ &= \frac{(x-y)}{\sqrt{2}} v_1 + \frac{(x+y-2z)}{\sqrt{6}} v_2 \\ &= \begin{pmatrix} (x-y)/2 \\ (-x+y)/2 \\ 0 \end{pmatrix} + \begin{pmatrix} (x+y-2z)/6 \\ (x+y-2z)/6 \\ -2x-2y+4z/6 \end{pmatrix} \\ &= \begin{pmatrix} (2x-y-z)/3 \\ (-x+2y-z)/3 \\ (-x-y+2z)/3 \end{pmatrix}. \end{aligned}$$

Autre méthode

Le vecteur $n(1,1,1)$ est un vecteur normal au plan W, on retire de v sa projection sur l'orthogonal de W donc

$$p_W(v) = v - \frac{\langle n, v \rangle}{\|n\|^2} n = \begin{pmatrix} x \\ y \\ z \end{pmatrix} - \frac{x+y+z}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{2x-y-z}{3} \\ \frac{-x+2y-z}{3} \\ \frac{-x-y+2z}{3} \end{pmatrix}.$$

4.5.2 Régression linéaire

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 (avec $n \geqslant 2$) $P_1 := (x_1, y_1), \ldots, P_n := (x_n, y_n)$ de $\mathbb{R}^2$ représentant par exemple le résultat de n expérimentations. On suppose que les x_i sont deux à deux distincts. Supposons que la théorie nous dise que y varie de façon affine en fonction de x. À cause des erreurs de manipulation, 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 $y = ax + b$ telle que les points théoriques $Q_1 := (x_1, ax_1 + b), \ldots, Q_n := (x_n, ax_n + b)$ soient le plus proche possible des points expérimentaux $P_1, \ldots, P_n$?

Plus précisément, comment choisir la droite $y = ax + b$ telle que l'erreur quadratique

$$e := P_1Q_1^2 + \cdots + P_nQ_n^2$$

soit minimale ?

On veut donc trouver $(a, b) \in \mathbb{R}^2$ tels que

$$e := (y_1 - (ax_1 + b))^2 + \cdots + (y_n - (ax_n + b))^2$$

soit minimale. Posons

$$\underline{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \underline{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ et } \underline{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

On a facilement que

$$\underline{Y} - (a\underline{X} + b\underline{1}) = \begin{pmatrix} y_1 - (ax_1 + b) \\ \vdots \\ y_n - (ax_n + b) \end{pmatrix},$$

et donc

$$e = \|\underline{Y} - (a\underline{X} + b\underline{1})\|^2,$$

où nous utilisons la norme associée au produit scalaire canonique sur $\mathbb{R}^n$. Soit W le sous-espace vectoriel dans $\mathbb{R}^n$ formé de tous les vecteurs de la forme $a\underline{X} + b\underline{1}$ lorsque (a, b) décrit $\mathbb{R}^2$. On veut donc minimiser $\|\underline{Y} - w\|$, lorsque w décrit W. D'après les propriétés de la projection orthogonale, le minimum est obtenu pour $w = p_W(\underline{Y})$.

On doit donc calculer $p_W(\underline{Y})$. Les coefficients a et b seront alors donnés par la relation

$$p_W(\underline{Y}) = a\underline{X} + b\underline{1}$$

car $(\underline{X}, \underline{1})$ est une base de W. Posons

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}, \overline{y} = \frac{y_1 + \cdots + y_n}{n}.$$

Appliquons l'algorithme de Gram-Schmidt à la base $e_1 = \underline{1}, e_2 = \underline{X}$ de W. On a $v_1 = \underline{1}/\|\underline{1}\| = \frac{1}{\sqrt{n}}\underline{1}$. On a aussi

$$f_2 = e_2 - \langle v_1 | e_2 \rangle v_1 = \underline{X} - \overline{x}\underline{1}$$

et $v_2 = f_2/\|f_2\|$. On a alors

$$\begin{aligned} p_W(\underline{Y}) &= \langle v_1 | \underline{Y} \rangle v_1 + \langle v_2 | \underline{Y} \rangle v_2 \\ &= \langle v_1 | \underline{Y} \rangle v_1 + \langle v_2 | \underline{Y} - \overline{y}\underline{1} \rangle v_2 \quad \text{car } \langle v_2 | \underline{1} \rangle = 0 \\ &= \overline{y}\underline{1} + \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} (\underline{X} - \overline{x}\underline{1}) \\ &= a\underline{X} + (\overline{y} - a\overline{x})\,\underline{1} \text{ où } a = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}. \end{aligned}$$

La droite a donc pour coefficient directeur le rapport entre la covariance des (x_i, y_i) et la variance des x_i et passe par le point de coordonnées moyenne des x, moyenne des y.

4.5.3 Résolution au sens des moindres carrés. ♠

On généralise l'exemple précédent, il s'agit de « résoudre » des systèmes linéaires $n \times m$ qui ont plus d'équations (n) que d'inconnues (m). Matriciellement, on considère l'équation d'inconnue v :

$$Av = b, \quad v \in \mathbb{R}^m, b \in \mathbb{R}^n, n > m$$

où A est une matrice « mince », avec moins de colonnes que de lignes.

Par exemple pour la régression linéaire, v a deux composantes : le coefficient directeur α de la droite cherchée et son ordonnée à l'origine β. On a donc $m = 2$, on essaie de faire passer une droite par n points $(x_1, y_1), \ldots, (x_n, y_n)$, Le système s'écrit

$$\begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

et n'a en général pas de solutions.

On peut alors chercher v qui minimise $\|Av - b\|^2$. Soit $\mathrm{Im}(A)$, le sous-espace vectoriel parcouru par Av pour $v \in \mathbb{R}^n$. Le problème revient à chercher la projection orthogonale de b sur $\mathrm{Im}(A)$. Pour cela, on pourrait chercher une base orthonormale de $\mathrm{Im}(A)$ comme précédemment. On peut aussi utiliser la propriété du projeté orthogonal Av de b sur $\mathrm{Im}(A)$,

$$\forall w, \quad \langle Av - b | Aw \rangle = 0.$$

Notons A^* la transposée de la matrice A (ou sa transconjuguée dans le cas complexe), on a :

$$\langle Av - b | Aw \rangle = \langle A^*(Av - b) | w \rangle,$$

donc

$$\forall w, \quad \langle A^*(Av - b) | w \rangle = 0$$

donc v est solution de

$$A^*(Av - b) = 0 \Leftrightarrow (A^*A)v = A^*b,$$

qui est un système de m équations à m inconnues. Par exemple pour la régression linéaire, on a un système 2×2 :

$$\begin{pmatrix} x_1 & \ldots & x_n \\ 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} x_1 & \ldots & x_n \\ 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ \ldots \\ y_n \end{pmatrix}$$

On peut faire le calcul du produit de matrice formellement :

$$\begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{pmatrix}$$

et vérifier qu'on retrouve la solution de la section précédente. En effet, la deuxième équation nous dit que la droite de régression passe par le point de coordonnées les moyennes ($\overline{x} = \frac{1}{n}\sum_i x_i, \overline{y} = \frac{1}{n}\sum_i y_i$), et l'opération $\frac{1}{n}L_1 - \frac{\overline{x}}{n}L_2$ élimine β et permet de trouver le coefficient directeur :

$$(\frac{1}{n}\sum_i x_i^2 - \overline{x}^2)\alpha = \frac{1}{n}\sum_i x_i y_i - \overline{xy}.$$

Exercice 4.5.1. Faire de même pour une régression avec 3 séries statistiques (donc une série dépendant des deux autres) $z_n = \alpha x_n + \beta y_n + \gamma$. Indication de solution : la matrice A s'obtient en mettant dans la première colonne les x_i, dans la deuxième colonne les y_i et dans la troisième colonne des 1.

4.5.4 Approcher une fonction continue par une fonction affine

Si on veut trouver la meilleure approximation d'une fonction continue $f : [a,b] \to \mathbb{R}$ par une fonction affine (dont le graphe est une droite : $y = \alpha x + \beta$), la méthode précédente ne marche plus, puisque l'on doit considérer une infinité de points.

On peut adapter la méthode en considérant un grand nombre de points sur le graphe de f, dont les abscisses sont régulièrement espacés, $P_1 = (x_1, f(x_1)), \dots, P_n = (x_n, f(x_n))$, avec $x_i = a + \frac{(b-a)i}{n}$, et déterminer la droite de meilleure approximation pour ces points. Bien sûr, plus n est grand, meilleure est l'approximation. L'entier n étant fixé, on doit donc minimiser

$$d := (f(x_1) - (\alpha x_1 + \beta))^2 + \cdots + (f(x_n) - (\alpha x_n + \beta))^2.$$

Ceci revient aussi à minimiser

$$S_n := \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - (\alpha x_i + \beta))^2, \text{ avec } x_i = a + \frac{(b-a)i}{n}.$$

On peut démontrer rigoureusement que S_n converge vers $\int_a^b (f(t) - (\alpha t + \beta))^2 \, \mathrm{d}t$ (en fait S_n peut s'interpréter comme une somme de Riemann). 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

$$\int_a^b (f(t) - (\alpha t + \beta))^2 \, \mathrm{d}t.$$

Ce genre d'intégrale s'interprète souvent comme l'énergie d'un système. Ainsi, le problème de minimisation précédent revient à demander de minimiser cette énergie. Il se trouve que cette intégrale peut être interprétée comme un carré scalaire pour les fonctions continues, et qu'on trouve donc directement la meilleure approximation en projetant sur le sous-espace vectoriel des applications affines, qui est de dimension 2. Ceci correspond à une notion de distance entre fonctions qui est l'intégrale du carré de la différence (donc la surface de l'espace entre les courbes, mais pondérée par un coefficient qui minimise les petites différences). Il existe d'autres façons de mesurer la distance entre deux fonctions, qui donneraient des notions de « meilleure approximation » différentes. Par exemple, on pourrait regarder la surface de l'espace entre les courbes, ou l'écart maximal entre les points du graphe de même abscice, etc. Les problèmes d'optimisation ne font pas appel aux mêmes méthodes selon les propriétés de chaque distance fournissant la mesure de ce qui est « mieux approché ». En fait, le problème est considérablement simplifié si la distance provient d'un produit scalaire, car une projection orthogonale donne la réponse.

Exemple 4.5.2. Considérons le problème de minimisation suivant : trouver $a, b \in \mathbb{R}$ qui minimise

$$\int_0^{\frac{\pi}{2}} (\cos(x) - a - bx)^2 \, \mathrm{d}x.$$

Soit V l'espace des fonctions continues sur $[0, \frac{\pi}{2}]$ muni de la forme bilinéaire

$$\langle f | g \rangle = \int_0^{\frac{\pi}{2}} f(x) g(x) \, \mathrm{d}x.$$

On vérifie que $\langle \cdot | \cdot \rangle$ est un produit scalaire sur V. Considérons maintenant le sous-espace W de V défini par

$$W = \mathrm{Vect}(1, x)^{2} = \{ f \in W \mid f : x \mapsto a + bx, a, b \in \mathbb{R} \}.$$

Le problème de minimisation se reformule alors ainsi :

$$\text{Trouver } g \in W \text{ tel que } \langle \cos - g | \cos - g \rangle \text{ soit minimal.}$$

2. Ces notations sont un peu bancales. Elles signifient l'espace engendré par la fonction constante égale à 1 et la fonction $x \mapsto x$. En toute rigueur, il faut éviter de confondre une fonction et l'expression d'une valeur de cette fonction. Nous nous permettrons cet abus de notation quand il n'y a pas d'ambiguïté comme ici.

Autrement dit, on cherche $g \in W$ tel que $\|\cos(x) - g(x)\|$ soit minimal. On connait la solution, c'est le projeté du cosinus sur W : $g = p_W(\cos(x))$. On cherche donc à calculer la projection orthogonale de $\cos(x)$ sur $W = \text{Vect}(1,x)$.

Appliquons le procédé de Gram-Schmidt à la base $e_1 = 1, e_2 = x$ de W :

$$v_1 = \frac{e_1}{\|e_1\|} = \sqrt{\frac{2}{\pi}},$$

$$f_2 = e_2 - \langle v_1|e_2\rangle v_1 = (x - \frac{\pi}{4}),$$

$$v_2 = \frac{x - \frac{\pi}{4}}{\|x - \frac{\pi}{4}\|}.$$

On a alors

$$g = p_W(\cos(x)) = \frac{\langle 1|\cos(x)\rangle}{\langle 1|1\rangle}1 + \frac{\langle x - \frac{\pi}{4}|\cos(x)\rangle}{\langle x - \frac{\pi}{4}|x - \frac{\pi}{4}\rangle}(x - \frac{\pi}{4}) = ax + b.$$

le calcul donne $a = (\frac{24}{\pi^2} - \frac{96}{\pi^3})$ et $b = (\frac{-4}{\pi} + \frac{24}{\pi^2})$.

4.5.5 Projection sur les polynômes trigonométriques

La méthode qu'on vient de voir se généralise sans difficulté pour approcher une fonction $f : [a,b] \to \mathbb{R}$ par un type de fonction particulier, pas seulement une fonction affine. Par exemple, on peut vouloir approcher f par une fonction g appartenant à un sous-espace vectoriel W des fonctions continues sur $[a,b]$, de façon à ce que l'intégrale

$$\int_a^b (f(t) - g(t))^2 \, \mathrm{d}t$$

soit minimale lorsque g décrit W.

Considérons le problème posé dans l'introduction, celui d'approcher une fonction par des sommes trigonométriques. Soit $f : [-L,L] \to \mathbb{R}$ une fonction que l'on supposera continue : on veut approcher f par une somme finie de fonctions trigonométriques

$$S_n(f) := a_0 + \sum_{k=1}^{n} a_k \cos\left(\frac{2k\pi x}{L}\right) + b_k \sin\left(\frac{2k\pi x}{L}\right).$$

On veut trouver les coefficients a_k et b_k tels que l'intégrale

$$\int_{-L}^{L} (f(t) - S_n(f)(t))^2 \, \mathrm{d}t$$

soit minimale.

Soit V l'espace vectoriel des fonctions continues sur $[-L,L]$ à valeurs réelles $\mathscr{C}^0([-L,L],\mathbb{R})$ et W le sous-espace vectoriel de V engendré par

$$1,\ \cos\left(\frac{2k\pi x}{L}\right),\ \sin\left(\frac{2k\pi x}{L}\right),\ k = 1,\ldots,n.$$

Autrement dit, W est l'ensemble de fonctions de la forme

$$g(x) = a_0 + \sum_{k=1}^{n} a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \sin\left(\frac{k\pi x}{L}\right).$$

Considérons le produit scalaire sur V

$$\langle f|g\rangle = \int_{-L}^{L} f(t)g(t) \, \mathrm{d}t.$$

Le raisonnement précédent montre que la meilleure approximation $S_n(f)$ est donnée par $p_W(f)$. Or, on peut vérifier que

$$\frac{1}{\sqrt{2L}}, \sqrt{\frac{1}{L}}\cos\left(\frac{2k\pi x}{L}\right), \sqrt{\frac{1}{L}}\sin\left(\frac{2k\pi x}{T}\right),\ k=1,\dots,n$$

fournit une base orthonormée de W — nous reviendrons en détail sur ce calcul dans le dernier chapitre.

La formule pour la projection orthogonale $p_W(f)$ nous donne alors

$$\begin{aligned} p_W(f) &= \langle 1|f\rangle \frac{1}{2L} + \sum_{k=1}^{n} \frac{1}{L}\langle \cos\left(\frac{k\pi x}{L}\right)|f\rangle \cos\left(\frac{k\pi x}{L}\right) + \frac{1}{L}\langle \sin\left(\frac{k\pi x}{L}\right)|f\rangle \sin\left(\frac{k\pi x}{L}\right) \\ &= \frac{1}{2L}\int_{-L}^{L} f(t)dt + \frac{1}{L}\int_{-L}^{L} f(t)\cos\left(\frac{k\pi t}{L}\right)\mathrm{d}t \cos\left(\frac{k\pi x}{L}\right) + \frac{1}{L}\int_{-L}^{L} f(t)\sin\left(\frac{k\pi t}{L}\right)\mathrm{d}t \sin\left(\frac{k\pi x}{L}\right). \end{aligned}$$

Les choix de coefficients a_0, a_k, b_k qui minimisent cette intégrale sont donc donnés par

$$\begin{aligned} a_0 &= \frac{1}{2L}\int_{-L}^{L} f(t)dt, \\ a_k &= \frac{1}{L}\int_{-L}^{L} f(t)\cos\left(\frac{k\pi t}{L}\right)\mathrm{d}t, \\ b_k &= \frac{1}{L}\int_{-L}^{L} f(t)\sin\left(\frac{k\pi t}{L}\right)\mathrm{d}t. \end{aligned}$$

4.6 Diagonalisation orthogonale des matrices symétriques

Nous présentons ici un théorème sur la diagonalisation des matrices symétriques. On commence par un lemme.

Lemme 4.6.1. *Soit $(V, \langle\cdot|\cdot\rangle)$ un espace préhilbertien de dimension n, et soit $\mathbf{e} = (e_1,\dots,e_n)$ une base orthonormée. Soit $\mathbf{v} = (v_1,\dots,v_n)$ une autre base de V, et soit P la matrice de passage correspondante*[3]*. La base $(v_1,\dots,v_n)$ est orthonormée si et seulement si ${}^tPP = I_n$, c'est-à-dire si $P^{-1} = {}^tP$.*

Preuve. Soient M et N les matrices de la forme $\langle\cdot|\cdot\rangle$ dans les bases $\mathbf{e}$ et $\mathbf{v}$. On sait que $N = {}^tPMP$: puisque $\mathbf{e}$ est supposée orthonormée nous avons $M = I_n$ et $N = {}^tPP$. La base $\mathbf{v}$ est orthonormée si et seulement si $N = I_n$ c'est-à-dire si et seulement si

$$I_n = {}^tPP.$$

Théorème 4.6.2. *Soit $M \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique, c'est-à-dire vérifiant ${}^tM = M$. Alors il existe une base de $\mathbb{R}^n$ formée de vecteurs propres de M qui est orthonormée pour le produit scalaire usuel sur $\mathbb{R}^n$.*

Ce théorème important signifie que toute matrice symétrique réelle est diagonalisable, et qu'on peut trouver une base de vecteurs propres qui est orthonormée. La démonstration repose sur la remarque suivante.

Lemme 4.6.3. *Soit M une matrice carrée $n \times n$ et $\underline{X}, \underline{Y} \in \mathbb{R}^n$. On a*

$$\langle \underline{X}|M\underline{Y}\rangle = \langle {}^tM\underline{X}|\underline{Y}\rangle.$$

où $\langle\cdot|\cdot\rangle$ est le produit scalaire canonique.

En particulier, si M est symétrique, $\langle \underline{X}|M\underline{Y}\rangle = \langle M\underline{X}|\underline{Y}\rangle$.

Preuve. On calcule :

$$\langle \underline{X}|M\underline{Y}\rangle = {}^t\underline{X}M\underline{Y} = {}^t\underline{X}\,{}^t({}^tM)\underline{Y} = {}^t({}^tM\underline{X})\underline{Y} = \langle {}^tM\underline{X}|\underline{Y}\rangle.$$

3. C'est-à-dire la matrice dont les colonnes sont les vecteurs coordonnées de $(v_1,\dots,v_n)$ dans la base $(e_1,\dots,e_n)$.

Lemme 4.6.4. *Soit M une matrice symétrique réelle. Toutes ses valeurs propres (les racines complexes du polynôme caractéristique) sont réelles, et les espaces propres associés à des valeurs propres distinctes sont deux à deux orthogonaux.*

Preuve. Observons tout d'abord qu'un polynôme est toujours scindé sur $\mathbb{C}$, donc le polynôme caractéristique de M a n valeurs propres (pas nécessairement distinctes). Soit λ une valeur propre de M et $x \neq 0$ un vecteur propre associé à λ (si λ n'est pas réelle, les coordonnées de x non plus, puisque M est à coefficients réels). On a $Mx = \lambda x$ d'où $\overline{Mx} = \overline{\lambda x}$ donc $\bar{x}$ est vecteur propre de $\overline{M} = M$ associé à $\overline{\lambda}$ et

$$\lambda\langle\bar{x}|x\rangle = \langle\bar{x}|Mx\rangle = \langle M\bar{x}|x\rangle = \overline{\lambda}\langle\bar{x}|x\rangle$$

donc $\lambda = \overline{\lambda}$ car $\langle\bar{x}|x\rangle = \sum_{i=1}^{n} |x_i|^2 \neq 0$ puisque $x \neq 0$. On a bien que toutes les valeurs propres sont réelles.

Soient maintenant λ et μ deux valeurs propres distinctes de M, $\underline{X}$ un vecteur propre de M associé à λ ($M\underline{X} = \lambda\underline{X}$), et $\underline{Y}$ un vecteur propre de M associé à μ ($M\underline{Y} = \mu\underline{Y}$). Alors

$$\lambda\langle\underline{X}|\underline{Y}\rangle = \langle M\underline{X}|\underline{Y}\rangle = \langle\underline{X}|M\underline{Y}\rangle = \mu\langle\underline{X}|\underline{Y}\rangle$$

Donc si $\lambda \neq \mu$ alors $\underline{X}$ et $\underline{Y}$ doivent être orthogonaux.

Preuve du Théorème 4.6.2. Soit M une matrice symétrique réelle $n \times n$. On note $\lambda_1, \ldots, \lambda_k$ les valeurs propres distinctes de M et E_{λ_i} le sous-espace propre associé à λ_i. Nous avons alors que E_{λ_i} est orthogonal à E_{λ_j} si $i \neq j$: la somme des sous-espaces propres est une somme directe orthogonale.

Soit $E = E_{\lambda_1} \overset{\perp}{\oplus} \ldots \overset{\perp}{\oplus} E_{\lambda_k}$ et F son orthogonal dans $\mathbb{R}^n$. Il s'agit de montrer que $F = \{0\}$.

E est stable par M. Remarquons que F est aussi stable par M. En effet, si $x \in F$ et $y \in E$, alors $\langle Mx|y\rangle = \langle x|My\rangle = 0$ puisque $My \in E$ et $x \in F = E^{\perp}$.

Supposons que $F \neq \{0\}$. En prenant une base adaptée à la décomposition en sous-espaces $E \overset{\perp}{\oplus} F$ on voit que M est semblable à une matrice diagonale par blocs

$$\begin{pmatrix} M_E & 0 \\ 0 & M_F \end{pmatrix}$$

où M_E (resp. M_F) représente l'endomorphisme défini par M restreint à E (resp. F). Mais alors, le polynôme caractéristique de M est le produit des polynômes caractéristiques de M_E et M_F. Donc le polynôme caractéristique de M_F est scindé à racines réelles, et ses racines sont des valeurs propres de M. Il existerait donc dans F un vecteur propre pour un des λ_i, ce qui est impossible puisque $F \cap E = \{0\}$ et tous les vecteurs propres sont déjà dans E par construction. On a donc $\mathbb{R}^n = E$.

Pour produire une base orthonormée de diagonalisation de M, il suffit de choisir dans chaque E_{λ_i} une base orthonormée $\mathbf{e_i}$ et de considérer la concaténation $\mathbf{e} = (\mathbf{e}_1, \ldots, \mathbf{e}_k)$. Par le Lemme 3.4.6, $\mathbf{e}$ est une base orthonormée pour $\mathbb{R}^n$ composée de vecteurs propres de M.

Ceci se traduit en termes de formes bilinéaires de la façon suivante :

Théorème 4.6.5. *Soit $(V, \langle\cdot|\cdot\rangle)$ un espace préhilbertien de dimension finie, et soit $\varphi : V \times V \to \mathbb{R}$ une forme bilinéaire symétrique. Alors il existe une base orthonormée pour $\langle\cdot|\cdot\rangle$ qui est aussi φ-orthogonale.*

Preuve. Soit $\mathbf{e} = (e_1, \ldots, e_n)$ orthonormée pour $\langle\cdot|\cdot\rangle$, et soit B la matrice de φ dans cette base. Alors B est une matrice symétrique d'après le Lemme 3.3.3. D'après le théorème précédent, il existe une base $(\underline{V}_1, \ldots, \underline{V}_n)$ de $\mathbb{R}^n$ formée de vecteurs propres de B qui est orthonormée pour le produit scalaire usuel de $\mathbb{R}^n$.

Si $\underline{V}_j = \begin{pmatrix} v_{1j} \\ \vdots \\ v_{nj} \end{pmatrix}$, posons $v_j = \sum_{i=1}^{n} v_{ij} e_i$, de telle façon que $\underline{V}_j$ est le vecteur de coordonnées de v_j dans la base $\mathbf{e}$. Nous allons montrer que $\mathbf{v} = (v_1, \ldots, v_n)$ est une base de V qui possède les propriétés voulues.

Comme $\mathbf{e}$ est orthonormée, on a

$$\langle v_i|v_j\rangle = {}^t\underline{V}_i\underline{V}_j$$

d'après le Lemme 3.4.8. Comme $(\underline{V}_1,\dots,\underline{V}_n)$ est orthonormée, on en déduit que

$$\langle v_i|v_j\rangle = \begin{cases} 0 \text{ si } i \neq j \\ 1 \text{ si } i = j \end{cases}$$

Il reste à voir que $\mathbf{v}$ est φ-orthogonale. Soit P la matrice de passage de $\mathbf{v}$ à $\mathbf{e}$. La matrice N qui représente φ dans la base $\mathbf{v}$ est donc ${}^tPBP$. Or $\mathbf{v}$ étant orthonormée, on a ${}^tPP = I_n$. On a ainsi $N = P^{-1}BP$.

Mais $\mathbf{v}$ étant formée de vecteurs propres de B, nous avons que $P^{-1}BP$ est diagonale. N est donc diagonale, ce qui revient à dire que $\mathbf{v}$ est φ-orthogonale.

Remarque 4.6.6. Cette démonstration permet de voir que nous pouvons construire une telle base, orthonormée et φ-orthogonale, en diagonalisant B.

Méthode pratique pour trouver une base de vecteurs, orthonormée et φ-orthogonale

1) Pour diagonaliser une matrice symétrique réelle M dans une base orthonormée de $\mathbb{R}^n$.
   - — Pour chaque valeur propre $\lambda \in \mathbb{R}$ de M, on calcule une base de $E_\lambda = \mathrm{Ker}(M - \lambda I_n)$. Si λ est une valeur propre simple, on normalise le vecteur propre de la base, si λ est une valeur propre multiple, on applique l'algorithme de Gram-Schmidt pour obtenir une base orthonormée de E_λ.
   - — On recolle les bases orthonormées précédentes pour obtenir une base $(\underline{V}_1,\dots,\underline{V}_n)$ de $\mathbb{R}^n$ formée de vecteurs propres de M, orthonormée *pour le produit scalaire usuel* sur $\mathbb{R}^n$.
2) Pour trouver une base d'un espace préhilbertien de dimension finie V muni du produit scalaire $\langle\cdot|\cdot\rangle$, qui soit à la fois orthonormée pour $\langle\cdot|\cdot\rangle$ et orthogonale pour une forme bilinéaire symétrique $\varphi : V \times V \to \mathbb{R}$.
   On choisit une base $\mathbf{e}$ de V *orthonormée* pour $\langle\cdot|\cdot\rangle$. Soit M la matrice de φ dans la base $\mathbf{e}$. M est une matrice symétrique. On applique la méthode ci-dessus pour obtenir une base $(\underline{V}_1,\dots,\underline{V}_n)$ orthonormée de $\mathbb{R}^n$ formée de vecteurs propres de M. On prend v_i, l'unique vecteur dans V qui admet pour coordonnées dans la base $\mathbf{e}$ le vecteur $\underline{V}_i$. La base $\mathbf{v} = (v_1,\dots,v_n)$ est alors la base recherchée.

Remarques 4.6.7. Dans ce qui précède, le fait que $\mathbf{e}$ soit orthonormée se traduit par le fait que, en coordonnées dans $\mathbf{e}$, le produit scalaire $\langle\cdot|\cdot\rangle$ dans V se calcule comme le produit scalaire usuel de $\mathbb{R}^n$. Travailler dans $(V,\langle\cdot|\cdot\rangle)$ muni de la base $\mathbf{e}$ équivaut à travailler dans $\mathbb{R}^n$ muni du produit scalaire usuel.

On peut remarquer que la matrice de passage P de $\mathbf{e}$ vers $\mathbf{v}$ transforme une base orthonomée en une autre base orthonomée, donc que c'est une matrice orthogonale (${}^tP = P^{-1}$). C'est la seule situation où diagonaliser la matrice M au sens habituel (trouver une matrice diagonale semblable à M) ou en tant que matrice de forme bilinéaire (matrice dans une base orthogonale, donc dont tous les coefficients non diagonaux sont nuls) reviennent au même, parce que les formules de changement de base coïncident : $N = {}^tPMP = P^{-1}MP$.

Exemples 4.6.8.

1) Soit $B = \begin{pmatrix} 3 & 4 \\ 4 & -3 \end{pmatrix}$.
   On vérifie que les valeurs propres sont 5 et -5, et que

$$E_5 = \mathrm{Vect}\left\{\begin{pmatrix}2\\1\end{pmatrix}\right\}, E_{-5} = \mathrm{Vect}\left\{\begin{pmatrix}1\\-2\end{pmatrix}\right\}.$$

   Une base orthonormée pour E_5 est donc

$$\frac{1}{\sqrt{5}}\begin{pmatrix}2\\1\end{pmatrix},$$

   et une base orthonormée pour E_{-5} est donc

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

La base recherchée est donc donnée par

$$\left(\frac{1}{\sqrt{5}}\begin{pmatrix}2\\1\end{pmatrix}, \frac{1}{\sqrt{5}}\begin{pmatrix}1\\-2\end{pmatrix}\right).$$

2) Munissons $\mathbb{R}^3$ de son produit scalaire usuel, et soit

$$\varphi\left(\begin{pmatrix}x_1\\\vdots\\x_3\end{pmatrix}, \begin{pmatrix}y_1\\\vdots\\y_3\end{pmatrix}\right) = \sum_{i,j\leqslant 3} x_i y_j = (x_1+x_2+x_3)(y_1+y_2+y_3)$$

Soit $\mathbf{e}$ la base canonique de $\mathbb{R}^3$. C'est une base orthonormée pour le produit scalaire usuel. La matrice M de φ dans la base canonique est alors

$$M = \begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix}.$$

On vérifie que les valeurs propres sont 3 et 0, que E_3 est une droite engendrée par le vecteur $\begin{pmatrix}1\\1\\1\end{pmatrix}$ et que E_0 admet comme base la famille

$$\left(\begin{pmatrix}1\\-1\\0\end{pmatrix}, \begin{pmatrix}1\\0\\-1\end{pmatrix}\right).$$

Une base orthonormée pour E_1 est donc

$$\frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}.$$

Pour trouver une base orthonormée de E_0, on applique Gram-Schmidt. On pose

$$v_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}$$

Ensuite on pose

$$f_2 = \begin{pmatrix}1\\0\\-1\end{pmatrix} - \frac{1}{2}\begin{pmatrix}1\\-1\\0\end{pmatrix} = \begin{pmatrix}1/2\\1/2\\-1\end{pmatrix}.$$

Enfin on pose

$$v_2 = \frac{f_2}{\|f_2\|} = \sqrt{\frac{2}{3}}\begin{pmatrix}1/2\\1/2\\-1\end{pmatrix}$$

Une base orthonormée pour E_0 est donc

$$\left(\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}, \sqrt{\frac{2}{3}}\begin{pmatrix}\frac{1}{2}\\\frac{1}{2}\\-1\end{pmatrix}\right).$$

La base recherchée est donc donnée par

$$v_1 = \frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}, v_2 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}, v_3 = \sqrt{\frac{2}{3}}\begin{pmatrix}\frac{1}{2}\\\frac{1}{2}\\-1\end{pmatrix}.$$

Si $x = x'_1 v_1 + x'_2 v_2 + x'_3 v_3$ et $y = xy'_1 v_1 + y'_2 v_2 + y'_3 v_3$, on a

$$b(x,y) = 3x'_3 y'_3.$$

4.7 Matrices orthogonales

Soit M une matrice, on note M^* sa transposée si elle est réelle (ou sa transconjuguée, i.e. transposée conjuguée si elle est complexe). Nous avons vu ci-dessus que les matrices réelles M telles que $M^*M = I_n$ sont très importantes puisqu'elle représentent des changements de bases orthonormées.

Proposition 4.7.1. *Soit M une matrice de taille $n \times n$. Les conditions suivantes sont équivalentes :*

1) *$M^*M = I_n$;*
2) *pour tous v, w nous avons $\langle Mv|Mw\rangle = \langle v|w\rangle$, où $\langle \cdot | \cdot \rangle$ est le produit scalaire canonique.*
3) *pour tout v nous avons $\|Mv\| = \|v\|$, où $\|v\|$ est la norme de v pour le produit scalaire canonique. On parle d'*isométrie.

On dit qu'une matrice réelle qui satisfait à ces conditions est orthogonale *(*unitaire *pour une matrice complexe, en utilisant le produit scalaire hermitien canonique).*

Preuve. Si $M^*M = I_n$ alors

$$\langle Mv|Mw\rangle = (Mv)^*Mv = v^*M^*Mw = v^*I_nw = v^*w = \langle v|w\rangle.$$

Donc (1) implique (2). (2) implique (3) en prenant $v = w$ et (3) implique (2) par la formule de polarisation. Reste à montrer que (2) implique (1). Si $\langle Mv|Mw\rangle = \langle v|w\rangle$ alors pour tout v, w nous avons

$$v^*M^*Mw = v^*I_nw = v^*w = \langle v|w\rangle$$

et donc $M^*M = I_n$.

Nous finissons cette section avec une étude des matrices orthogonales de taille 2×2, i.e. des isométries de $\mathbb{R}^2$. Nous allons démontrer le théorème suivant :

Proposition 4.7.2. *Soit M une matrice 2×2 orthogonale. Alors l'application $\mathbb{R}^2 \mapsto \mathbb{R}^2$ donnée par $v \mapsto Mv$ est*

1) *une rotation autour de l'origine ou*
2) *une symétrie orthogonale par rapport à une droite passant par l'origine.*

Soit $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ une matrice orthogonale. On a alors

$$\left\| M \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\|^2 = a^2 + c^2 = 1.$$

Il existe donc un θ tel que $\begin{pmatrix} a \\ c \end{pmatrix} = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}$. De même $\begin{pmatrix} b \\ d \end{pmatrix} = \begin{pmatrix} \cos(\phi) \\ \sin(\phi) \end{pmatrix}$ et on peut écrire

$$M = \begin{pmatrix} \cos\theta & \cos\phi \\ \sin\theta & \sin\phi \end{pmatrix}.$$

On a alors

$$\begin{aligned} {}^tMM &= \begin{pmatrix} \cos^2\theta + \sin^2\theta & \cos\theta\cos\phi + \sin\theta\sin\phi \\ \cos\theta\cos\phi + \sin\theta\sin\phi & \cos^2\phi + \sin^2\phi \end{pmatrix} \\ &= \begin{pmatrix} 1 & \cos(\theta - \phi) \\ \cos(\theta - \phi) & 1 \end{pmatrix} \end{aligned}$$

et nous avons donc M orthogonale si et seulement si $\cos(\theta - \phi) = 0$, c'est-à-dire si et seulement si

$$\phi = \theta + \pi/2 \text{ ou } \phi = \theta - \pi/2.$$

Dans le premier cas nous avons

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

et on reconnait la matrice d'une rotation d'angle θ autour de l'origine. Dans le second cas on a

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

On calcule le polynôme caractéristique de M qui vaut $\lambda^2 - 1$, donc les valeurs propres sont 1 et -1. On trouve comme vecteurs propres de valeurs propres 1 et -1 respectivement :

$$e_1 = \begin{pmatrix} \cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix}, e_2 = \begin{pmatrix} -\sin\frac{\theta}{2} \\ \cos\frac{\theta}{2} \end{pmatrix}.$$

sont des vecteurs propres de M de valeur propre 1 et -1 respectivement.

Autrement dit, on a $Me_1 = e_1$ et $Me_2 = -e_2$. On vérifie que $\langle e_1 | e_2 \rangle = 0$, donc que les vecteurs e_1 et e_2 sont bien orthogonaux. M représente la symétrie orthogonale par rapport à la droite engendrée par e_1.

Chapitre 5

Séries numériques

Un paradoxe bien connu énonce que si l'on avance de pas à chaque fois réduits de moitié par rapport au précédent, on cumule une infinité de déplacements stricts vers l'avant, mais on n'ira jamais plus loin que si on avait fait un deuxième pas de la même longueur que le premier. Cela provient de la formule

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots=2$$

où le symbole « ... » se comprend comme « et ainsi de suite jusqu'à l'infini ». Quel sens donner à cette équation, et en particulier, quel sens donner aux points de suspension de son membre de gauche ? Ça ne peut pas signifier « le résultat qu'on obtient en effectuant une infinité d'additions » puisqu'il est impossible de faire une infinité d'additions.

La somme infinie à gauche doit être comprise comme une limite. En écrivant cette équation, nous disons la chose suivante :

En prenant n assez grand, nous pouvons rendre la somme finie (que l'on va appeler une somme partielle, puisqu'elle ne comporte qu'un nombre fini de termes)

$$1+\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^n}$$

aussi proche qu'on veut de 2.

La somme infinie

$$1+\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\cdots,$$

que l'on écrit aussi $\sum_{n=0}^{\infty}\frac{1}{2^n}$, doit être comprise comme la *limite de la suite des sommes partielles*.

5.1 Introduction aux séries

Définition 5.1.1. Soit $(u_n)_{n\in\mathbb{N}}$ une suite de nombres réels ou complexes.

La *série* de terme général u_n est en fait la même suite, mais considérée pour les propriétés de la somme de ses termes.

La *somme partielle* d'indice k

$$s_k = u_0+u_1+u_2+\cdots+u_k=\sum_{n=0}^{k}u_n$$

forme une suite $(s_k)_k$.

On dit que la série de terme général (u_n) est convergente si la suite (s_k) l'est, et dans ce cas on appelle *somme de la série* la limite de (s_k), et on la note $\sum_{n=0}^{\infty}u_n$.

On dit que la série diverge si la suite (s_k) n'a pas de limite finie.

Notation. Pour alléger l'écriture, et pour préciser de quelle limite on parle, on peut utiliser le formalisme suivant : si on parle de (u_n), cela signifie qu'on étudie les propriétés de la *suite* (u_n) ; si on parle de $\sum u_n$, cela signifie qu'on étudie les propriétés de la *série* de terme général (u_n), donc en fait de la suite (s_k). Dans cette notation, le $\sum$ n'a pas le sens d'une sommation, c'est juste un symbole pour dire qu'on s'intéresse aux propriétés de la série, et non de la suite.

Remarque 5.1.2. On peut évidemment accepter dans la définition des sommes qui commencent à un rang $m \neq 0$, par exemple $m = 1$ si u_0 n'est pas défini, ou si on ne veut calculer la somme qu'à partir d'un certain terme. Dans l'exposé qui suit, nous énonçons tous les résultats pour une série indexée à partir de $m = 0$ mais toutes les propriétés s'adaptent sans difficulté au cas $m \neq 0$.

Exemples 5.1.3.

1) Si on pose comme ci-dessus $u_n = \frac{1}{2^n}$ et on considère la série $\sum u_n$ alors la somme partielle s_k est donnée par la formule bien connue pour les sommes des termes d'une suite géométrique :
$$s_k = 1 + \frac{1}{2} + \cdots + \frac{1}{2^k} = \frac{1-(1/2)^{k+1}}{1-1/2} = 2 - 2^{-k}.$$

2) Si on pose $u_n = 1$ pour tout n et on considère la série $\sum u_n$ alors les sommes partielles $s_k = \sum_{n=0}^{k} u_n$ sont données pour tout k par
$$s_k = 1 + 1 + \cdots + 1 = k + 1.$$

3) Si on pose $u_n = (-1)^n$ et on considère la série $\sum u_n$ alors les sommes partielles $s_k = \sum_{n=0}^{k} u_n$ sont données par
$$s_0 = 1$$
$$s_1 = 1 - 1 = 0$$
$$s_2 = 1 - 1 + 1 = 1$$
et ainsi de suite, c'est-à-dire que, pour tout k pair, $s_k = 1$ et, pour tout k impair, $s_k = 0$.

4) Si on pose $u_n = \frac{1}{n^2}$ et on considère la série $(\sum_{n \geqslant 1} u_n)$ alors la somme partielle s_k est le nombre réel
$$s_k = 1 + \frac{1}{4} + \frac{1}{9} + \cdots + \frac{1}{k^2}.$$
Contrairement aux autres cas, nous ne disposons d'aucune formule générale pour cette somme partielle.

Attention ! Les deux notations
$$\sum u_n \quad \text{et} \quad \sum_{n=m}^{\infty} u_n,$$
qui sont très proches, désignent des choses différentes. La première est juste un raccourci pour dire « la série de terme général u_n » tandis que la seconde est la valeur de la somme de cette série (si elle converge). Il existe aussi dans certains textes une notation du genre $(\sum_{n \geqslant m} u_n)$ pour la suite de sommes partielles $(s_k)_{k \geqslant m}$ mais elle est à éviter car il est difficile de la distinguer de la valeur de la somme, et l'indice k dont elle dépend n'apparaît même pas dans l'expression, c'est donc une source de confusion. Dans ce qui suit, on utilisera systématiquement (s_k).

Passons aux premiers résultats.

Proposition 5.1.4. *Le terme général d'une série convergente tend vers 0.*

En effet, soit $\sum u_n$ une série et soit $(s_k)_{k \geqslant 0}$ la suite de ses sommes partielles. Si la série $\sum u_n$ est convergente, et sa somme est s, alors on a que $s_k \to s$ et $s_{k-1} \to s$ lorsque $k \to +\infty$, donc $s_k - s_{k-1} \to 0$. Or $s_k - s_{k-1} = u_k$ donc $u_k \to 0$.

Remarque 5.1.5. Par contraposition, si le terme général d'une suite ne tend pas vers 0 alors la série diverge. Par exemple, on peut voir sans calculer les sommes partielles que la série $\sum(-1)^n$ diverge parce que son terme général $(-1)^n$ ne tend pas vers 0. De même une suite géométrique de raison λ diverge lorsque $|\lambda| \geqslant 1$.

Attention ! La réciproque est fausse. Il existe des séries divergentes dont le terme général tend vers 0, par exemple on montre que $\sum \frac{1}{n+1}$ diverge alors que son terme général $\frac{1}{n}$ tend vers 0 (cf. proposition 5.2.10 plus bas).

Exemples 5.1.6.

1) Pour la série $\sum \frac{1}{2^n}$ nous avons que la somme partielle

$$s_k = 2 - 2^{-k} \to 2.$$

On peut donc écrire

$$\sum_{n=0}^{\infty} \frac{1}{2^n} = 2.$$

2) Plus généralement, soit λ un nombre réel ou complexe tel que $|\lambda| < 1$, et considérons la série $\sum \lambda^n$. La somme partielle est calculée par la même formule que le cas particulier précédent :

$$s_k = 1 + \lambda + \cdots + \lambda^k = \frac{1 - \lambda^{k+1}}{1 - \lambda}.$$

Puisque $|\lambda| < 1$ nous avons que $\lambda^k \to 0$ donc

$$s_k \to \frac{1}{1 - \lambda}.$$

Autrement dit, la série géométrique de raison $|\lambda| < 1$ converge et on a

$$\sum_{n=0}^{\infty} \lambda^n = \frac{1}{1 - \lambda}.$$

3) La série $\sum 1$ a pour sommes partielles $s_k = k + 1$. Cette suite n'est pas convergente : sa limite n'est pas finie. C'est donc une série *divergente* [1].

4) La série $\sum(-1)^n$ a pour sommes partielles

$$s_k = 1 \text{ si } k \text{ est pair, } s_k = 0 \text{ si } k \text{ est impair.}$$

Cette suite de sommes partielles, bien que bornée, diverge.

5) Même si nous ne disposons pas de formule pour les sommes partielles $s_k = \sum_{n=1}^{k} \frac{1}{n^2}$ il est possible de montrer que cette suite converge vers une limite finie. Nous verrons à la fin du semestre que

$$\lim_{k \to \infty} s_k = \frac{\pi^2}{6}$$

que nous pouvons aussi écrire

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

1. On aurait pu aussi remarquer que la suite u_n ne tend pas vers 0 et faire usage de la remarque 5.1.4.

5.2 Convergence des séries

La remarque suivante, qui se déduit des propriétés de linéarité pour la convergence des suites, est souvent utile dans l'étude des séries.

Proposition 5.2.1 (Linéarité de la convergence des séries)**.** *Soient $\sum u_n$ et $\sum v_n$ deux séries convergentes réelles ou complexes, de sommes u et v respectivement. Alors pour tout $\lambda, \mu \in \mathbb{C}$, la série*

$$\sum(\lambda u_n + \mu v_n)$$

est convergente, avec pour somme $\lambda u + \mu v$.

Le cas des séries réelles *à termes positifs* est assez simple.

Lemme 5.2.2. *Soit $\sum u_n$ une série réelle dont tous les termes u_n sont positifs. Pour tout $k \geqslant 0$, soit s_k la somme partielle*

$$s_k = \sum_{n=0}^{k} u_n.$$

Il y a deux possibilités

1) *la suite (s_k) converge vers une limite finie s. Autrement dit, la série $\sum u_n$ est convergente ;*
2) *la suite de sommes partielles (s_k) tend vers $+\infty$.*

En effet, $s_k - s_{k-1} = u_k \geqslant 0$ donc la suite s_k est croissante. Si elle est majorée, elle converge vers une limite finie (toute suite croissante majorée est convergente). Sinon, elle n'est pas majorée et tend donc vers $+\infty$. On peut remarquer que, dans le cas où la série converge, les sommes partielles sont toutes majorées par la somme de la série :

$$\sum_{n=0}^{k} u_n \leqslant \sum_{n=0}^{\infty} u_n.$$

On déduit de ce lemme une propriété très utile :

Proposition 5.2.3 (Critère de comparaison)**.** *Soient $\sum u_n$ et $\sum v_n$ deux séries à termes positifs. On suppose que, $u_n \leqslant v_n$ pour tout n.*

1) *Si la série $\sum u_n$ diverge, alors la série $\sum v_n$ diverge ;*
2) *si la série $\sum v_n$ converge, alors la série $\sum u_n$ converge.*

Preuve. Soit $(s_k)_{k\in\mathbb{N}}$ la suite de sommes partielles de la série $\sum u_n$. Soit $(t_k)_{k\in\mathbb{N}}$ la suite de sommes partielles de la série $\sum v_n$. La propriété découle d'une remarque assez basique : si $\forall n,\ u_n \leqslant v_n$, alors $\forall k,\ s_k \leqslant t_k$.

1) Si $\sum u_n$ diverge, $s_k \to +\infty$, ce qui implique $t_k \to +\infty$.
2) Si $\sum v_n$ converge, la suite (t_k) converge, donc est majorée. Or (s_k) est majorée par tout majorant de (t_k). Il s'ensuit que la série $\sum u_n$ converge.

Pour appliquer le lemme, il sera utile de se ramener à des séries à termes positifs. On peut d'abord observer que, s'il n'y a qu'un nombre fini de termes négatifs, on peut appliquer ce résultat, car la nature d'une série ne dépend pas de ses premiers termes (mais bien sûr la somme en dépend si la série est convergente). Dans la suite, quand on énonce un résultat pour une série à termes positifs, on pourra évidemment l'appliquer à une série dont les termes sont positifs *à partir d'un certain rang*. S'il y a un nombre infini de termes positifs et négatifs, on peut d'abord regarder la nature de la série des valeurs absolues du terme général.

Définition 5.2.4. Soit $\sum u_n$ une série. On dit que $\sum u_n$ est *absolument convergente* si la série $\sum |u_n|$ est convergente.

On a le résultat suivant.

Proposition 5.2.5. *Toute série absolument convergente est convergente.*

Idée de la preuve (hors programme) : cela résulte de l'inégalité triangulaire sur les sommes partielles

$$|\sum_{n=N}^{M} u_n| \leqslant \sum_{n=N}^{M} |u_n|$$

Comme $\sum |u_n|$ est convergente, le terme de droite peut être rendu aussi petit que l'on veut pourvu que l'on choisisse N assez grand. Cela permet d'établir rigoureusement la convergence de la suite des sommes partielles de u_n (c'est ce qu'on appelle une suite de Cauchy).
Attention : la réciproque de cette proposition est fausse : il existe des séries réelles convergentes qui ne sont pas absolument convergentes. Leur comportement est parfois surprenant — par exemple, en permutant les termes d'une telle série on peut la rendre divergente, ou la faire converger vers n'importe quel nombre réel. De plus ces séries convergent lentement, il faut calculer des sommes partielles à des rangs d'indice élevé pour avoir une valeur approchée de la somme. Les séries absolument convergentes sont donc plus sympathiques ! Mais on n'a pas toujours le choix (par exemple certaines séries de Fourier).

Remarque 5.2.6. Le comportement de la série de terme général $u_n = (-1)^n$, qui diverge sans tendre vers $+\infty$, n'est possible que parce que certains termes de cette série sont négatifs.

Le critère de d'Alembert traite le cas des séries qui se comparent facilement à des séries géométriques.

Proposition 5.2.7 (Critère de d'Alembert)**.** *Soit $\sum u_k$ une série telle que $\frac{|u_{k+1}|}{|u_k|} \underset{k\to\infty}{\longrightarrow} \lambda$. Si $\lambda < 1$ alors la série $\sum u_k$ est absolument convergente. Si $\lambda > 1$ alors la série $\sum u_k$ diverge.*

Preuve. Si $\lambda > 1$, le terme général de la série ne tend pas vers 0, donc elle diverge. Si $\lambda < 1$, on observe que $0 \leqslant \lambda < \frac{\lambda+1}{2} < 1$. Comme la suite $|u_{n+1}/u_n|$ converge vers $\lambda < \frac{\lambda+1}{2}$, il existe un rang N tel que

$$\forall n \geqslant N, \quad |u_{n+1}|/|u_n| \leqslant \frac{\lambda+1}{2}.$$

Donc pour tout $n \geqslant N$, on a :

$$|u_n| \leqslant C\left(\frac{\lambda+1}{2}\right)^{n-N}, \quad C = |u_N|$$

. Comme $\frac{\lambda+1}{2} < 1$ la série géométrique

$$\sum \left(\frac{\lambda+1}{2}\right)^{n-N}$$

converge (on ne considère que les termes $n \geqslant N$). En appliquant le critère de comparaison 5.2.3, comme $|u_m|$ est une série à termes positifs, on en déduit que la série $\sum |u_m|$ converge.

Les séries géométriques convergent assez rapidement. Mais toutes les séries ne convergent pas aussi rapidement, par exemple les séries de Fourier qui seront abordées en fin de cours. Pour déterminer leur nature, on commence par utiliser un critère plus fin, le critère des équivalents.

La proposition suivant est très pratique :

Proposition 5.2.8 (Critère d'équivalence)**.** *Soient $\sum u_n$, $\sum v_n$ deux séries à termes* positifs.
Si $u_n \underset{n\to\infty}{\sim} v_n$ alors la série $\sum u_n$ converge si et seulement si la série $\sum v_n$ converge.

Preuve. La propriété d'équivalence des *suites* u_n et v_n se traduit par

$$\forall \varepsilon > 0 \ \exists N \in \mathbb{N} \text{ t. q. } n \geqslant N \Rightarrow (1-\varepsilon)u_n \leqslant v_n \leqslant (1+\varepsilon)u_n.$$

(quand l'une des deux suites, p. ex. u_n, ne s'annule jamais à partir d'un certain rang, il suffit de vérifier $v_n/u_n \to 1$).

En posant $\varepsilon = 1/2$, on voit qu'il existe un $N \in \mathbb{N}$ tel que, pour tout $n \geqslant N$,

$$v_n \leqslant 2u_n; \quad u_n \leqslant 2v_n.$$

Nous avons donc par la proposition 5.2.3 que, chaque fois que l'une converge, l'autre aussi, et chaque fois que l'une diverge, l'autre aussi.

Remarque 5.2.9. Si la suite u_n est strictement positive à partir d'un certain rang, toute suite équivalente à u_n sera également strictement positive à partir d'un certain rang (pas nécessairement le même !). Il suffit donc de vérifier cette condition sur une seule des deux suites.

Les développements limités permettent de trouver des équivalents pour de nombreuses suites (de façon générale, le premier terme non nul d'un développement limité est un équivalent). Un équivalent permet souvent de se ramener à des séries plus simples. Il faut donc connaitre les propriétés de convergence d'un certain nombre de séries de référence. Nous avons déjà vu le cas des séries géométriques. En voici un autre :

Proposition 5.2.10 (Critère de Riemann)**.** *Pour tout nombre réel positif $s > 0$ la suite infinie*

$$\left(\sum_{n\geqslant 1} \frac{1}{n^s}\right)$$

diverge si $s \leqslant 1$ et converge si $s > 1$.

Remarque 5.2.11. Le cas $s \leqslant 0$ ne présente pas d'intérêt, puisque le terme de la série $\sum n^{-s}$ ne tend alors pas vers 0 (la série diverge *gorssièrement)*.

Preuve. On doit déterminer quand la suite de sommes partielles

$$s_k = \sum_{n=1}^{k} \frac{1}{n^s}$$

converge. Puisque la suite (u_n) est à termes positifs il suffit par le lemme 5.2.2 de savoir quand la suite s_k est majorée. Nous allons faire cela par une technique très puissante : comparaison d'une somme avec une intégrale. Il y a en effet un lien fort entre l'intégrale $\int_1^k f(x)\,\mathrm{d}x$ et la somme $\sum_{n=1}^k f(n)$.

Puisque $s > 0$, la fonction $f(x) = x^{-s}$ est (strictement) décroissante sur $\mathbb{R}_+^*$, donc pour tout x tel que $x \in [n, n+1]$, nous avons que

$$\frac{1}{n^s} \geqslant \frac{1}{x^s} \geqslant \frac{1}{(n+1)^s}.$$

Il s'ensuit que

$$\int_n^{n+1} \frac{1}{n^s}\,\mathrm{d}x \geqslant \int_n^{n+1} \frac{1}{x^s}\,\mathrm{d}x \geqslant \int_n^{n+1} \frac{1}{(n+1)^s}\,\mathrm{d}x.$$

c'est à dire que pour tout entier strictement positif n nous avons que

$$\frac{1}{n^s} \geqslant \int_n^{n+1} \frac{1}{x^s}\,\mathrm{d}x \geqslant \frac{1}{(n+1)^s}.$$

En sommant ces inégalités pour $n = 1, \ldots, k$, nous obtenons que

$$1 + \frac{1}{2^s} + \cdots + \frac{1}{k^s} \geqslant \int_1^2 \frac{1}{x^s}\,\mathrm{d}x + \int_2^3 \frac{1}{x^s}\,\mathrm{d}x + \cdots + \int_k^{k+1} \frac{1}{x^s}\,\mathrm{d}x \geqslant \frac{1}{2^s} + \frac{1}{3^s} + \cdots \frac{1}{(k+1)^s},$$

ce qui peut s'écrire

$$\sum_{n=1}^{k} \frac{1}{n^s} \geqslant \int_1^{k+1} \frac{1}{x^s}\,\mathrm{d}x \geqslant \sum_{n=2}^{k+1} \frac{1}{n^s},$$

c'est-à-dire

$$s_k \geqslant \int_1^{k+1} \frac{1}{x^s}\,\mathrm{d}x \geqslant s_{k+1} - 1.$$

En réorganisant ces équations, nous obtenons enfin

$$\int_1^k \frac{1}{x^s}\,\mathrm{d}x \leqslant s_k \leqslant \int_1^k \frac{1}{x^s}\,\mathrm{d}x + 1.$$

Pour calculer l'intégrale, il suffit de connaître une primitive de la fonction $f(x) = x^{-s}$ en fonction de s. Or, nous savons que

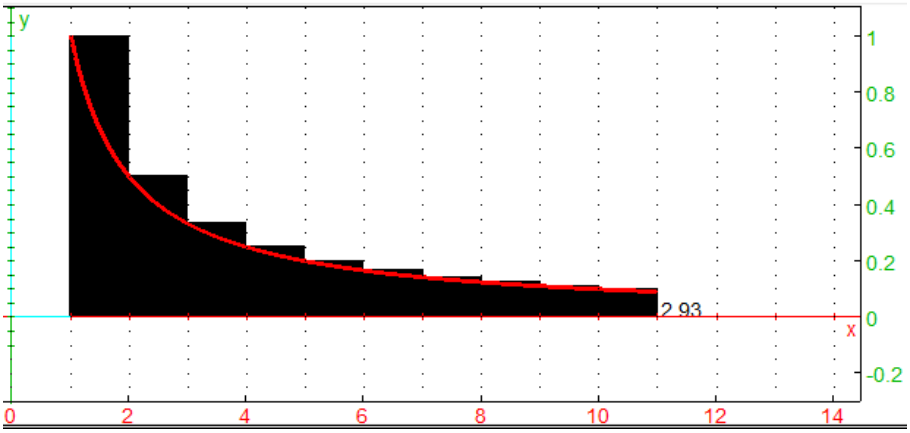


FIGURE 5.1 – Illustration graphique de la comparaison d'une série et de l'intégrale correspondante pour une fonction décroissante telle que $f(x) = 1/x$. En noir l'union des rectangles dont l'aire est une somme partielle de la série $\sum f(n)$ ($n \geqslant 1$), en rouge la courbe de f. On voit que l'aire sous la courbe est plus petite que la surface noire.

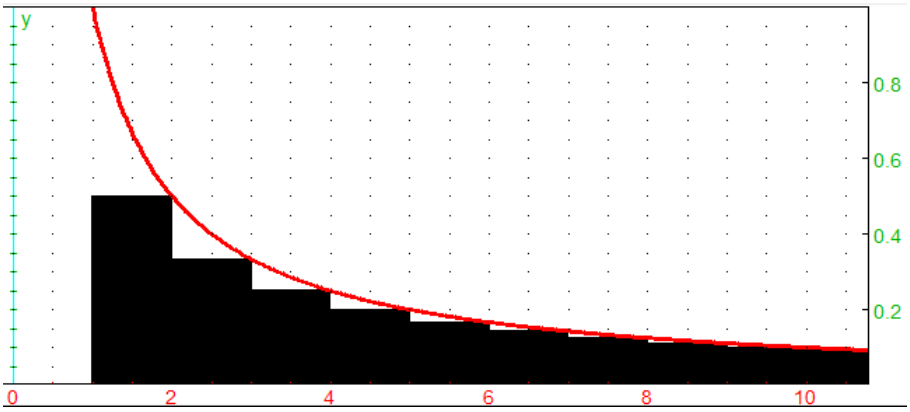


FIGURE 5.3 – En décalant les rectangles sur la gauche, on voit que l'aire sous la courbe est plus grande que la surface noire.

— si $s = 1$, une primitive de $f(x) = 1/x$ est le logarithme néperien $\ln x$.
— si $s \neq 1$, une primitive de $f(x) = 1/x$ est la fonction $F(x) = x^{1-s}/(1-s)$.

Nous allons maintenant regarder les différents cas, en fonction des valeurs de s.

1) Cas 1 : Si $s > 1$ (donc $1 - s < 0$), nous avons que

$$\int_1^k x^{-s}\,\mathrm{d}x = \left[\frac{x^{1-s}}{1-s}\right]_1^k = \frac{1-k^{1-s}}{s-1} \leqslant \frac{1}{s-1}.$$

On a donc que pour tout k

$$s_k \leqslant \frac{1}{s-1} + 1.$$

La suite s_k est donc majorée et la série $\sum \frac{1}{n^s}$ converge.

2) Cas 2 : Si $s = 1$, nous avons que

$$\int_1^k \frac{1}{x}\,\mathrm{d}x = [\ln(x)]_1^k = \ln(k).$$

Nous avons donc que

$$s_k \geqslant \ln(k) \underset{k\to\infty}{\longrightarrow} +\infty$$

donc la suite s_k n'est pas majorée et par le lemme 5.2.2 la série $(\sum \frac{1}{n}$ diverge.

3) Cas 3 : Si $s < 1$, on a pour tout entier positif non nul n que $\frac{1}{n^s} \geqslant \frac{1}{n} > 0$. D'après le cas précédent, $\sum \frac{1}{n}$ ne converge pas, donc il résulte du lemme 5.2.2 que $(\sum \frac{1}{n^s}$ diverge également.

Ceci termine démonstration de la proposition 5.2.10

Exemples 5.2.12. Les exemples qui suivent montrent à quel point, en combinant les propositions 5.2.8 et 5.2.10, on dispose d'un outil puissant pour déterminer si des séries convergent ou divergent.

1) Soit $u_n = \sin(\frac{1}{n})$ pour tout $n \geqslant 1$. Nous avons le développement limité au voisinage de 0 $\sin x = x + o(x)$ d'où l'on déduit

$$\sin\left(\frac{1}{n}\right) = \frac{1}{n} + o\left(\frac{1}{n}\right),$$

c'est-à-dire que

$$u_n \underset{n\to\infty}{\sim} \frac{1}{n}.$$

Puisque la série $\sum \frac{1}{n}$ diverge par la proposition 5.2.10, il résulte de la proposition 5.2.8 que $\sum u_n$ diverge aussi.

2) Soit $u_n = 1 - \cos(\frac{1}{n})$ pour tout $n \geqslant 1$. Nous savons que $u_n \underset{n\to\infty}{\sim} \frac{1}{2n^2}$. Puisque la série $\sum \frac{1}{n^2}$ converge par la proposition 5.2.10, il résulte de la proposition 5.2.8 que $(\sum_{n\geqslant 1} u_n)$ converge aussi.

3) Soit $u_n = \frac{1}{\sqrt{n}}(1 - \cos(\frac{1}{n}))$. Nous avons que

$$u_n \underset{n\to\infty}{\sim} \frac{1}{2n^2\sqrt{n}} = \frac{1}{2} n^{-\frac{5}{2}}.$$

Puisque la série $\sum \frac{1}{n^{5/2}}$ converge par la proposition 5.2.10, il résulte de la proposition 5.2.8 que $\sum u_n$ converge aussi.

4) Soit $u_n = \sin\left(\frac{1}{n}\right)\left(e^{\frac{1}{\sqrt{n}}} - 1\right)$. Par les développements limités, on a que

$$\sin(\frac{1}{n}) \underset{n\to\infty}{\sim} \frac{1}{n}$$

et

$$e^{\frac{1}{\sqrt{n}}} - 1 \underset{n\to\infty}{\sim} n^{-1/2}.$$

Il vient

$$u_n \underset{n\to\infty}{\sim} \frac{1}{n}\frac{1}{n^{1/2}} = n^{-3/2}.$$

Puisque la série $\sum \frac{1}{n^{3/2}}$ converge par la proposition 5.2.10, il résulte de la proposition 5.2.8 que $\sum u_n$ converge aussi.

Que se passe-t-il pour les séries qui ont un nombre infini de termes négatifs et positifs ? Si la série converge absolument, on a vu que la série convergeait. Sinon, il se peut que la série converge quand même. On peut montrer par exemple que $\sum \frac{(-1)^n}{n}$ converge alors que $\sum \frac{1}{n}$ diverge. Intuitivement, cela vient du fait qu'il peut y avoir des compensations entre des termes de signes opposés alors que, si le signe est constant, tous les termes s'ajoutent et la convergence dépend uniquement de la vitesse à laquelle le terme d'une série tend vers 0.

L'étude de la nature des séries ayant une infinité de termes positifs et négatifs qui ne sont pas absolument convergentes sort du cadre de ce chapitre, dont l'objectif est de clarifier la notion de série et de somme infinie par un survol des propriétés des séries à termes positifs.

Nous verrons dans le chapitre suivant certaines séries de Fourier qui ne sont pas absolument convergentes, on admettra donc qu'elles convergent en appliquant le théorème donnant la valeur de leur somme.

Chapitre 6

Séries de Fourier

Nous allons maintenant revenir sur la question posée en début de semestre. Rappelons que nous cherchions à résoudre l'équation de la chaleur :

$$\frac{\partial T}{\partial t} = D\frac{\partial^2 T}{\partial x^2}$$

sur le domaine $\{(x,t)|x \in [0,L], t \geqslant 0\}$ en respectant les conditions initiales

$$T(x,0) = \phi(x)$$

(où ϕ est une fonction donnée) et les conditions aux bords

$$\frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(L,t) = 0$$

pour tout $t > 0$. Nous avions remarqué que lorsque la condition initiale ϕ était une somme finie de cosinus

$$\phi = a_0 + \sum_{k=1}^{n} a_k \cos\left(\frac{k\pi x}{L}\right)$$

cette équation possède une solution

$$T(x,t) = a_0 + \sum_{k=1}^{n} a_k \cos\left(\frac{k\pi x}{L}\right) e^{-Dk^2\pi^2 x/L^2}.$$

Nous allons maintenant chercher à résoudre cette équation pour une condition initiale ϕ quelconque en approchant ϕ par des sommes finies de la forme

$$\phi = a_0 + \sum_{k=1}^{n} a_k \cos\left(\frac{k\pi x}{L}\right).$$

Pour l'équation des ondes, on cherche plutôt à approcher ϕ par une somme de sinus. On va présenter une méthode pour approcher une fonction quelconque ϕ par une somme trigonométrique — c'est-à-dire, une fonction g de la forme

$$g(x) = a_0 + \sum_{k=1}^{n} a_k \cos\left(\frac{k\pi x}{L}\right) + b_k \sin\left(\frac{k\pi x}{L}\right).$$

Le cas de l'équation de la chaleur correspond au cas où tous les b_k sont nuls, celui de l'équation des ondes au cas où les a_k sont nuls.

6.1 Approximants de Fourier, coefficients de Fourier et séries de Fourier : définitions et exemples

Dans ce paragraphe, nous allons appliquer la méthode de la projection orthogonale pour approcher une fonction par une série trigonométrique.

On se donne une fonction f, continue, réelle, et définie sur une intervalle $[-L,L]$. On cherche à approcher $f(x)$ par une somme de fonctions trigonométriques fondamentales $2L$-périodiques :

$$a_0 + \sum_{k=1}^{n} a_k \cos\left(k\frac{\pi}{L}x\right) + b_k \sin\left(k\frac{\pi}{L}x\right).$$

Si f est une fonction du temps périodique de période T définie pour $t \in [-T/2, T/2]$, il faut remplacer x par t et L par $T/2$. Pour éviter de traîner des notations trop lourdes, on va poser

$$\omega = \frac{\pi}{L} = \frac{2\pi}{T},$$

de sorte que, pour une fonction f périodique dépendant du temps, ω est une pulsation, pour une fonction périodique dépendant de la position, la longueur d'onde est $2L$.

Pour simplifier, on pourra poser dans la suite :

$$L = \pi, \quad T = 2\pi, \quad \omega = 1.$$

Pour chaque n, nous allons chercher la fonction $S_n^{\mathbb{R}}(f)$, qui sera le meilleur approximant de f de la forme

$$S_n^{\mathbb{R}}(f)(x) = a_0 + \sum_{k=1}^{n} a_k \cos(k\omega x) + b_n \sin(k\omega x)$$

par rapport à la distance définie par le produit scalaire

$$\langle f|g\rangle = \int_{-L}^{L} f(x)g(x)\,\mathrm{d}x.$$

Dans ce chapitre nous aurons souvent besoin de travailler avec des fonctions qui ne sont pas continues sur $[-L,L]$ mais presque, au sens où elles peuvent avoir un nombre fini de « sauts ». On introduit donc la

Définition 6.1.1. Soit $i \geqslant 0$ un entier et $a < b \in \mathbb{R}$. Une fonction $f : [a,b] \to \mathbb{C}$ est dite $\mathscr{C}^i$ *par morceaux* s'il existe p réels a_k ($a = a_0 < a_1 < \ldots < a_p = b$) tels que

1) la fonction f est de classe $\mathscr{C}^i$ sur chaque intervalle $]a_k, a_{k+1}[$.
2) Pour tout $m = 0, \ldots, i$, et
   — pour tout $k = 0, \ldots, p-1$ les limites à droite $\lim_{x \to a_k^+} f^{(m)}(x)$,
   — pour tout $k = 1, \ldots, p$ les limites à gauche $\lim_{x \to a_k^-} f^{(m)}(x)$

   existent et sont finies.

Une fonction $f : \mathbb{R} \to \mathbb{C}$ est dite $\mathscr{C}^i$ par morceaux si elle est $\mathscr{C}^i$ par morceaux sur $[-n,n]$ pour tout n (i.e. les points de discontinuité sont en nombre fini sur un intervalle borné).

Une fonction $f : [a,b] \to \mathbb{C}$ est $\mathscr{C}^\infty$ par morceaux si elle est $\mathscr{C}^i$ par morceaux pour tout $i \geqslant 0$. On note l'espace vectoriel de toute les fonctions réelles (resp. complexes) $\mathscr{C}^i$ par morceaux sur une intervalle $[a,b]$ par $\mathscr{C}^i_{\mathrm{mor}}([a,b],\mathbb{R})$ (resp. $\mathscr{C}^i_{\mathrm{mor}}([a,b],\mathbb{C})$.)

Remarque 6.1.2. Il s'agit bien d'un espace vectoriel parce qu'on peut combiner les subdivisions associées à deux fonctions pour en obtenir une plus fine adaptée aux deux simultanément. De même, le produit de deux fonctions $\mathscr{C}^i$ par morceaux sur $[a,b]$ est $\mathscr{C}^i$ par morceaux sur $[a,b]$.

Remarque 6.1.3. Si $f : [a,b] \to \mathbb{C}$ est continue par morceaux, alors $\int_a^b f(x)\,\mathrm{d}x$ est bien définie.

Si $a = a_0, a_1, \ldots, a_p = b$ est la subdivision correspondante, on a

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{j=0}^{p-1} \int_{a_j}^{a_{j+1}} f(x)\,\mathrm{d}x.$$

Cela s'applique au produit de deux fonctions, donc on peut définir sur $\mathscr{C}^i_{\mathrm{mor}}([a,b],\mathbb{R})$ la forme bilinéaire

$$\langle f|g\rangle = \int_{-L}^{L} f(x)g(x)\,\mathrm{d}x.$$

Cette forme est « presque » un produit scalaire sur $\mathscr{C}^i_{\mathrm{mor}}([a,b],\mathbb{R})$. En effet, si $\int_a^b f^2(x)\,\mathrm{d}x = 0$ implique que $f \equiv 0$ sur un intervalle sur lequel f est continue, une fonction essentiellement nulle sur un intervalle mais avec quelques valeurs aberrantes isolées aurait ausi une intégrale nulle. Pour obtenir une définition de produit scalaire rigoureuse, on pourrait se restreindre au sous-espace vectoriel des fonctions $\mathscr{C}^i$ par morceaux qui sont nulles aux points de discontinuité, mais ce serait un peu trop restrictif. Dans la pratique cette question n'a pas beaucoup d'importance : on considérera comme essentiellement égales deux fonctions qui sont égales sauf en des points isolés.

Soit maintenant $f : [-L,L] \to \mathbb{R}$ une fonction continue par morceaux. On considère $\mathscr{C}^0_{\mathrm{mor}}([-L,L],\mathbb{R})$ avec son produit scalaire :

$$\langle f|g\rangle = \int_{-L}^{L} f(x)g(x)\,\mathrm{d}x.$$

Pour tout n, notons W_n l'ensemble des fonctions $g(x)$ de la forme

$$g(x) = a_0 + \sum_{k=1}^{n} a_k \cos(k\omega x) + b_k \sin(k\omega x).$$

On vérifie facilement que W_n est un sous-espace vectoriel de $\mathscr{C}^0_{\mathrm{mor}}([-L,L],\mathbb{R})$.

Définition 6.1.4. Le n-ième approximant trigonométrique de Fourier de f, noté $S_n^{\mathbb{R}}(f)$, est défini par la projection de f sur W_n

$$S_n^{\mathbb{R}}(f) = p_{W_n}(f).$$

Autrement dit, $S_n^{\mathbb{R}}(f)$ est la fonction dans W_n qui minimise la distance euclidienne

$$d(S_n^{\mathbb{R}}(f), f) = \sqrt{\int_{-L}^{L} (S_n^{\mathbb{R}}(f)(x) - f(x))^2\,\mathrm{d}x}.$$

Nous allons maintenant procéder au calcul explicite des approximants de Fourier. Il faut pour cela commencer par identifier une base orthonormée de W_n. Par définition de cet espace,

$$B = \{1, \cos(\omega x), \sin(\omega x), \cos(2\omega x) \ldots, \cos(n\omega x), \sin(n\omega x)\},$$

où $\omega = \frac{\pi}{L}$ est une famille génératrice de W_n.

Lemme 6.1.5. *La famille B est orthogonale pour W_n par rapport au produit scalaire $\langle f|g\rangle = \int_{-L}^{L} f(x)g(x)\,\mathrm{d}x$. C'est donc une base orthogonale de W_n.*

Preuve. Il nous faut démontrer que

1) $\int_{-L}^{L} 1 \times \sin(k\omega x)\,\mathrm{d}x = \int_{-L}^{L} 1 \times \cos(k\omega x)\,\mathrm{d}x = 0$ pour tout $k \neq 0$.
2) $\int_{-L}^{L} \cos(j\omega x)\sin(k\omega x)\,\mathrm{d}x = 0$ pour tout $k, j > 0$.
3) $\int_{-L}^{L} \sin(j\omega x)\sin(k\omega x)\,\mathrm{d}x = \int_{-L}^{L} \cos(j\omega x)\cos(k\omega x)\,\mathrm{d}x = 0$ pour $k \neq j > 0$.

On a $\omega L = \pi$, donc :

$$\int_{-L}^{L} \sin(k\omega x)\,\mathrm{d}x = \left[\frac{-1}{k\omega}\cos(k\omega x)\right]_{-L}^{L} = 0$$
$$\int_{-L}^{L} \cos(k\omega x)\,\mathrm{d}x = \left[\frac{1}{k\omega}\sin(k\omega x)\right]_{-L}^{L} = 0$$

donc (1) est vrai. Ensuite,

$$\int_{-L}^{L} \cos(j\omega x)\sin(k\omega x)\,\mathrm{d}x = \int_{-L}^{L} \frac{1}{2}(\sin((k+j)\omega x) + \sin((k-j)\omega x)\,\mathrm{d}x.$$

Par (1) cet intégrale est nulle si $k \neq j$. Lorsque $k = j$ on a que

$$\int_{-L}^{L} \cos(k\omega x)\sin(k\omega x)\,\mathrm{d}x = \int_{-L}^{L} \frac{1}{2}\sin(2k\omega x)\,\mathrm{d}x = 0.$$

Dans tous les cas, (2) est vérifié. Enfin, pour $k \neq j$ et $k, j > 0$ on a que

$$\int_{-L}^{L} \sin(j\omega x)\sin(k\omega x)\,\mathrm{d}x = \int_{-L}^{L} \frac{1}{2}(-\cos((k+j)\omega x) + \cos((k-j)\omega x)\,\mathrm{d}x = 0$$

par (1). De même

$$\int_{-L}^{L} \cos(j\omega x)\cos(k\omega x)\,\mathrm{d}x = \int_{-L}^{L} \frac{1}{2}(\cos((k+j)\omega x) + \cos((k-j)\omega x)\,\mathrm{d}x = 0$$

par (1). La condition (3) est donc vérifiée. Ceci termine la démonstration de l'orthogonalité de B.

Pour obtenir une base orthonormée pour $W_n(\mathbb{R})$ il suffira donc de normaliser la base

$$B = (1, \cos(\omega x), \sin(\omega x), \cos(2\omega x) \ldots, \cos(n\omega x), \sin(n\omega x))$$

On a :

$$\|1\|^2 = \int_{-L}^{L} 1^2\,\mathrm{d}x = 2L$$

puis pour les cosinus :

$$\|\cos(k\omega x)\|^2 = \int_{-L}^{L} \cos^2(k\omega x)\,\mathrm{d}x = \int_{-L}^{L} \frac{1}{2}(1 + \cos(2k\omega x))\,\mathrm{d}x = L.$$

Par un calcul similaire $\|\sin(k\omega x)\|^2 = L$. Nous résumons les calculs précédents dans la

Proposition 6.1.6 (Base orthonormée pour les séries de Fourier)**.** *Une base orthonormée $\tilde{B}$ de W_n est donnée par*

$$\tilde{B} = \left(\frac{1}{\sqrt{2L}}, \frac{\cos(\omega x)}{\sqrt{L}}, \frac{\sin(\omega x)}{\sqrt{L}}, \frac{\cos(2\omega x)}{\sqrt{L}}, \ldots, \frac{\cos(n\omega x)}{\sqrt{L}}, \frac{\sin(n\omega x)}{\sqrt{L}}\right).$$

Nous pouvons donc calculer l'approximant de Fourier trigonométrique $S_n^{\mathbb{R}}(f)$ en utilisant le formule de la projection orthogonale. Cette formule s'écrit comme suit :

$$\begin{aligned} S_n^{\mathbb{R}}(f)(x) &= \left\langle \frac{1}{\sqrt{2L}} \middle| f \right\rangle \frac{1}{\sqrt{2L}} + \sum_{k=1}^{n} \left\langle \frac{\cos(k\omega x)}{\sqrt{L}} \middle| f \right\rangle \frac{\cos(k\omega x)}{\sqrt{L}} + \left\langle \frac{\sin(k\omega x)}{\sqrt{L}} \middle| f \right\rangle \frac{\sin(k\omega x)}{\sqrt{L}} \\ &= \frac{1}{2L}\int_{-L}^{L} f(x)\,\mathrm{d}x + \sum_{k=1}^{n} a_k(f)\cos(k\omega x) + b_k(f)\sin(k\omega x) \end{aligned}$$

où les coefficients $a_k(f)$ et $b_k(f)$ sont définis par

$$a_k(f) = \frac{1}{L}\int_{-L}^{L} \cos(k\omega x) f(x)\,\mathrm{d}x, \quad b_k(f) = \frac{1}{L}\int_{-L}^{L} \sin(k\omega x) f(x)\,\mathrm{d}x.$$

Définition 6.1.7. Soit f une fonction réelle ou complexe continue par morceaux définie sur une intervalle $[-L,L]$ et $\omega = \pi/L$. Les coefficients de Fourier trigonométriques de f sont les nombres $a_0(f)$, $a_k(f), b_k(f)$ $(k>0)$ définis par

$$\begin{aligned} a_0(f) &= \frac{1}{2L}\int_{-L}^{L} f \,\mathrm{d}x, \\ a_k(f) &= \frac{1}{L}\int_{-L}^{L} \cos(k\omega x)\, f(x)\,\mathrm{d}x, \\ b_k(f) &= \frac{1}{L}\int_{-L}^{L} \sin(k\omega x)\, f(x)\,\mathrm{d}x. \end{aligned}$$

Le n-ième approximant de Fourier trigonométrique est alors donné par

$$S_n^{\mathbb{R}}(f)(x) = a_0(f) + \sum_{k=1}^{n} a_k(f)\cos(k\omega x) + b_k \sin(k\omega x).$$

Remarque 6.1.8. Il résulte des formules d'Euler :

$$e^{ik\omega x} = \cos(k\omega x) + i\sin(k\omega x)$$

$$\begin{cases} \cos(k\omega x) = \frac{1}{2}(e^{ik\omega x} + e^{-ik\omega x}) \\ \sin(k\omega x) = \frac{1}{2i}(e^{ik\omega x} - e^{-ik\omega x}) \end{cases}$$

qu'approcher f par une somme de fonctions trigonométriques équivaut à l'approcher par une somme d'exponentielles complexes. Plus précisément, considérons la somme trigonométrique

$$a_0 + \sum_{k=1}^{n} \left(a_k \cos(k\omega x) + b_k \sin(k\omega x)\right).$$

Si on pose $c_0 = a_0$ et pour tout $k>0$

$$c_k = \frac{1}{2}(a_k - ib_k),\ c_{-k} = \frac{1}{2}(a_k + ib_k)$$

alors on a

$$a_0 + \sum_{k=1}^{n} \left(a_k \cos(k\omega x) + b_k \sin(k\omega x)\right) = \sum_{k=-n}^{n} c_k e^{ik\omega x}.$$

Pour des fonctions complexes, cette version des approximants de Fourier est particulièrement utile.

Définition 6.1.9. Soit $f : [-L,L] \to \mathbb{C}$ une fonction continue par morceaux. On définit le n-ième approximant exponentiel de Fourier de f, noté $S_n^{\mathbb{C}}(f)$, par

$$S_n^{\mathbb{C}}(f)(x) = \sum_{k=-n}^{n} c_k e^{ik\omega x}$$

où c_k sont les coefficients exponentiels de Fourier de f définis par

$$c_k = \frac{1}{2}(a_k - ib_k),\ c_{-k} = \frac{1}{2}(a_k + ib_k)$$

où a_k et b_k sont les coefficients trigonométriques de Fourier de f.

En utilisant le formule pour les a_k et b_k ci-dessus, on obtient la formule suivante.

Définition 6.1.10. Soit f une fonction réelle ou complexe continue par morceaux et définie sur $[-L,L]$. Les coefficients de Fourier exponentiels de f sont les nombres $c_k(f)$ définis par

$$c_k(f) = \frac{1}{2L}\int_{-L}^{L} e^{-(ik\omega x)} f(x)\,\mathrm{d}x$$

Nous avons donc que l'approximant de Fourier trigonométrique est

$$S_n^{\mathbb{R}}(f)(x) = a_0(f) + \sum_{k=1}^{n} a_k(f)\cos(k\omega x) + b_k(f)\sin(k\omega x),$$

et que l'approximant exponentiel est

$$S_n^{\mathbb{C}}(f)(x) = \sum_{k=-n}^{n} c_k(f)e^{ik\omega x}.$$

Puisque l'on peut calculer les a_k, b_k, c_k pour toutes les valeurs de k, on peut poser, de façon relativement formelle :

Définition 6.1.11. Soit f une fonction continue par morceaux sur une intervalle $[-L,L]$ et $\omega = \pi/L$. La *série de Fourier trigonométrique* de f est la somme infinie

$$S^{\mathbb{R}}(f)(x) = a_0(f) + \sum_{k=1}^{\infty} a_k(f)\cos(k\omega x) + b_k(f)\sin(k\omega x).$$

La *série de Fourier exponentielle* de f est la somme infinie

$$S^{\mathbb{C}}(f)(x) = \sum_{k=-\infty}^{\infty} c_k(f)e^{ik\omega x}.$$

Remarque 6.1.12. À ce stade, il faut entendre ces notations comme une écriture formelle regroupant toutes les informations de tous les approximants de Fourier. Nous ne savons pas si cette série converge pour tout f ou x pour laquelle nous pouvons l'écrire (pour cela, il suffit qu'on puisse calculer les intégrales qui définissent ses coefficients, ce qui est une condition très faible sur f !).

Exemples 6.1.13.

1) On considère la fonction $f(x) = x$ sur l'intervalle $[-\pi,\pi]$. Nous avons ici $L = \pi, \omega = 1$ et

$$a_0(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} x\,\mathrm{d}x = 0.$$

Ensuite, pour tout $k > 0$, en intégrant par parties, on voit que

$$\begin{aligned} a_k(f) &= \frac{1}{\pi}\int_{-\pi}^{\pi} x\cos(kx)\,\mathrm{d}x \\ &= \frac{1}{\pi}\left[\frac{x}{k}\sin(kx)\right]_{-\pi}^{\pi} - \frac{1}{k\pi}\int_{-\pi}^{\pi}\sin(kx)\,\mathrm{d}x \\ &= -\frac{1}{k\pi}\int_{-\pi}^{\pi}\sin(kx)\,\mathrm{d}x = 0. \end{aligned}$$

On pouvait d'ailleurs éviter ce calcul en appliquant un argument de parité (voir plus bas). De même, une intégration par parties montre que

$$\begin{aligned} b_k(f) &= \frac{1}{\pi}\int_{-\pi}^{\pi} x\sin(kx)\,\mathrm{d}x \\ &= \left[-\frac{x}{k\pi}\cos(kx)\right]_{-\pi}^{\pi} + \frac{1}{k\pi}\int_{-\pi}^{\pi}\cos(kx)\,\mathrm{d}x \\ &= -\frac{2(-1)^k}{k} + \frac{1}{k\pi}\int_{-\pi}^{\pi}\cos(kx)\,\mathrm{d}x = \frac{2(-1)^{k+1}}{k}. \end{aligned}$$

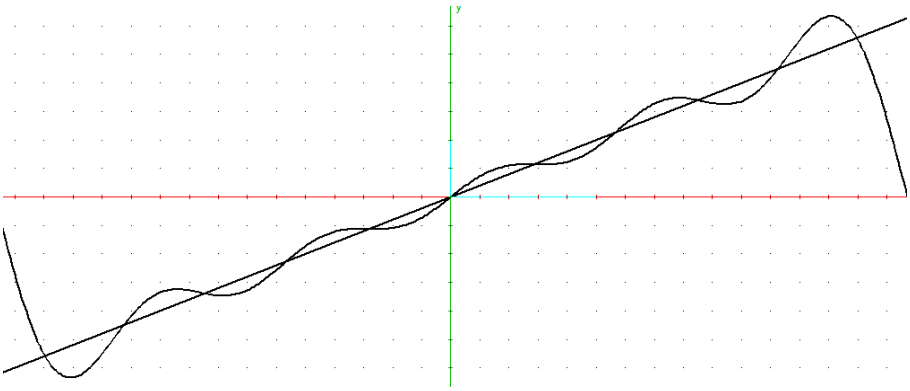


FIGURE 6.1 – Représentation graphique de $f(x) = x$ et de son approximant de Fourier à l'ordre 5.

La série de Fourier trigonométrique de f est donc la somme infinie

$$S^{\mathbb{R}}(f)(x) = \sum_{k=1}^{\infty} \frac{2(-1)^{k+1}}{k} \sin(kx).$$

(Attention, ce n'est pas une série absolument convergente, par exemple en $x = \pi/2$).
Calculons maintenant la série de Fourier exponentielle de f. L'intégration par parties nous donne que

$$c_k(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ikx} x \,\mathrm{d}x = \frac{1}{2\pi}\left[\frac{x}{-ik}e^{-ikx}\right]_{-\pi}^{\pi} + \int_{-\pi}^{\pi} \frac{1}{2ik\pi} e^{-ikx}\,\mathrm{d}x$$

$$= \frac{(-1)^k}{-2ik} = \frac{i(-1)^k}{2k}.$$

La série des Fourier exponentielle de f est donc

$$S^{\mathbb{C}}(f)(x) = \sum_{k=-\infty}^{\infty} \frac{i(-1)^k}{2k} e^{ikx}.$$

2) On considère la fonction f définie sur $[-\pi,\pi]$ telle que $f(x) = 1$ pour $x \geqslant 0$ et $f(x) = 0$ pour $x < 0$.
On a que

$$a_0(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\,\mathrm{d}x = \frac{1}{2}.$$

Par ailleurs, pour tout $k > 0$,

$$\begin{aligned} a_k(f) &= \frac{1}{\pi}\int_{-\pi}^{\pi} \cos(kx) f(x)\,\mathrm{d}x, \\ &= \frac{1}{\pi}\int_{0}^{\pi} \cos(kx)\,\mathrm{d}x \\ &= \frac{1}{k\pi}[\sin(kx)]_0^{\pi} \\ &= 0 \end{aligned} \qquad \begin{aligned} b_k(f) &= \frac{1}{\pi}\int_{-\pi}^{\pi} \sin(kx) f(x)\,\mathrm{d}x \\ &= \frac{1}{\pi}\int_{0}^{\pi} \sin(kx)\,\mathrm{d}x \\ &= \frac{1}{k\pi}[-\cos(kx)]_0^{\pi} \\ &= \frac{1-(-1)^k}{k\pi}. \end{aligned}$$

On note que $(1-(-1)^k) = 2$ si k est impair et 0 si k est pair. En écrivant tout k impair dans la forme $k = 2l+1$, on obtient que la séries de Fourier trigonométrique de f est

$$S^{\mathbb{R}}(f)(x) = \frac{1}{2} + \sum_{l=0}^{+\infty} \frac{2}{(2l+1)\pi} \sin((2l+1)x).$$

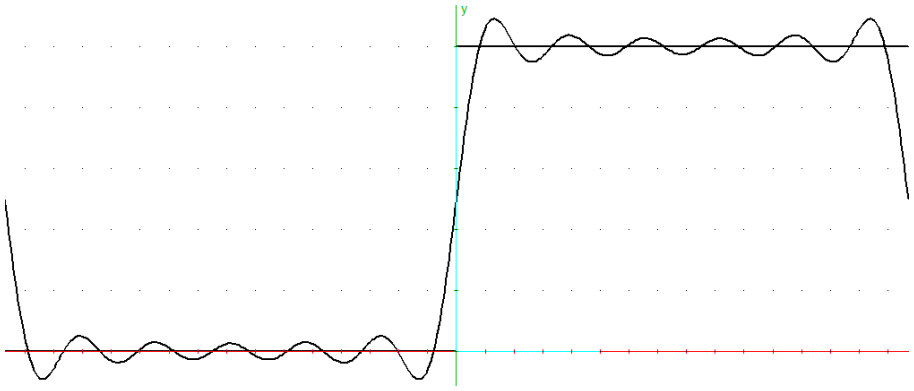


FIGURE 6.2 – Représentation graphique de la fonction créneau et de son approximant de Fourier à l'ordre 5.

Calculons maintenant la série de Fourier exponentielle de f. Pour $k \neq 0$ on a que

$$c_k(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{(-ikx)} f(x)\,\mathrm{d}x = \frac{1}{2\pi}\int_{0}^{\pi} e^{(-ikx)}\,\mathrm{d}x$$

$$= \frac{1}{-2ik\pi}[e^{-(ikx)}]_0^{\pi} = \frac{-1+(-1)^k}{-2ik\pi}.$$

Comme nous avons déjà calculé que $c_0 = a_0 = \frac{1}{2}$, la série de Fourier exponentielle de f est alors

$$S^{\mathbb{C}}(f)(x) = \frac{1}{2} + \sum_{l=-\infty}^{\infty} \frac{-i}{(2l+1)\pi} e^{i(2l+1)x}.$$

6.2 Séries en sinus et cosinus

La série de Fourier est une série qui mélange des termes en sinus et en cosinusinus. Nous montrerons dans cette section comment modifier cette construction pour obtenir des séries en sinus seulement, ou en cosinusinus seulement, pour approcher une fonction f donnée. Notre point de départ sera la proposition suivante :

Proposition 6.2.1. *Soit $f : [-L,L] \to \mathbb{R}$ une fonction réelle.*

1) Si f est paire alors $b_k(f) = 0$ pour tout k.

2) Si f est impaire alors $a_k(f) = 0$ pour tout k.

Preuve. On note que pour toute fonction g, impaire sur $[-L,L]$, nous avons que

$$\int_{-L}^{L} g(x)\,\mathrm{d}x = 0.$$

Par ailleurs,

1) Le produit d'une fonction impaire par une fonction paire est lui-même une fonction impaire.
2) Pour tout $k > 0$ la fonction $\cos(kx)$ est une fonction paire et la fonction $\sin(kx)$ est une fonction impaire.

Si la fonction $f : [-L,L] \to \mathbb{R}$ est une fonction paire alors pour tout k la fonction $f(x)\sin(k\pi x)$ est une fonction impaire et donc

$$b_k(f) = \frac{1}{L}\int_{-L}^{L} f(x)\sin(k\omega x)\,\mathrm{d}x = 0.$$

Par contre, si la fonction $f : [-L, L] \to \mathbb{R}$ est une fonction impaire alors pour tout k la fonction $f(x)\cos(k\omega x)$ est une fonction impaire et donc

$$a_k(f) = \frac{1}{L}\int_{-L}^{L} f(x)\sin(k\omega x)\,\mathrm{d}x = 0.$$

Ceci termine la démonstration de la proposition.

Avec cette proposition, nous pourrons construire des sommes de cos (resp. de sin) approchant une fonction donnée. Notre méthode sera la suivante :

1) Étant donnée une fonction $f : [0, L] \to \mathbb{R}$, on construit une extension g sur $[-L, L]$ qui est paire (si on veut construire une séries en cosinus) ou impaire (si on veut construire une série en sinus.)
2) Nous calculons alors la séries de Fourier de cette nouvelle fonction g.
3) Puisque g est paire (resp. impaire) sa séries de Fourier ne contient que des termes en cosinus (resp. en sinus). Cela donnera bien une série de Fourier en cosinus (resp. en sinus) pour f.

Définition 6.2.2. Soit $f : [0, L] \to \mathbb{R}$ une fonction continue par morceaux. On définit sur $[-L, L]$ les extensions impaire et paire de f par

$$f_{\text{paire}}(x) = \begin{cases} f(x) & \text{si } x \geqslant 0 \\ f(-x) & \text{si } x < 0, \end{cases}$$

$$f_{\text{impaire}}(x) = \begin{cases} f(x) & \text{si } x > 0 \\ -f(-x) & \text{si } x < 0 \\ 0 & \text{si } x = 0. \end{cases}$$

Notons que f_{paire} est paire et f_{impaire} est impaire par construction. il résulte de la proposition 6.2.1 que

(a) pour tout k, $a_k(f_{\text{impaire}}) = 0$,
(b) pour tout k, $b_k(f_{\text{paire}}) = 0$.

Nous pouvons maintenant définir les séries en sinus et cosinus de notre fonction f.

Définition 6.2.3. Soit $f : [0, L] \to \mathbb{R}$ une fonction continue par morceaux. La série en sinus de f est la série trigonométrique

$$S^{\sin}(f)(x) = S^{\mathbb{R}}(f_{\text{impaire}})(x) = \sum_{k=1}^{\infty} b_k(f_{\text{impaire}})\sin(k\omega x).$$

La série en cosinus de f est la série trigonométrique

$$S^{\cos}(f)(x) = S^{\mathbb{R}}(f_{\text{paire}})(x) = a_0(f_{\text{paire}}) + \sum_{k=1}^{\infty} a_k(f_{\text{paire}})\cos(k\omega x).$$

Remarque 6.2.4. Notons que par parité

$$b_k(f_{\text{impaire}}) = \frac{1}{L}\int_{-L}^{L} f_{\text{impaire}}(x)\sin(k\omega x)\,\mathrm{d}x = \frac{2}{L}\int_{0}^{L} f(x)\sin(k\omega x)\,\mathrm{d}x.$$

De même

$$a_k(f_{\text{paire}}) = \frac{1}{L}\int_{-L}^{L} f_{\text{paire}}(x)\cos(k\omega x)\,\mathrm{d}x = \frac{2}{L}\int_{0}^{L} f(x)\cos(k\omega x)\,\mathrm{d}x.$$

et

$$a_0(f) = \frac{1}{L}\int_{0}^{L} f(x)\,\mathrm{d}x.$$

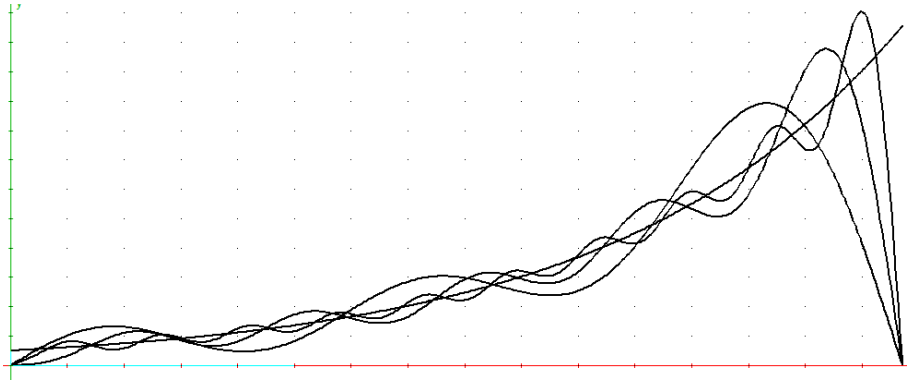


FIGURE 6.3 – Représentation graphique de la fonction exponantielle et de son approximant de Fourier (en sinus) aux ordres 5, 10 et 20.

Exemple 6.2.5. On considère la fonction $f(x) = e^x$ sur l'intervalle $[0,\pi]$. Nous cherchons à calculer sa série en sinus. Nous avons que

$$S^{\sin}(f)(x) = \sum_{k=1}^{\infty} b_k \sin(kx)$$

avec

$$\begin{aligned}
b_k &= \frac{2}{\pi}\int_0^{\pi} e^x \sin(kx)\,\mathrm{d}x \\
&= \frac{2}{\pi}\int_0^{\pi} e^x \frac{e^{ikx}-e^{-ikx}}{2i}\,\mathrm{d}x = \\
&= \frac{1}{i\pi}\int_0^{\pi} e^{x(1+ik)} - e^{x(1-ik)}\,\mathrm{d}x \\
&= \frac{1}{i\pi}\left[\frac{e^{x(1+ik)}}{1+ik} - \frac{e^{x(1-ik)}}{1-ik}\right]_0^{\pi} \\
&= \frac{1}{i\pi}\left[\frac{(1-ik)e^{x(1+ik)} - (1+ik)e^{x(1-ik)}}{1+k^2}\right]_0^{\pi} \\
&= \frac{1}{i\pi}\left[\frac{e^x(-2ik\cos(kx) + 2i\sin(kx))}{1+k^2}\right]_0^{\pi} \\
&= \frac{2k(1-(-1)^k e^{\pi})}{(1+k^2)\pi}.
\end{aligned}$$

La série en sinus de f est donc

$$S^{\sin}(f)(x) = \sum_{k=0}^{\infty} \frac{2k(1-(-1)^k e^{\pi})}{(1+k^2)\pi}\sin(kx).$$

6.3 Convergence des séries de Fourier

Nous avons donc créé, pour chaque fonction f continue par morceaux définie sur une intervalle $[-L,L]$, une suite de fonctions $S_n^{\mathbb{R}}(f)$. Chaque élément dans cette suite de fonctions est « plus proche » de la fonction f que celle qui la précède. Mais pour nous être utile, il faudrait s'assurer que, quitte à prendre n très grand, la fonction $S_n^{\mathbb{R}}(f)$ est aussi proche que l'on veut de la fonction f. Cette question sera le sujet de ce paragraphe.

Le théorème suivant, dont la démonstration dépasse le cadre de ce cours, nous assure que si la fonction f est continument dérivable alors en tout point x la série de Fourier converge vers f, plus précisément :

Théorème 6.3.1 (Dirichlet)**.** *Soit $f : [-L,L] \to \mathbb{C}$ une fonction $\mathscr{C}^1$ par morceaux. Alors pour tout point $x \in]-L,L[$ où f est continue, on a :*

$$\lim_{n\to\infty} S_n^{\mathbb{R}}(f)(x) \to f(x)$$

Pour les valeurs de $x \in]-L,L[$ où f effectue un saut, on a convergence de la série vers la moyenne des limites à droite et à gauche :

$$\lim_{n\to\infty} S_n^{\mathbb{R}}(f)(x) = \frac{1}{2}\left(\lim_{y\to x^+} f(y) + \lim_{y\to x^-} f(y)\right).$$

Ce résultat s'étend aux extrémités :

$$\lim_{n\to\infty} S_n^{\mathbb{R}}(f)(-L) = \lim_{n\to\infty} S_n^{\mathbb{R}}(f)(L) = \frac{1}{2}\left(\lim_{x\to L^-} f(x) + \lim_{x\to -L^+} f(x)\right).$$

En particulier, si f est continue et $\mathscr{C}^1$ par morceaux, $S_n^{\mathbb{R}}(f)$ converge vers f sur $]-L,L[$.

Attention ! Ce résultat est *faux* si la fonction f n'est pas supposée dérivable.

Idée de la preuve : on se place pour simplifier en un point x où f est continue. On peut supposer que $f(x) = 0$ en observant que la série de Fourier d'une fonction constante c est $a_0 = c, a_k = b_k = 0$. Il s'agit donc de montrer que

$$\lim_{n\to\infty}\left(a_0 + \sum_{k=1}^{n} a_k(f)\cos(k\omega x) + b_k(f)\sin(k\omega x)\right) = 0$$

ou, avec les coefficients de Fourier exponentiels ;

$$\lim_{n\to\infty} \sum_{k=-n}^{n} c_k(f) e^{ik\omega x} = 0.$$

En remplaçant les c_k par leur valeur

$$\lim_{n\to\infty} \sum_{k=-n}^{n} \left(\int_{-L}^{L} f(t) e^{-ik\omega t}\,\mathrm{d}t\right) e^{ik\omega x} = 0,$$

on peut rentrer $e^{ik\omega x}$ dans l'intégrale en t. On cherche donc la limite lorsque n tend vers l'infini de

$$\sum_{k=-n}^{n} \int_{-L}^{L} f(t) e^{-ik\omega t} e^{ik\omega x}\,\mathrm{d}t = \int_{-L}^{L} f(t) \sum_{k=-n}^{n} e^{ik\omega(x-t)}\,\mathrm{d}t.$$

Dans l'intervalle $]-L,L[$, si $x \neq t$ alors $\rho = e^{i\omega(x-t)} \neq 1$ donc la sommation dans l'intégrale est la somme d'une série géométrique de raison différente de 1, elle vaut

$$\begin{aligned}
\sum_{k=-n}^{n} e^{ik\omega(x-t)} &= \sum_{k=-n}^{n} \rho^k \\
&= \rho^{-n}\frac{\rho^{2n+1}-1}{\rho-1} \\
&= \frac{\rho^{n+1/2} - \rho^{-n-1/2}}{\rho^{1/2} - \rho^{-1/2}} \\
&= \frac{e^{i(n+1/2)\omega(x-t)} - e^{-i(n+1/2)\omega(x-t)}}{e^{i(1/2)\omega(x-t)} - e^{-i(1/2)\omega(x-t)}} \\
&= \frac{2i\sin\left((n+1/2)\,\omega(x-t)\right)}{2i\sin\left((1/2)\omega(x-t)\right)} \\
&= \frac{\sin\left((n+1/2)\,\omega(x-t)\right)}{\sin\left((1/2)\omega(x-t)\right)}.
\end{aligned}$$

On observe que cette somme vaut $2n+1$ lorsque $x=t$ qui est la limite de l'expression ci-dessus lorsque x tend vers t. On est donc ramené à montrer que, lorsque $n\to\infty$, la limite ci-dessous est nulle :

$$\lim_{n\to\infty}\int_{-L}^{L} f(t)\frac{\sin\left((n+\frac{1}{2})\omega(x-t)\right)}{\sin\left(\frac{1}{2}\omega(x-t)\right)}\,\mathrm{d}t=0.$$

Ceci vient d'une mise en forme rigoureuse des observations suivantes :

1) Lorsque t est proche de x, $f(t)$ tend vers 0 (on a supposé $f(x)=0$, et f continue en x).
2) Dès qu'on séloigne de 0, $F_x(t)=f(t)/\sin\left(\frac{1}{2}\omega(x-t)\right)$ est $\mathscr{C}^1$ par morceaux. On peut intégrer par parties et faire apparaitre un facteur $1/(n+\frac{1}{2})$ qui tend vers 0 lorsque n tend vers l'infini :

$$\int F_x(t)\sin\left((n+\frac{1}{2})\omega(x-t)\right)\mathrm{d}t=\left[F_x(t)\frac{\cos\left((n+\frac{1}{2})\omega(x-t)\right)}{-(n+\frac{1}{2})\omega}\right] - \int\frac{dF_x(t)}{\mathrm{d}t}\frac{\cos\left((n+\frac{1}{2})\omega(x-t)\right)}{-(n+\frac{1}{2})\omega}\,\mathrm{d}t.$$

Il y a une autre forme de convergence qui nous sera utile, qui dit que la distance de $S_n^{\mathbb{R}}(f)$ à f (pour le produit scalaire intégral sur $\mathscr{C}^0_{\mathrm{mor}}([-L,L],\mathbb{R})$) tend vers 0.

Théorème 6.3.2 (Théorème de Parseval)**.** *Soit f une fonction réelle continue par morceaux sur une intervalle $[-L,L]$. Soit d_n la distance de f à la somme partielle d'ordre n de sa série de Fourier :*

$$d_n=d(f,S_n^{\mathbb{R}}(f))=\sqrt{\int_{-L}^{L}(f(x)-S_n^{\mathbb{R}}(f)(x))^2\,\mathrm{d}x}$$

entre f et $S_n^{\mathbb{R}}(f)$. Alors $d_n\to 0$ quand $n\to\infty$.

À nouveau, la démonstration de ce théorème dépasse le cadre de ce cours : on montre que le résultat est vrai lorsque f est assez régulière en appliquant Dirichlet puis on fait un raisonnement « par densité », en approchant pour le produit scalaire une fonction continue par morceaux par une fonction régulière.

En appliquant le fait que la somme partielle de la série de Fourier est une projection orthogonale de f pour la norme $\|g\|^2=\int_{-L}^{L}g(t)^2\,\mathrm{d}t$ et en appliquant l'inégalité triangulaire, on a

$$0\leqslant\|f\|-\|S_n^{\mathbb{R}}(f)\|\leqslant d_n\quad\Rightarrow\quad\|S_n^{\mathbb{R}}(f)\|\to\|f\|.$$

Comme $S_n^{\mathbb{R}}(f)$ s'exprime en fonction de $(1,\cos(\omega x),\sin(\omega x),\dots)$ qui sont orthogonales pour le produit scalaire, on peut appliquer le théorème de Pythagore à

$$S_n^{\mathbb{R}}(f)(x)=a_0+a_1\cos(\omega x)+\cdots+a_n\cos(n\omega x)+b_1\sin(\omega x)+\cdots+b_n\sin(n\omega x)$$

et on obtient

$$\|S_n^{\mathbb{R}}(f)\|^2=a_0^2\|1\|^2+a_1^2\|\cos(\omega x)\|^2+\cdots+a_n^2\|\cos(n\omega x)\|^2+b_1^2\|\sin(\omega x)\|^2+\cdots+b_n^2\|\sin(n\omega x)\|^2.$$

En utilisant les calculs de normes de la proposition 6.1.6, on en déduit :

Corollaire 6.3.3 (Égalité de Parseval)**.** *Soit f une fonction réelle continue par morceaux sur une intervalle $[-L,L]$. On a l'égalité*

$$\lim_{n\to\infty}\|S_n^{\mathbb{R}}(f)\|=\lim_{n\to\infty}\left(2La_0^2+\sum_{k=1}^{n}L(a_k^2+b_k^2)\right)=\int_{-L}^{L}f(x)^2\,\mathrm{d}x.$$

Autrement dit,

$$2La_0^2+\sum_{k=1}^{\infty}L(a_k^2+b_k^2)=\int_{-L}^{L}f(x)^2\,\mathrm{d}x.$$

Exemple 6.3.4. Considérons $f(x) = x$ sur $[-\pi, \pi]$. On a vu que, pour cette fonction, $a_k = 0$ pour tout k, et $b_k = \frac{2(-1)^k}{k}$. Par ailleurs

$$\int_{-\pi}^{\pi} x^2 \, \mathrm{d}x = \frac{2\pi^3}{3}.$$

Il s'ensuit

$$\pi \sum_{k=1}^{\infty} \frac{4}{k^2} = \frac{2\pi^3}{3},$$

donc

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}.$$

Corollaire 6.3.5. *Soient $f, g : \mathbb{R} \to \mathbb{C}$ deux fonctions réelles continues par morceaux et définies sur $[-L, L]$. Si $a_n(f) = a_n(g)$ et $b_n(f) = b_n(g)$ pour tout n, alors, pour tout point x où f et g sont toutes les deux continues, on a $f(x) = g(x)$.*

Preuve. L'égalité de Parseval implique que si $a_n(f) = a_n(g)$ et $b_n(f) = b_n(g)$ pour tout n alors

$$\int_{-L}^{L} |f(x) - g(x)|^2 \, \mathrm{d}x = 0.$$

Ceci n'est possible que si $f(x) = g(x)$ en tout point x où f et g sont continues.

6.4 Solutions d'équations aux dérivées partielles

Dans cette section, nous serons amenés à faire quelques manipulations dont nous ne pourrons pas donner une justification complète.

6.4.1 L'équation de la chaleur

On rappelle qu'il s'agit de déterminer l'évolution au cours du temps de la température $T(x,t)$ d'une barre de longueur L ($x \in [0, L]$), isolée à ses deux extrémités. On cherche donc une fonction $\mathscr{C}^\infty$

$$T : (x,t) \in [0,L] \times \mathbb{R}^+ \to T(x,t) \in \mathbb{R}$$

telle que

$$\frac{\partial T}{\partial t} = D \frac{\partial^2 T}{\partial x^2}, x \in [0,L], t > 0$$

avec les conditions aux bords en tout temps

$$\frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(L,t) = 0$$

et la condition initiale à l'instant $t = 0$,

$$T(x,0) = \varphi(x) \text{ pour tout } x \in [0,L].$$

Remarquons que, puisque nous cherchons une solution T qui est $\mathscr{C}^\infty$, la condition initiale φ doit aussi être $\mathscr{C}^\infty$ sur $[0,L]$.

On va simplifier le problème en cherchant la série de Fourier de T, ce qui transformera les dérivées en x en multiplications par $k\omega$. Plus précisément on remplace $\varphi(.)$ et $T(.,t)$ par leur série en cosinus sur $[0,L]$, ce qui permettra de satisfaire à la condition d'isolation au bord puisque la dérivée du cos est un sin qui s'annule aux bords. Posons donc :

$$T(x,t) = \sum_{k0}^{\infty} a_k(t) \cos(k\omega x), \quad \omega = \frac{\pi}{L}.$$

On remplace dans l'équation aux dérivées partielles

$$\frac{\partial}{\partial t}\left(\sum_{k\geqslant 0} a_k(t)\cos(k\omega x)\right) = D\frac{\partial^2}{\partial x^2}\left(\sum_{k\geqslant 0} a_k(t)\cos(k\omega x)\right).$$

On admet qu'on peut intervertir la dérivation partielle avec la somme, on obtient donc

$$\left(\sum_{k\geqslant 0} \frac{\partial}{\partial t}(a_k(t))\cos(k\omega x)\right) = D\left(\sum_{k\geqslant 0} a_k(t)\frac{\partial^2}{\partial x^2}\cos(k\omega x)\right),$$

soit

$$\left(\sum_{k\geqslant 0} a_k'(t)\cos(k\omega x)\right) = D\left(\sum_{k\geqslant 0} a_k(t)(-k^2\omega^2)\cos(k\omega x)\right).$$

Deux fonctions régulières ayant le même développement en séries de Fourier sont égales, donc on obtient l'égalité

$$a_k'(t) = -Dk^2\omega^2 a_k(t),$$

équation différentielle ordinaire dont la solution est

$$a_k(t) = e^{-Dk^2\omega^2 t}a_k(0).$$

À l'instant $t = 0$, on a

$$\varphi(x) = T(x,0) = \sum_{k\geqslant 0} a_k(0)\cos(k\omega x),$$

donc les $a_k(0)$ sont les coefficients de la série en cosinus de la condition initiale. D'où le :

Théorème 6.4.1. *L'équation de la chaleur*

$$\frac{\partial T}{\partial t} = D\frac{\partial^2 T}{\partial x^2}, \quad x\in[0,L], t\geqslant 0$$

admet une unique solution $\mathscr{C}^\infty$ vérifiant les conditions aux bords

$$\frac{\partial T}{\partial x}(0,t) = \frac{\partial T}{\partial x}(L,t) = 0$$

et la condition initiale à l'instant $t = 0$

$$T(x,0) = \varphi(x) \quad x\in[0,L]$$

où φ est une fonction $\mathscr{C}^\infty$ sur $[0,L]$ telle que $\varphi'(0) = \varphi'(L) = 0$.

Si les a_k sont les coefficients de la série en cosinus de φ, on a :

$$T(x,t) = a_0 + \sum_{k=1}^{\infty} a_k\cos(k\omega x)\,e^{-k^2\omega^2 Dt}$$

avec

$$\omega = \frac{\pi}{L}, \quad a_0 = \frac{1}{L}\int_0^L \varphi(x)\ \mathrm{d}x, \quad a_k = \frac{2}{L}\int_0^L \varphi(x)\cos(k\omega x)\ \mathrm{d}x\ (k>0).$$

Remarque 6.4.2. La démonstration que la fonction T est bien définie et $\mathscr{C}^\infty$ dépasse le cadre de ce cours. La régularité d'une fonction périodique peut se lire sur la décroissance de ses coefficients de Fourier a_k et b_k lorsque k tend vers l'infini. Si f est régulière, on peut intégrer par parties dans le calcul de a_k et b_k et faire apparaitre autant de puissances de k négatives que l'on veut, a_k et b_k tendent vers 0 plus vite que n'importe quelle puissance négative de k. Réciproquement, les puissances de k qui apparaissent quand on dérive sous le signe somme ne gêneront pas la convergence de la série. On peut d'ailleurs observer qu'il suffit de prendre la condition initiale $\mathscr{C}^2$ en x pour obtenir une solution $\mathscr{C}^\infty$ en tout instant non nul grâce à l'exponentielle qui décroit plus vite que toute puissance négative de k (l'équation de la chaleur régularise la solution pour $t > 0$).

6.4.2 L'équation des ondes

On rappelle que l'on cherche une fonction

$$T : [0,L] \times \mathbb{R}^+ \to \mathbb{R}$$

telle que

$$\frac{\partial^2 T}{\partial t^2} = D\frac{\partial^2 T}{\partial x^2}, x \in [0,L], t > 0$$

avec les conditions aux bords

$$T(0,t) = T(L,t) = 0$$

et les conditions initiales

$$T(x,0) = \varphi(x) \text{ pour tout } x \in [0,L], t = 0,$$

$$\frac{\partial T}{\partial t}(x,0) = 0 \text{ pour tout } x \in [0,L], t = 0.$$

Remarquons que, puisque la solution T est $\mathscr{C}^\infty$, la condition initiale φ doit aussi être $\mathscr{C}^\infty$ sur $[0,L]$.
Nous allons montrer le théorème suivant :

Théorème 6.4.3. *Soit $\varphi : [0,L] \to \mathbb{R}$ une fonction $\mathscr{C}^\infty$ sur $[0,L]$ telle que $\varphi(0) = \varphi(L) = 0$. Alors l'équation des ondes*

$$\frac{\partial T^2}{\partial t^2} = D\frac{\partial^2 T}{\partial x^2}, x \in [0,L], t \geqslant 0$$

admet une unique solution $\mathscr{C}^\infty$

$$T : [0,L] \times \mathbb{R}^+ \to \mathbb{R}$$

vérifiant les conditions aux bords

$$T(0,t) = T(L,t) = 0$$

et les conditions initiales

$$T(x,0) = \varphi(x) \textit{ pour tout } x \in [0,L], t = 0$$

$$\frac{\partial T}{\partial t}(x,0) = 0 \textit{ pour tout } x \in [0,L], t = 0.$$

Cette solution est donnée par la formule

$$T(x,t) = \sum_{k=1}^{\infty} b_k \sin(k\omega x)\cos\left(k\sqrt{D}\omega t\right),$$

où pour tout $k > 0$ on a que $b_k = \frac{2}{L}\int_0^L \varphi(x)\sin(k\omega x)\,\mathrm{d}x$.

Puisque φ est $\mathscr{C}^1$, et $\varphi(0) = \varphi(L) = 0$, son extension impaire, φ_{impaire} est $\mathscr{C}^1$. Par le théorème de Dirichlet il suit que sur $[-L,L]$ nous avons que

$$\varphi_{\text{impaire}}(x) = \sum_{k=1}^{\infty} b_k(\varphi_{\text{impaire}})\sin(k\omega x).$$

et en particulier pour tout $x \in [0,L]$

$$\varphi(x) = \sum_{k=1}^{\infty} b_k \sin(k\omega x).$$

où $b_k = \frac{2}{L}\int_0^L \varphi(x)\sin(k\omega x)\,\mathrm{d}x$.

Autrement dit, le théorème de Dirichlet nous dit dans ce cas que φ est égale à la somme (infinie) de sa série en sinus.

Par analogie avec le cas ou φ est donnée par une somme finie de sinus, nous allons poser

$$T(x,t) = \sum_{k\geqslant 1} b_k \sin(k\omega x)\cos\left(k\omega\sqrt{D}t\right),$$

et vérifier que T est solution de l'équation des ondes. Pour tout $x \in [0,L]$, on a

$$T(x,0) = \sum_{k\geqslant 1} b_k \sin(k\omega x) = \varphi(x)$$

et la première condition initiale est donc vérifiée. De plus, $T(0,t) = T(L,T) = 0$ d'après les propriétés du sinus : les conditions aux bords sont donc vérifiées.

Posons

$$u_k(x,t) = b_k \sin(k\omega x)\cos\left(k\omega\sqrt{D}t\right),$$

de telle façon que $T(x,t) = \sum_k u_k(x,t)$. Comme pour l'équation de la chaleur [1], on admettra qu'on peut intervertir la dérivation et le signe $\sum$, i.e. en posant

$$T(x,t) = \sum_k u_k(x,t)$$

on a bien que

$$\frac{\partial T}{\partial x} = \sum_{k=1}^{\infty} \frac{\partial u_k}{\partial x}$$

et

$$\frac{\partial T}{\partial t} = \sum_{k=1}^{\infty} \frac{\partial u_k}{\partial t}$$

Vérifions maintenant la deuxième condition initiale. Puisque pour tout k, $\frac{\partial u_k}{\partial t}(x,0) = 0$, il suit que $\frac{\partial T}{\partial t}(x,0) = 0$ pour tout $x \in [0,L]$.

Il reste à vérifier que T satisfait l'équation $\frac{\partial^2 T}{\partial t^2} = D\frac{\partial^2 T}{\partial x^2}$. On a que

$$\frac{\partial^2 T}{\partial t^2} = \sum_{k\geqslant 1} b_k \frac{\partial^2}{\partial t^2}\left(\sin(k\omega x)\cos\left(k\omega\sqrt{D}t\right)\right)$$

$$= \sum_{k\geqslant 1} -D\omega^2 k^2 b_k \left(\sin(k\omega x)\cos\left(k\omega\sqrt{D}t\right)\right).$$

Mais on a aussi

$$\frac{\partial^2 T}{\partial x^2} = \sum_{k\geqslant 1} b_k \frac{\partial^2}{\partial x^2}\left(\sin(k\omega x)\cos\left(\omega k\sqrt{D}t\right)\right)$$

$$= \sum_{k\geqslant 1} -\omega^2 k^2 b_k \left(\sin(k\omega x)\cos\left(k\omega\sqrt{D}t\right)\right).$$

On a donc bien

$$\frac{\partial^2 T}{\partial t^2} = D\frac{\partial^2 T}{\partial x^2}$$

et la fonction donnée est donc une solution de notre équation.

Montrons maintenant que cette solution est unique. Pour cela, supposons que l'on ait deux solutions $\mathscr{C}^\infty$ du problème, disons T_1 et T_2, et posons $u = T_1 - T_2$. Il est facile de vérifier que l'on a

$$\frac{\partial u}{\partial t} = D\frac{\partial^2 u}{\partial x^2}, x \in [0,L], t \in \mathbb{R}^+,$$

1. Mais contrairement à l'équation de la chaleur, on a un facteur en cosinus qui oscille au lieu d'une exponentielle décroissante, il n'y a donc pas de régularisation en temps $t > 0$ pour une condition initiale éventuellement non régulière

et que

$$u(0,t) = u(L,t) = u(x,0) = 0 \text{ pour tout } x \in [0,L], t \in \mathbb{R}^+.$$

Considérons la fonction

$$b_k(t) = \frac{1}{2D} \int_0^L u(x,t) \sin(k\omega x)\, \mathrm{d}x.$$

Autrement dit, $b_k(t)$ est le k-ième coefficient dans la série en sinus de la fonction $x \to u(x,t)$. Par le corollaire 6.3.5 il suffira de montrer que $b_k(t)$ est nul pour tout t. Par les conditions initiales on a que

$$b_k(0) = b_k,\ \frac{\partial b_k}{\partial t}(0) = \int_0^L \frac{\partial u_k(x,0)}{\partial t} \sin(k\omega x)\, \mathrm{d}x = 0.$$

Calculons

$$\begin{aligned} b_k''(t) &= \int_0^L \frac{\partial^2 u(x,t)}{\partial t^2} \sin(k\omega x)\, \mathrm{d}x \\ &= D \int_0^L \frac{\partial^2 u(x,t)}{\partial x^2} \sin(k\omega x)\, \mathrm{d}x = 0 \end{aligned}$$

ce qui est égal, après une double intégration par parties et en utilisant les conditions aux bords, à

$$-\frac{Dk^2\pi^2}{L^2} \int_0^L \frac{\partial^2 u(x,t)}{\partial x^2} \sin(k\omega x)\, \mathrm{d}x = -Dk^2\omega^2 b_k(t).$$

Autrement dit,

$$b_k''(t) = -Dk^2\omega^2 b_k(t).$$

Mais la seule solution $\mathscr{C}^\infty$ de cette fonction telle que $b_k(0) = b_k'(0) = 0$ est la fonction nulle. On a donc $b_k(t) = 0$ pour tout k et tout t, ce qui donne bien que $u(x,t) = 0$ pour tout x,t.

Annexe A

Appendice : Espace-temps, bases et forme de Minkowski

La théorie de la relativité pose que la séparation que nous observons entre l'espace et le temps est une illusion et qu'en réalité les événements sont placés dans un continuum de dimension 4 que l'on appelle l'espace-temps. Cet espace-temps dont les points représentent des événements est de dimension 4, puisque pour préciser un événement il faut donner :

1) le lieu ou il s'est produit (précisé dans des coordonnées cartésiennes par 3 données numériques) et
2) l'heure à laquelle il s'est produit (précisée par 1 donnée numérique).

Donc, pour préciser un événement il faut 4 coordonnées.

Contrairement à la physique newtonienne, il n'est plus possible dans la relativité restreinte de donner une décomposition de l'espace-temps en une partie spatiale plus une partie temps. Plus précisément, la physique newtonienne, si elle place bien les événements dans un continuum de dimension 4 (lieu ou l'événement a eu lieu, plus l'heure à laquelle il s'est produit) pose aussi l'existence une décomposition intrinsèque

$$\text{Espace-temps} \;=\; \text{espace} \oplus \text{temps}.$$

Deux observateurs newtoniens, même en mouvement, seront toujours d'accord pour dire que deux événements ont lieu au même moment (c'est-à-dire, que le vecteur qui les sépare dans l'espace-temps est contenu dans le sous-espace « espace ») ou qu'ils ont lieu au même endroit (c'est-à-dire, que le vecteur qui les sépare dans l'espace-temps est contenu dans le sous-espace « temps »).

Ceci n'est plus vrai dans la relativité restreinte : selon les observateurs la décomposition de l'espace-temps en espace et temps va varier. Lorsqu'un observateur O, non soumis à une accélération, observe une événement, il va le mesurer en utilisant son *référentiel*, et le résultat de cette mesure sera un *quadrivecteur* (T_O, X_O, Y_O, Z_O) où

— $\frac{T_O}{c}$ est l'heure à laquelle l'évènement s'est produit, mesuré par l'observateur O,

— (X_O, Y_O, Z_O) est la position de l'événement, mesuré par O.

Ces mesures n'ont plus rien d'absolu ; elles varieront selon l'observateur.

Qu'est-ce qu'un référentiel ? *C'est la donnée d'une origine A dans l'espace-temps* [1] **et d'une base de l'espace-temps**. Après avoir fixé une origine A, l'observateur O mesure l'espace-temps en utilisant une base qui lui est propre (t_O, x_O, y_O, z_O). Si v est un élément de l'espace-temps (que l'on considère comme un espace vectoriel avec origine A), les coordonnées de v dans la base (t_O, x_O, y_O, z_O) sont précisément les coefficients du quadri-vecteur (T_O, X_O, Y_O, Z_O).

1. Ce qui permet de considérer l'espace-temps comme un espace vectoriel, en identifiant un point P avec le vecteur $\overrightarrow{AP}$

Attention : si tous les référentiels ont des bases, seulement certaines bases spéciales peuvent être utilisées dans des référentiels.

Comment parler des événements ? On peut, bien sûr, choisir un observateur et identifier un événement avec son quadrivecteur dans le référentiel de cet observateur. Cette solution est peu satisfaisante, puisqu'elle nous oblige à choisir un référentiel spécial auquel nous donnons une signification particulière, alors que le principe fondamental de la relativité restreinte est que toutes les référentiels se valent. Mais il est aussi possible avec les éléments ci-dessus de donner une description de la relativité restreinte sans référentiel distingué.

Récapitulons :

1) L'espace-temps, est un espace *affine* [2] à 4 dimensions dont les points représentent des événements. Il existe indépendamment de la choix du référentiel et du système de coordonnées. Nous l'appellerons ET.
2) Chaque observateur, O, mesure l'espace-temps utilisant son propre référentiel, consistant en le choix d'une origine A et d'une base de ET, (t_O, x_O, y_O, z_O). (ET est un espace vectoriel après choix de A.)
3) Soit $v \in ET$ un événement et soient (T_O, X_O, Y_O, Z_O) les coordonnées de v dans la base (t_O, x_O, y_O, z_O). L'observateur O verra l'événement v en un temps T_O et une position (X_O, Y_O, Z_O). *Le temps et la position d'un événement dans le référentiel de O sont simplement ses coordonnées dans cette base particulière.*

Une question naturelle se pose :

Question. Y a-t-il des propriétés d'un vecteur $v \in ET$ qui ne dépendent pas du choix d'observateur ?

Si v est un vecteur de ET — pensons-le comme la séparation entre deux événements, A et B — alors ni le temps ni la distance représentés par v ne sont indépendants de l'observateur. Mais il y une notion intrinsèque, au moins pour les vecteurs de type temporel, c'est celui du *temps propre*.

Définition A.1. Soit $v = \overrightarrow{AB}$ un vecteur de ET. Le temps propre de v est *le temps vécu par un observateur qui voyage de A et B sans accélération.*

De nombreuses expériences ont établie comme donnée expérimentale la relation suivante entre le temps propre $TP(v)$ d'un vecteur temporel et ses coordonnées (T_O, X_O, Y_O, Z_O) mesurés par un observateur O :

$$TP(v)^2 = T_O^2 - X_O^2 - Y_O^2 - Z_O^2.$$

On reconnaît dans le membre de droite la forme quadratique associée à une forme bilinéaire, auxquelles nous donnons le nom de « produit scalaire de Minkowski ».

Définition A.2. Soit $v, v' \in ET$, soit O un observateur, soit le référentiel de O donné par (t_O, x_O, y_O, z_O). Soient (T_O, X_O, Y_O, Z_O) et (T'_O, X'_O, Y'_O, Z'_O) les quadrivecteurs de v et w mesurés par O. Le produit scalaire de Minkowski est la forme bilinéaire sur ET donnée par

$$M(v, v') = T_O T'_O - X_O X'_O - Y_O Y'_O - Z_O Z'_O.$$

Sa forme quadratique associée $q_M(v) = M(v, v)$ a la propriété que pour tout v temporel

$$q_M(v) = (\text{temps propre de } v)^2.$$

Attention : il y ici un abus de notation. Puisqu'il existe des vecteurs pour lesquels $q_M(v) < 0$, la forme de Minkowski n'est pas un produit scalaire au sens des mathématiciens. C'est une forme bilinéaire de signature (1,3).

Plusieurs notions de la relativité restreinte admettent une interprétation en termes de la forme de Minkowski.

1) Un vecteur $v \in ET$ est temporel si $q_M(v) > 0$, luminaire si $q_M(v) = 0$ et spatial si $q_M(v) < 0$.
2) Les transformations de Lorentz sont des matrices P de changement de base (ou de changement de référentiel en gardant le même origine) qui laissent invariante la forme de Minkowski.

2. c'est-à-dire, un espace qui devient un espace vectoriel après choix d'une origine.

Annexe B

Appendice : Les coniques et quadriques

On va maintenant appliquer les résultats précédents à l'étude des coniques et des quadriques.

Jusqu'à la fin de ce paragraphe, on se place dans l'espace affine $\mathbb{R}^2$ ou $\mathbb{R}^3$, muni de son repère orthonormé usuel.

Définition B.1. Une *conique* est le lieu géométrique de $\mathbb{R}^2$ défini par une équation de la forme

$$ax^2 + by^2 + 2cxy + dx + ey + f = 0,$$

où au moins un des termes quadratiques est non nul.

Une *quadrique* est le lieu géométrique de $\mathbb{R}^3$ défini par une équation de la forme

$$ax^2 + by^2 + cz^2 + 2dxy + 2exz + 2fyz + \alpha x + \beta y + \gamma z + \delta = 0,$$

où au moins un des termes quadratiques est non nul.

On veut classer les différents types de coniques et quadriques. Puisque l'on veut conserver le lieu géométrique, on ne s'autorise qu'à faire des changements de variables qui remplacent le repère canonique par un autre repère orthonormée (pour les produit scalaire canonique), donc des translations ou des isométries (i.e. qui conservent les distances et les angles non orientés).

Ceci est nécessaire : en effet, l'équation

$$\frac{x^2}{9} + \frac{y^2}{4} = 1$$

représente une ellipse de centre O, alors que le changement de variables $x' = \frac{x}{3}$ et $y' = \frac{y}{2}$ donne l'équation

$$x'^2 + y'^2 = 1$$

représente le cercle unité, qui n'est pas le même lieu géométrique.

Cas des coniques

Considérons la conique

$$ax^2 + by^2 + 2cxy + dx + ey + f = 0,$$

et soit

$$q : \mathbb{R}^2 \to \mathbb{R}, (x,y) \mapsto ax^2 + by^2 + 2cxy.$$

D'après le théorème 3.4.9, il existe une base orthonormée (v_1, v_2) qui est ϕ_q-orthogonale, donc q-orthogonale.

Dans le nouveau repère (O, v_1, v_2), l'équation de la conique s'écrit

$$a'x'^2 + c'y'^2 + d'x + e'y + f' = 0.$$

On se débarrasse ensuite d'un ou deux termes linéaires en complétant les carrés et en effectuant une translation d'origine.

Après éventuellement permutation des nouvelles variables ou changements du type $X \leftrightarrow -X$ ou $Y \leftrightarrow -Y$ on obtient une équation d'un des types suivants, selon la signature de q, pour des réels $U, V > 0$.

1) Signature $(2,0)$ ou $(0,2)$:
   (a) $\frac{X^2}{U^2} + \frac{Y^2}{V^2} = 1$: c'est une ellipse (ou un cercle si $U = V$).
   (b) $\frac{X^2}{U^2} + \frac{Y^2}{V^2} = 0$: c'est le point $(X,Y) = (0,0)$.
   (c) $\frac{X^2}{U^2} + \frac{Y^2}{V^2} = -1$: c'est l'ensemble vide.
2) Signature $(1,1)$:
   (a) $\frac{X^2}{U^2} - \frac{Y^2}{V^2} = 1$: c'est une hyperbole.
   (b) $\frac{X^2}{U^2} - \frac{Y^2}{V^2} = 0$: c'est la réunion des deux droites d'équation $X/U = Y/V$ et $X/U = -Y/V$.
3) Signature $(1,0)$ ou $(0,1)$: $Y^2 = 2\lambda X$: c'est une parabole.

Exemple B.2. Considérons la conique d'équation

$$3x^2 - 3y^2 + 8xy + 6\sqrt{5}x + 2\sqrt{5}y + 5 = 0.$$

Soit $q : \mathbb{R}^2 \to \mathbb{R}, \begin{pmatrix} x \\ y \end{pmatrix} \mapsto 3x^2 - 3y^2 + 8xy$.

Sa matrice représentative dans la base canonique est

$$\begin{pmatrix} 3 & 4 \\ 4 & -3 \end{pmatrix}.$$

D'après un exemple précédent, une base orthonormée qui est aussi q-orthogonale est donnée par

$$\frac{1}{\sqrt{5}}\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{5}}\begin{pmatrix} 1 \\ -2 \end{pmatrix},$$

les vecteurs étant respectivement des vecteurs propres pour 5 et -5.

Soient x', y' les coordonnées dans cette nouvelle base. On a donc

$$\begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{\sqrt{5}}\begin{pmatrix} 2 & 1 \\ 1 & -2 \end{pmatrix}\begin{pmatrix} x' \\ y' \end{pmatrix} = \frac{1}{\sqrt{5}}\begin{pmatrix} 2x' + y' \\ x' - 2y' \end{pmatrix}.$$

Par construction de cette base, la forme q dans cette base s'écrit

$$5x'^2 - 5y'^2.$$

Elle est donc de signature $(1,1)$, et on a donc une hyperbole (sauf cas dégénéré).

On a alors

$$5x'^2 - 5y'^2 + 10x' + 10y' + 5 = 0,$$

soit

$$x'^2 - y'^2 + 2x' + 2y' + 1 = 0.$$

On a donc

$$(x' + 1)^2 - (y' - 1)^2 + 1 = 0.$$

En posant $X = x' - 1, Y = y' - 1$, on obtient $X^2 - Y^2 = -1$. En posant $X' = Y$ et $Y' = X$, on obtient finalement l'équation réduite de l'hyperbole

$$X'^2 - Y'^2 = 1.$$

Cas des quadriques

Comme précédemment, on se ramène au cas d'une équation sans termes croisés, et on se débarasse d'un ou plusieurs termes linéaires.

Après éventuellement permutation des nouvelles variables ou changements du type $X \leftrightarrow -X$, $Y \leftrightarrow -Y$, $Z \leftrightarrow -Z$, et éventuellement une nouvelle translation-rotation, on obtient la classification suivante :

1) Signature $(3,0)$ ou $(0,3)$. On obtient 3 cas :
   (a) $\frac{X^2}{U^2}+\frac{Y^2}{V^2}+\frac{Z^2}{W^2}=1$: c'est un ellipsoïde.
   (b) $\frac{X^2}{U^2}+\frac{Y^2}{V^2}+\frac{Z^2}{W^2}=0$: c'est le point $(X,Y,Z)=(0,0,0)$.
   (c) $\frac{X^2}{U^2}+\frac{Y^2}{V^2}+\frac{Z^2}{W^2}=-1$: c'est l'ensemble vide.
2) Signature $(2,1)$ ou $(1,2)$. On obtient 3 cas :
   (a) $\frac{X^2}{U^2}+\frac{Y^2}{V^2}-\frac{Z^2}{W^2}=-1$: c'est un hyperboloïde à deux nappes.
   (b) $\frac{X^2}{U^2}+\frac{Y^2}{V^2}-\frac{Z^2}{W^2}=1$: c'est un hyperboloïde à une nappe.
   (c) $\frac{X^2}{U^2}+\frac{Y^2}{V^2}=\frac{Z^2}{W^2}$: c'est un cône.
3) Signature $(2,0)$ ou $(0,2)$. On obtient quatre cas :
   (a) $\frac{X^2}{U^2}+\frac{Y^2}{V^2}=1$: c'est un cylindre elliptique.
   (b) $\frac{X^2}{U^2}+\frac{Y^2}{V^2}=0$: c'est la droite $X=Y=0$.
   (c) $\frac{X^2}{U^2}+\frac{Y^2}{V^2}=-1$: c'est l'ensemble vide.
   (d) $\frac{X^2}{U^2}+\frac{Y^2}{V^2}=\frac{Z}{W}$: c'est un paraboloïde elliptique.
4) Signature $(1,1)$:
   (a) $\frac{X^2}{U^2}-\frac{Y^2}{V^2}=1$: c'est un cylindre hyperbolique.
   (b) $\frac{X^2}{U^2}-\frac{Y^2}{V^2}=0$: c'est la réunion des deux plans d'équation $X/U-Y/V=0$ et $X/U+Y/V=0$.
   (c) $\frac{X^2}{U^2}-\frac{Y^2}{V^2}=-1$: c'est l'ensemble vide.
   (d) $\frac{X^2}{U^2}-\frac{Y^2}{V^2}=\frac{Z}{W}$: c'est un paraboloïde hyperbolique.
5) Signature $(1,0)$ ou $(0,1)$:
   (a) $X^2=2pY$: cylindre parabolique.
   (b) $X^2/U^2=1$: réunion de deux plans parallèles d'équation $X=1$ et $X=-1$.
   (c) $X^2/U^2=0$: plan $X=0$.
   (d) $X^2/U^2=-1$: ensemble vide.

Exemple B.3. Soit la quadrique

$$x^2+y^2+z^2+2xy+2xz+2yz+\sqrt{3}x+\sqrt{3}y+2=0.$$

Soit $q:\mathbb{R}^3\to\mathbb{R}, \begin{pmatrix}x\\y\end{pmatrix}\mapsto x^2+y^2+z^2+2xy+2xz+2yz$.

Sa matrice représentative dans la base canonique est

$$\begin{pmatrix}1&1&1\\1&1&1\\1&1&1\end{pmatrix}.$$

D'après un exemple précédent, une base orthonormée qui est aussi q-orthogonale est donnée par

$$\frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}, \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}, \sqrt{\frac{2}{3}}\begin{pmatrix}\frac{1}{2}\\\frac{1}{2}\\-1\end{pmatrix},$$

ces vecteurs étant respectivement des vecteurs propres pour $1,0$ et 0.

Soient x',y',z' les coordonnées dans cette nouvelle base.

Par construction de cette base, la forme q dans cette base s'écrit

$$3x'^2.$$

Elle est donc de signature $(1,0)$, et on a donc un cylindre parabolique, un ensemble vide, un plan ou une réunion de deux plans.

On vérifie que l'équation de cette quadrique dans cette base est

$$3x'^2+2x'+\sqrt{2}z'+2=0,$$

soit

$$3(x'+\frac{1}{3})^2+\sqrt{2}z'+\frac{5}{3}=0.$$

Si on pose $X=x'+\frac{1}{3},Y=z'+\frac{5}{3\sqrt{2}},Z=y'$, on obtient

$$X^2=-\sqrt{2}Y.$$

Annexe C

Appendice : Formes hermitiennes

Dans beaucoup d'applications en physique — notamment en mécanique quantique, mais pas exclusivement — nous avons besoin d'utiliser des espaces complexes. Par exemple la fonction d'onde qui représente un particule dans la représentation de Schrödinger est un élément de $\mathscr{C}^0(\mathbb{R}^3,\mathbb{C})$, c'est-à-dire une fonction complexe sur l'espace $\mathbb{R}^3$.

Mais si on essaie de définir une notion de longueur sur un espace complexe V en utilisant des formes bilinéaires complexes on se rend rapidement compte que c'est impossible. En effet, aucune forme bilinéaire complexe φ ne peut avoir une forme quadratique associée qui est réelle positive partout, puisque si

$$\varphi(v,v) > 0$$

alors on a

$$\varphi(iv,iv) = i^2\varphi(v,v) = -\varphi(v,v) < 0.$$

Par contre, on sait que la fonction

$$f(z) = \overline{z}z$$

est partout réelle et positive sur $\mathbb{C}$: d'ailleurs, la distance euclidienne sur $\mathbb{C}$, vue comme un $\mathbb{R}$-espace vectoriel, est donnée par

$$d(z_1,z_2) = \sqrt{(\overline{z_1-z_2})(z_1-z_2)}.$$

Nous allons donc essayer de définir des distances sur des espaces complexes en utilisant des formes *hermitiennes*, c'est-à-dire des fonctions de deux variables se comportant comme la fonction

$$(z_1,z_2) \mapsto \overline{z_1}z_2.$$

Définition C.1. Soit V un espace vectoriel complexe. Une fonction

$$h : V \times V \to \mathbb{C}$$

est une forme hermitienne si et seulement si

1) $h(x+y,z) = h(x,z) + h(y,z)$
2) $h(x,\lambda y) = \lambda h(x,y)$
3) $h(x,y) = \overline{h(y,x)}$.

Notez qu'il résulte de (3) que $h(x,x)$ est réel pour tout x. Nous avons par ailleurs que $h(\lambda x,y) = \overline{\lambda}h(x,y)$.

Exemples C.2.

1) La forme h définie sur $\mathbb{C}^n \times \mathbb{C}^n$ par

$$h\left(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}\right) = \sum \overline{x_i} y_i$$

est une forme hermitienne. Celle-ci s'appelle la forme hermitienne canonique sur $\mathbb{C}^n$.

2) La forme h définie sur $\mathscr{C}^0([a,b],\mathbb{C}) \times \mathscr{C}^0([a,b],\mathbb{C})$ par

$$h(f,g) = \int_a^b \overline{f}(x) g(x) dx$$

est une forme hermitienne sur $\mathscr{C}^0([-1,1],\mathbb{C})$.

3) La forme h définie sur $M_n(\mathbb{C}) \times M_n(\mathbb{C})$ par

$$h(M,N) = \mathrm{Tr}({}^t\overline{M}N)$$

est une forme hermitienne.

On dit qu'une forme hermitienne h est définie positive si pour tout $x \in V \setminus 0_V$ nous avons que

$$h(x,x) > 0.$$

Définition C.3. Un espace hermitien (V,h) est la donnée d'un espace vectoriel complexe V et d'une forme hermitienne h, définie positive sur E.

Bien sûr, tout ce que nous avons fait pour les formes bilinéaires symétriques peut aussi se faire pour les formes hermitiennes.

Proposition C.4. *Soit E un espace vectoriel complexe et soit h une forme hermitienne sur E. Soit $\mathbf{e} = (e_1 \dots e_n)$ une base pour E et soit M la matrice définie par $M_{i,j} = h(e_i, e_j)$. La matrice M est appelée la matrice de h dans la base $\mathbf{e}$. Soient $x,y \in E$ et soient $\underline{X}, \underline{Y}$ leurs vecteurs de coordonnées dans la base $\mathbf{e}$. Alors nous avons*

$$h(x,y) = {}^t\overline{\underline{X}} M \underline{Y}$$

Proposition C.5 (Règles de changement de la base)**.** *Soit E un espace vectoriel complexe et soit h une forme hermitienne sur E. Soient $\mathbf{e} = (e_1 \dots e_n)$ et $\mathbf{f} = (f_1 \dots f_n)$ deux bases pour E. Soit M la matrice de h dans la base $\mathbf{e}$ et soit N la matrice de h dans la base $\mathbf{f}$. Soit P la matrice de passage de $\mathbf{e}$ vers $\mathbf{f}$. Alors*

$$N = {}^t\overline{P} M P.$$

Nous notons que condition (3) de la définition des formes hermitiennes implique la proposition suivante.

Proposition C.6. *Soit h une forme hermitienne sur un espace complexe V. Soit M la matrice de h dans une base $\mathbf{e}$. Alors*

$${}^t\overline{M} = M$$

Cette proposition inspire la définition suivante.

Définition C.7. Soit M une matrice complexe. L'adjointe M^* de M est la matrice définie par

$${}^t\overline{M} = M^*.$$

Définition C.8. Une matrice complexe M de taille $n \times n$ est dite *hermitienne* (ou *auto-adjointe*) si et seulement si

$$M^* = {}^t\overline{M} = M.$$

Nous pouvons définir le longueur d'un vecteur et la distance entre deux éléments dans un espace hermitien comme pour un espace préhilbertien réel.

Définition C.9. Soit (V,h) un espace hermitien et soient $v,w \in V$. On définit la longueur de v par

$$\|v\| = \sqrt{h(v,v)},$$

et la distance entre v et w par $d(v,w) = \|v-w\|$.

Avec cette notion de distance et de longueur, une version du procédé de Gram-Schmidt et la projection orthogonale sont valables aussi sur des espaces hermitiens.

Proposition C.10 (Projection orthogonale dans les espaces hermitiens.)**.** *Soit (V,h) un espace hermitien et soit $W \in E$ un sous-espace vectoriel de dimension finie. Soit $(v_1,\ldots,v_n)$ une base orthonormée pour W. Alors pour tout $v \in V$ on définit la projection orthogonale de v sur W par*

$$p_W(v) = \sum_{i=1}^{n} h(v_i,v)v_i.$$

La projection $p_W(v)$ est alors l'élément de W qui minimise la distance $d(v,w)$ lorsque w parcourt W.

La démonstration est identique à celle donnée dans le cas des espaces préhilbertiens réels.

Proposition C.11. *Soit (V,h) un espace hermitien de dimension finie et soit $\mathbf{e} = (e_1,\ldots,e_n)$ une base. On construit une nouvelle base $(v_1,\ldots,v_n)$ récursivement par l'algorithme suivant :*

1) *On pose $v_1 = \frac{e_1}{\|e_1\|}$.*
2) *La famille $(v_1,\ldots,v_k)$ étant construite, nous posons*

$$f_{k+1} = e_{k+1} - \sum_{i=1}^{k} h(v_i,e_{k+1})v_i.$$

3) *On pose $v_{k+1} = \frac{f_{k+1}}{\|f_{k+1}\|}$.*
4) *On a maintenant construit $(v_1,\ldots,v_{k+1})$ et on revient à l'étape (2) pour construire v_{k+1}.*

La base de (V,h) ainsi construite est orthonormée.

La démonstration est identique à celle donnée dans le cas des espaces préhilbertiens réels.

Remarque C.12 (Notation bra-ket)**.** C'est une notation qui permet de retrouver plus facilement certains résultats lorsqu'on travaille avec des produits scalaires hermitiens et est particulièrement adaptée aux calculs en mécanique quantique.

Si on note $<v|w>$ le produit scalaire canonique de deux vecteurs v et w dans $\mathbb{C}^n$, on note $|w>$ le vecteur colonne des coordonnées de w et $<v| = |v>^*$ le vecteur ligne obtenu par transposition et conjugaison, alors on peut en quelque sorte séparer la notation du produit scalaire en son milieu :

$$<v|w> = <v||w>$$

Si $(e_1,\ldots,e_n)$ est une base orthonormée, alors la relation :

$$v = \sum_i <e_i,v> e_i$$

s'écrit :

$$|v> = \sum_{i=1}^{n} <e_i|v> |e_i> = \sum_{i=1}^{n} |e_i><e_i|v> = (\sum_{i=1}^{n} |e_i><e_i|)|v>$$

autrement dit le calcul des coordonnées dans une base orthonormale s'écrit comme la décomposition de l'application identité de $\mathbb{C}^n$

$$\mathbf{1}_n = \sum_{i=1}^{n} |e_i><e_i|$$

On ne peut pas se tromper de sens ci-dessus car $< e_i||e_i >= 1$ est un scalaire. La projection p sur un sous-espace vectoriel de base orthonormée $(f_1,\ldots,f_k)$ s'écrit de manière analogue :

$$p = \sum_{i=1}^{k} |f_i >< f_i|$$

le calcul de $p(v)$ est alors obtenu par

$$p|v >= (\sum_{i=1}^{k} |f_i >< f_i|)|v >= \sum_{i=1}^{k} |f_i >< f_i|v >$$

Gram-Schmidt s'écrit :

$$|v_1 >= \frac{|e_1 >}{\|e_1\|}, \quad |f_{k+1} >= (\mathbf{1}_n - \sum_{i=1}^{k} |v_i >< v_i|)|e_{k+1} >, \quad |v_{k+1} >= \frac{f_{k+1}}{\|f_{k+1}\|}$$

Si A est une matrice hermitienne, alors on peut placer A au milieu du produit scalaire et on peut l'appliquer indifféremment à v ou w :

$$< v|Aw >=< Av|w >=< v|A|w > .$$

La généralisation de cette notation aux espaces hilbertiens complexes de dimension infinie est populaire en mécanique quantique où on manipule constamment des espaces de Hilbert et des matrices hermitiennes (les observables). Une notation qui intègre les propriétés des objets manipulés permet de simplifier le travail !

Index