Mathématiques en technologie de l'Information

Orestis Malaspinas

Table des matières

Chapitre 1

Rappel

1.1 Fonctions

Une fonction f de façon générale est un objet qui prend un (ou plusieurs) paramètres et qui lui (leur) associe un résultat

$$\text{résultat} = f(\text{paramètres}). \tag{1.1}$$

Nous pouvons aussi exprimer cette notion de la manière suivante. Considérons deux ensembles A et B. Supposons qu'à chaque élément $x \in A$ est associé un élément dans B que nous notons par $f(x)$. Alors on dit que f est une fonction ou une application (de A dans B). A ce niveau A et B sont arbitraires mais dans la suite nous allons nous intéresser surtout du cas où $A \subseteq \mathbb{R}$. A est le *domaine de définition* de f. Les valeurs de f constituent les *images* de x.

Illustration 1 *(Fonctions, généralités)*

1. La tension U est une fonction de la résistance R et du courant I

$$U = f(R, I) = R \cdot I. \tag{1.2}$$

2. Une fonction peut être quelque chose de beaucoup plus général (qu'on ne peut pas forcément représenter simplement avec des opérateurs mathématiques). Prenons le cas de la fonction qui pour un nombre entier x rend le prochain entier dont le nom commence par la même lettre que x.

$$f(2) = 10,\ f(3) = 13,\ \ldots \tag{1.3}$$

Dans ce cours nous allons nous intéresser à des fonctions à un seul paramètre (aussi appelé variable). Si on note la variable x et le résultat y, de façon générale on peut écrire

$$y = f(x). \tag{1.4}$$

Si par ailleurs on a une fonction g et une fonction f, on peut effectuer des compositions de fonction, qu'on note $g \circ f$, ou encore

$$y = g(f(x)). \tag{1.5}$$

Illustration 2 *(Fonctions)*

1. Soit $f(x) = 2 \cdot x$ et $g(x) = \sqrt{x}$, alors la composition des deux fonctions

$$(f \circ g)(x) = f(g(x)) = f(\sqrt{x}) = 2\sqrt{x}. \tag{1.6}$$

2. On peut composer un nombre arbitraire de fonctions. Voyons le cas avec trois fonctions $f(x) = 2x^2 + 3$, $g(x) = \cos(2 \cdot x)$, et $h(x) = 1/x$

$$f(g(h(x))) = f(g(1/x)) = f(\cos(2/x)) = 2\cos^2(2/x) + 3. \tag{1.7}$$

Pour certaines fonctions, notons les $f(x)$, on peut également définir une fonction inverse que l'on note $f^{-1}(x)$ dont la composition donne la variable de départ

$$f(f^{-1}(x)) = x. \tag{1.8}$$

Illustration 3 *(Fonction inverse)*

1. Soient $f(x) = 2 \cdot x$ et $f^{-1}(x) = x/2$, alors la composition des deux fonctions

$$f(f^{-1}(x)) = f(x/2) = 2x/2 = x. \tag{1.9}$$

2. Soient $f(x) = x^2$ et $f^{-1}(x) = \sqrt{x}$, alors la composition des deux fonctions

$$f(f^{-1}(x)) = f(\sqrt{x}) = |x|. \tag{1.10}$$

   On a donc que $\sqrt{x}$ est l'inverse de x^2 uniquement pour les réels positifs. $f(x) = x^2$ n'a pas d'inverse pour les x négatifs. On peut se convaincre qu'une fonction ne peu admettre une inverse que si elle elle satisfait la condition $x_1 \neq x_2 \rightarrow f(x_1) \neq f(x_2)$. Dans notre exemple $-1 \neq 1$ mais $(f(-1) = f(1) = 1$

1.2 Domaine de définition

Définition 1 *(Domaine de définition)*

Le domaine de définition, noté $D \subset \mathbb{R}$, d'une fonction f, est l'ensemble de valeurs où f admet une image.

Illustration 4 *(Domaine de définition)*

1. Le domaine de définition de $f(x) = x$ est $D = \mathbb{R}$.
2. Le domaine de définition de $f(x) = 1/x$ est $D = \mathbb{R}^*$.
3. Le domaine de définition de $f(x) = \sqrt{x+1}/(x-10)$ est $D = [-1; 10[\cup]10; \infty[$.

1.3 Limites

Soit f une fonction et $D \subseteq \mathbb{R}$ non-vide et soient a et b deux réels.

1.3.1 Limite

Définition 2 *(Limite)*

Pour f définie en D, on dit que b est la limite de x en a si si au fur et à mesure que x se rapproche de a, $f(x)$ se rapproche de b et nous notons $\lim_{x\to a} f(x) = b$. C'est-à-dire pour tout voisinage de b qui contient toutes les valeurs de $f(x)$ nous avons un voisinage de a qui contient les valeurs de x (suffisamment proches de a).

La définition mathématique plus stricte est:

Pour tout $\varepsilon > 0$, il existe un $\delta > 0$, tel que, pour tout $x \in D$ tel que $|x - a| < \delta$, on ait $|f(x) - a| < \varepsilon$.

Ou encore quand le but est d'écrire ça de la façon la plus compacte possible

$$\forall \varepsilon > 0, \exists \delta > 0 \mid \forall x \in D,\ |x - a| < \delta \Rightarrow |f(x) - b| < \varepsilon. \tag{1.11}$$

Remarque 1

Il n'est pas nécessaire que $a \in D$. Mais si c'est le cas et donc f est définie en a alors on a $\lim_{x\to a} f(x) = f(a)$.

Illustration 5 *(Limite)*

Si $f(x) = x$, alors $\lim_{x\to 0} f(x) = 0$.

Définition 3 *(Limite, asymptote)*

Pour f définie en D, on dit que la limite de $f(x)$ en a est égale à l'infini si pour tout $c > 0$ l'intervalle $[c; \infty[$ contient toutes les valeurs de $f(x)$ pour x suffisamment proche de a. On dit aussi que f tend vers l'infini.

Illustration 6 *(Limite, asymptote)*

Si $f(x) = 1/x^2$, alors $\lim_{x \to 0} f(x) = \infty$.

1.3.2 Limite à gauche, limite à droite

Il est possible que le comportement de certaines fonctions soit différent selon qu'on approche a par la gauche ou par la droite (i.e. $f(x) = 1/x$, pour $a = 0$).

On note la limite à droite $\lim_{x \to a^+} f(x)$ ou $\lim_{x \to a, x > a} f(x)$ et $\lim_{x \to a^-} f(x)$ ou $\lim_{x \to a, x < a} f(x)$ la limite à gauche de la fonction f en a.

Si la fonction f admet une limite en a, alors les deux limites sont égales.

Illustration 7 *(Limite à gauche/droite)*

Si $f(x) = 1/x$, alors $\lim_{x \to 0^+} f(x) = \infty$ et $\lim_{x \to 0^-} f(x) = -\infty$.

1.3.3 Comportement asymptotique

Dans certains cas il peut être intéressant d'étudier le comportement des fonctions quand $x \to \pm\infty$. Dans ces cas-là on dit qu'on s'intéresse au comportement *asymptotique* d'une fonction. Ce concept est particulièrement pertinent quand on étudie une fonction qui a la forme d'une fraction

$$h(x) = \frac{f(x)}{g(x)}. \tag{1.12}$$

Si on s'intéresse au comportement à l'infini de cette fonction on va prendre sa "limite" lorsque $x \to \infty$

$$\lim_{x \to \infty} h(x) = \lim_{x \to \infty} \left(\frac{f(x)}{g(x)} \right). \tag{1.13}$$

Un exemple peut être $f(x) = x - 1$, $g(x) = x + 1$ et donc $h(x) = (x-1)/(x+1)$

$$\lim_{x \to \infty} \frac{x-1}{x+1} = \lim_{x \to \infty} \frac{x(1 - 1/x)}{x(1 + 1/x)} = 1. \tag{1.14}$$

De même quand on a $f(x) = 3x^4 - 5x^3 + 1$, $g(x) = 1$ et donc $h(x) = 3x^4 - 5x^3 + 1$. Il vient donc

$$\lim_{x \to \infty} 3x^4 - 5x^3 + 1 = \lim_{x \to \infty} 3x^4 \left(1 - \frac{5}{3x} + \frac{1}{3x^4} \right) = \infty. \tag{1.15}$$

Si nous compliquons un peu l'exemple et que nous avons $f(x) = x^3 + 3x^2 + 1$, $g(x) = x^2$ et donc $h(x) = (x^3 + 3x^2 + 1)/x^2$

$$\lim_{x\to\infty} (x^3 + 3x^2 + 1)/x^2 = \lim_{x\to\infty} x = \infty. \tag{1.16}$$

Un cas encore un peu plus complexe serait $f(x) = 3x^3 + 1$, $g(x) = 4x^3 + 2x^2 + x$

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = \lim_{x\to\infty} \frac{3x^3(1 + 1/3x^3)}{4x^3(1 + 1/2x^+1/4x^2)} = \frac{3}{4}. \tag{1.17}$$

Ce genre d'estimations est important en informatique lors de l'analyse de performance des algorithmes. On peut prendre l'exemple des algorithmes de tri "bubble sort" et "quick sort". Leur complexité respective moyenne est de n^2 et de $n\log(n)$, quand n est le nombre d'éléments de la chaîne à trier. Si on fait le rapport pour de ces deux complexités on a

$$\lim_{n\to\infty} \frac{n^2}{n\log(n)} = \lim_{n\to\infty} \frac{n}{\log(n)}. \tag{1.18}$$

On peut simplement voir que ce rapport va tendre vers l'infini en dessinant la courbe $n/\log(n)$. Il existe un moyen "analytique" d'évaluer ce rapport. Tout nombre n peut s'écrire avec une précision p comme

$$n = A \cdot 10^{p-1}, \tag{1.19}$$

où p est le nombre de chiffres significatifs qu'on veut représenter, et $1 \leq A < 10$. On a également que [1]

$$\log(A) = \log\left(\frac{1+y}{1-y}\right) = 2\sum_{k=0}^{\infty} \frac{y^{2k+1}}{2k+1}, \tag{1.20}$$

avec $y = (A-1)/(A+1)$. On a finalement que

$$\log(n) = \log(A \cdot 10^{p-1}) = (p-1)\log(10) + 2\sum_{k=0}^{\infty} \frac{y^{2k+1}}{2k+1}. \tag{1.21}$$

La valeur de y étant quelque chose de proche de 0, la somme converge vite vers une valeur finie et on peut faire l'approximation

$$\log(n) \cong (p-1)\log(10), \tag{1.22}$$

pour n grand (ce qui est équivalent à p grand). On a donc que finalement le rapport $n/\log(n)$ va comme

$$\lim_{n\to\infty} \frac{n}{\log(n)} = \frac{A}{\log(10)} \cdot \lim_{p\to\infty} \frac{10^{p-1}}{(p-1)} = \frac{A}{\log(10)} \cdot \lim_{p\to\infty} \frac{10^{p-1}}{p} = \infty. \tag{1.23}$$

1. Pour ceux que ça intéresse cette série s'obtient à l'aide d'une série de Taylor.

1.4 Continuité

Définition 4 *(Continuité)*

Soit f une fonction définie sur un intervalle ouvert D contenant a. On dit que f est continue en a si et seulement si $\lim_{x \to a} f(x) = f(a)$.

Propriétés 1 *(Fonctions continues)*

Soient f et g deux fonctions continues en a et b un réel:

1. $f + g$ est continue en a.
2. bf est continue en a.
3. si $g(a) \neq 0$, f/g est continue en a.
4. $h = g \circ f$ est continue en a.

Définition 5 *(Continuité sur un intervalle)*

Une fonction f est dite continue dans un intervalle $D =]a; b[$ si et seulement si elle est continue en tout point de D. De plus, elle est continue sur $D = [a, b]$ si elle est continue sur $]a; b[$ et continue à droite en a et à gauche en b.

Théorème 1 *(Valeurs intermédiaires)*

Soit f une fonction continue sur D, et a, b deux points contenus dans D tels que $a < b$ et $f(a) < f(b)$, alors

$$\forall y \in [f(a); f(b)],\ \exists\ c \in [a, b] | f(c) = y. \tag{1.24}$$

Nous pouvons bien sûr énoncer un résultat similaire dans le cas $f(a9 > f(b)$.

1.5 Dérivées

Définition 6 *(Dérivée en un point)*

Soit f une fonction définie sur D et $a \in D$. On dit que f est dérivable en a s'il existe un b (appelé la dérivée de f en a) tel que

$$\begin{aligned} &\lim_{h\to 0} \frac{f(a+h)-f(a)}{h} = b, \text{ ou} \\ &\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = b. \end{aligned} \tag{1.25}$$

Définition 7 *(Dérivée sur un intervalle)*

Si f est dérivable en tout point de $D =]a; b[$, alors on définit f' la fonction dérivée de f dans l'intervalle D qui associe en tout point x de D la valeur dérivée de f.

Propriété 1

Si f est dérivable en a alors f est continue en a.

Propriétés 2

Soient f et g deux fonctions dérivables sur D (dont les dérivées sont f' et g'), et $a \in \mathbb{R}$, alors

1. $(f+g)' = f' + g'$.
2. $(af)' = af'$.
3. $(f \cdot g)' = f'g + fg'$.
4. Si g ne s'annule pas $(f/g)' = (f'g - fg')/g^2$.
5. $(g \circ f)' = (g' \circ f) \cdot f'$, autrement dit pour $x \in D$, $(g(f(x)))' = g'(f(x)) \cdot f'(x)$.

Il existe quelques dérivées importantes que nous allons utiliser régulièrement dans la suite de ce cours. En supposons que $C \in \mathbb{R}$, nous avons

1. $f(x) = x^n$, $f'(x) = nx^{n-1}$.
2. $f(x) = e^{Cx}$, $f'(x) = Ce^{Cx}$.
3. $f(x) = \ln(x)$, $f'(x) = 1/x$.
4. $f(x) = C$, $f'(x) = 0$.
5. $f(x) = \sin(x)$, $f'(x) = \cos(x)$.
6. $f(x) = \cos(x)$, $f'(x) = -\sin(x)$.

Définition 8 *(Dérivée seconde)*

Si f' est dérivable sur D, alors sa dérivée, notée f'', est appelée la dérivée seconde de f.

1.5.1 Variation des fonctions

Propriétés 3 *(Croissance/décroissance)*

Soit f' la fonction dérivée de f sur D

1. Si $f' > 0$ sur D, alors f est croissante sur D.
2. Si $f' < 0$ sur D, alors f est décroissante sur D.
3. Si $f' = 0$ sur D, alors f est constante sur D.

Définition 9 *(Maximum/minimum local)*

Une fonction admet un maximum local (respectivement minimum local) sur un intervalle $D =]a; b[$ s'il existe un $x_0 \in D$ tel que $f(x_0) \geq f(x)$ (respectivement $f(x_0) \leq f(x)$) pour tout $x \in D$.

Propriété 2 *(Maximum/minimum)*

Soient f une fonction dérivable sur $D =]a; b[$ et $x_0 \in D$. On dit que f admet un extremum en x_0 si $f'(x_0) = 0$. De plus si $f'(x_0) = 0$ et f' change de signe en x_0 alors $f(x_0)$ est un maximum ou un minimum de f.

1.6 Etude de fonction

Effectuer l'étude de fonction de la fonction suivante

$$f(x) = \frac{x^3}{x^2 - 4}. \tag{1.26}$$

1. Déterminer le domaine de définition.
2. Déterminer la parité de la fonction. Rappel:

$$\begin{aligned} f(-x) &= f(x), \text{ paire,} \\ f(-x) &= -f(x), \text{ impaire.} \end{aligned} \tag{1.27}$$

3. Trouver les zéros de la fonction (Indication: trouver les x tels que $f(x) = 0$).

4. Trouver les éventuelles asymptotes verticales ou discontinuités, ainsi que les asymptotes affines.
5. Calculer $f'(x)$ et déterminer sa croissance et points critiques (déterminer où la fonction est croissante, décroissante, atteint un extremum, etc).
6. Faire un croquis de $f(x)$.

Chapitre 2

Optimisation

2.1 La régression linéaire

Lors d'une régression linéaire, le but est de trouver la droite, $y(x) = a \cdot x + b$, qui passe au mieux au travers d'un nuage de N points (x_i, y_i), $i = 1, ..., N$ (voir fig. 2.1).

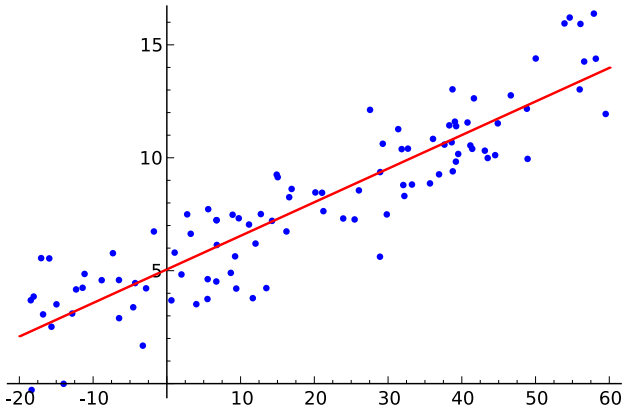

FIGURE 2.1 – Nous recherchons l'équation de la droite, en rouge, $y(x) = ax + b$ passant au plus proche des points (x_i, y_i) et bleu. Source: Wikipedia https://bit.ly/2SfiLzb

Pour déterminer l'équation de cette droite, nous devons donc trouver les coefficients a et b tels que la droite passe au plus proche des points. Nous devons d'abord définir ce que signifie mathématiquement "passe au mieux par au travers du nuage de points". Une façon de mesurer la "qualité" d'une droite est de mesurer la somme des distances au carré entre les points (x_i, y_i) et la droite $y(x) = a \cdot x + b$ pour des valeurs de a et b données, soit

$$E(a, b) = \sum_{i=1}^{N} (y(x_i) - y_i)^2. \tag{2.1}$$

Nous cherchons par conséquent à minimiser $E(a, b)$ sous la contrainte que $y(x)$ est une droite. Pour simplifier encore plus le problème mathématique, nous pouvons rajouter comme contrainte que la droite $y(x)$ passe par le point $(0, 0)$, on a donc que $y(x) = a \cdot x$ (l'ordonnée à l'origine est nulle, $b = 0$) et que

$$E(a) = \sum_{i=1}^{N} (y(x_i) - y_i)^2, \tag{2.2}$$

est indépendant de b. En résumé nous cherchons à résoudre le problème mathématique

$$\min_{a \in \mathbb{R}} E(a) = \min_{a \in \mathbb{R}} \sum_{i=1}^{N} (y(x_i) - y_i)^2, \tag{2.3}$$

$$\text{où } y(x) = a \cdot x, \quad \text{(contrainte)}. \tag{2.4}$$

On peut réécrire la fonction $E(a)$ comme

$$\begin{aligned} E(a) &= \sum_{i=1}^{N} \left(y^2(x_i) - 2 \cdot y_i \cdot y(x_i) + y_i^2 \right) = \sum_{i=1}^{N} \left(a^2 \cdot x_i^2 - 2 \cdot a \cdot x_i \cdot y_i + y_i^2 \right), \\ &= a^2 \sum_{i=1}^{N} x_i^2 + 2a \sum_{i=1}^{N} x_i y_i + \sum_{i=1}^{N} y_i^2. \end{aligned} \tag{2.5}$$

Les x_i et y_i étant connus, nous cherchons a, tel que $E(a)$ soit minimal. $E(a)$ est en fait l'équation d'une parabole: elle a la forme

$$E(a) = B \cdot a^2 - 2C \cdot a + D, \tag{2.6}$$

avec $B = \sum_{i=1}^{N} x_i^2$, $C = \sum_{i=1}^{N} x_i y_i$, et $D = \sum_{i=1}^{N} y_i^2$. B étant forcément positif cette parabole sera **convexe** et donc nous sommes assurés qu'il existe un minimum pour $E(a)$. Une façon de déterminer a, tel que $E(a)$ est minimal est d'utiliser la dérivée. On a l'équation $E'(a) = 0$ à résoudre:

$$\begin{aligned} E'(a) &= 0, \\ 2 \cdot B \cdot a - 2 \cdot C &= 0, \\ a = \frac{C}{B} &= \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}. \end{aligned} \tag{2.7}$$

Illustration 8

Soient les 4 points $(0, 0.1)$, $(1, 0.3)$, $(2, 0.3)$ et $(3, 0.4)$. La fonction d'erreur $E(a)$ s'écrit

$$E(a) = 14 \cdot a^2 - 4.2 \cdot a + 0.35. \tag{2.8}$$

On peut la représenter comme sur la fig. 2.2 et on constate qu'elle possède un minimum proche de $a = 0$.

En résolvant $E'(a) = 0$, on obtient $a = 4.2/28 = 0.15$. On a que l'équation de la droite passant par $(0, 0)$ et au plus proche de nos 4 points est

$$y(x) = 0.15 \cdot x. \tag{2.9}$$

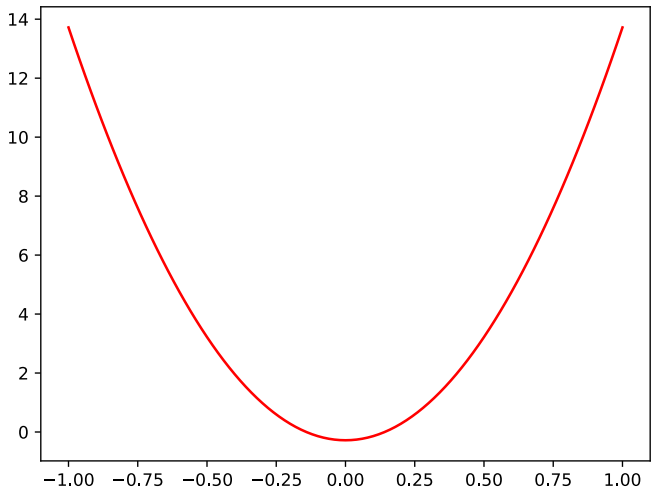

FIGURE 2.2 – La fonction $E(a) = 14a^2 - 4.2a + 0.35$ pour $a \in [-1, 1]$. On voit bien qu'elle possède un minimum proche de $a = 0$.

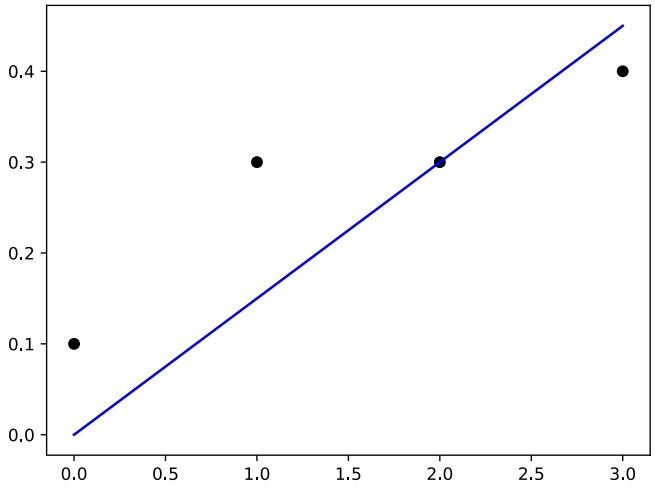

FIGURE 2.3 – Les 4 points $(0, 0.1)$, $(1, 0.3)$, $(2, 0.3)$ et $(3, 0.4)$ (en noir) et la droite obtenue par régression linéaire (en bleu).

On peut observer le résultat de la régression sur la fig. 2.3, où on voit les 4 points (en noir), ainsi que la droite obtenue (en trait bleu).

La régression linéaire est un problème **d'optimisation continu** (par opposition aux problèmes **d'optimisation discrets**). Ce genre de problème, bien que possédant un espace de recherche infini, est bien souvent plus simple à résoudre que les problèmes d'optimisation discrets, car il possède un cadre théorique mieux défini.

Pour le résoudre, nous avons commencé par construire un modèle mathématique. Nous avons défini une fonction à minimiser, $E(a)$, et ajouté une contrainte, la forme de $y(x)$. Puis, il a suffi de trouver le minimum de $E(a)$ sous la contrainte et le tour était joué.

2.2 L'optimisation mathématique

Suite à ces deux exemples, nous allons essayer de définir de façon assez théorique comment formuler mathématiquement un problème d'optimisation. Il existe deux types distincts de problèmes d'optimisation:

1. L'optimisation continue.
2. L'optimisation discrète (souvent appelée optimisation combinatoire).

Dans ce chapitre nous ne parlerons que de l'optimisation continue.

2.2.1 L'optimisation continue

L'optimisation continue ou *programme mathématique continu* est un programme d'optimisation soumis à certaines contraintes. On peut l'exprimer de la façon suivante.

Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction objectif (ou fonction de coût), on cherche $\vec{x}_0 \in \mathbb{R}^n$, tel que $f(\vec{x}_0) \leq f(\vec{x})$ pour $\vec{x}$ certaines conditions: **les contraintes**. Celles-ci sont en général des égalités strictes ou des inégalités qui peuvent s'exprimer de la façon suivante. Soient m fonctions $g_i : \mathbb{R}^n \to \mathbb{R}$

$$g_i(\vec{x}) \leq 0, \quad i = 1, ..., m. \tag{2.10}$$

Si $m = 0$ on a à faire à un problème d'optimisation sans contraintes. On peut résumer tout cela comme

$$\begin{aligned} &\min_{\vec{x} \in \mathbb{R}^n} f(\vec{x}), \\ &g_i(\vec{x}) \leq 0, \quad i = 1, ..., m, \\ &\text{pour } m \geq 0. \end{aligned}$$

Les contraintes limitent l'espace des solutions et forment un sous-ensemble, noté A, de $\mathbb{R}^n$ ($A \subseteq \mathbb{R}^n$).

Une des difficultés pour déterminer le minimum d'une fonction coût est l'existence de plusieurs minima locaux. Un **minimum local**, $\vec{x}^* \in A$, est tel que pour une

région proche de $\vec{x}^*$, on a que $f(\vec{x}) \geq f(\vec{x}^*)$. Un exemple d'une telle fonction, est une fonction de Ackley. En une dimension, elle est de la forme (voir la fig. 2.4)

$$f(x) = -20e^{-0.2*\sqrt{0.5x^2}} - e^{0.5(\cos(2\pi x))} + e + 20. \tag{2.11}$$

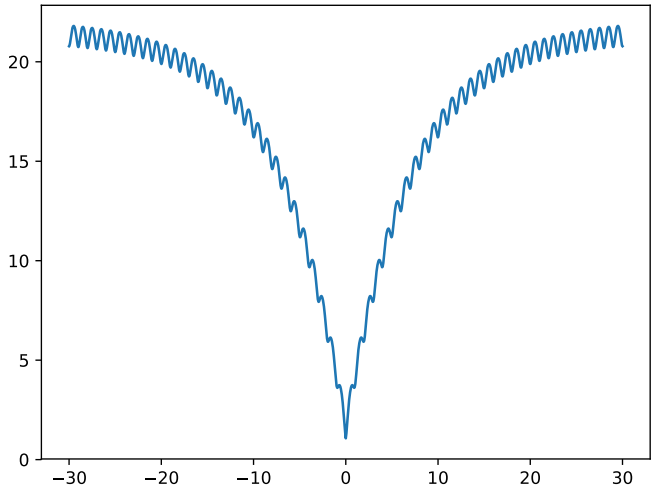

FIGURE 2.4 – La fonction d'Ackley en une dimension. Il est particulièrement compliqué de trouver le minimum global en raison de l'existence d'une multitude de minima locaux.

On constate la présence d'un grand nombre de minima locaux qui rendent la recherche du minimum global (se trouvant en $x = 0$) particulièrement compliqué à déterminer.

L'optimisation continue est très communément utilisée en apprentissage automatique (machine learning), en particulier pour optimiser les poids des réseaux de neurones.

2.3 Optimisation continue

Dans cette section, nous allons considérer des problèmes purement continus. Nous allons dans un premier temps considérer une fonction objectif, f,

$$f : D \to \mathbb{R}, \quad D \subseteq \mathbb{R}, \tag{2.12}$$

dont nous allons chercher le minimum (pour autant qu'il existe). Nous allons supposer que f est une fonction continue et dérivable.

2.3.1 Minimum local/global

Comme vous le savez, le minimum (ou le maximum) d'une fonction, se situe à un endroit où sa dérivée est nulle. On recherche donc, x, tel que

$$f'(x) = 0. \tag{2.13}$$

Mais cette contrainte sur $f'(x)$ n'est pas suffisante pour garantir de trouver un minimum. En effet, si $f'(x) = 0$, peut également vouloir dire qu'on se trouve sur un point d'inflexion ou sur un maximum. On peut assez facilement, discriminer ces deux cas, en considérant la deuxième dérivée de f. En effet, nous avons à faire à un minimum seulement si

$$f''(x) > 0. \tag{2.14}$$

Les cas où $f''(x) = 0$ est un point d'inflexion et $f''(x) < 0$ est un maximum.

Un autre problème beaucoup plus compliqué à résoudre est de déterminer un minimum **global**. En effet, comme pour la fonction de Ackley (voir la fig. 2.4), une fonction peut posséder un grand nombre de minima **locaux** (où $f'(x) = 0$ et $f''(x) > 0$) mais qui n'est pas un minimum global.

Mathématiquement un *minimum local* se définit comme x^* tel qu'il existe $\delta > 0$ et que $f(x^*) \leq f(x)$, pour $x \in [x^* - \delta, x^* + delta]$. Un *minimum global* est un x^* tel que $\forall x \in D$, $f(x^*) \leq f(x)$.

En fait, il n'existe pas de méthode pour déterminer un minimum global, pour n'importe quelle fonction. Nous somme assurés de le trouver, uniquement si f est une fonction convexe partout ($f''(x) > 0 \ \forall x$).

2.4 Algorithmes de recherche des zéros d'une fonction

Comme nous venons de le voir, lors de la recherche d'un minimum, il est nécessaire de trouver le point x^* où $f'(x^*) = 0$. Le problème est donc de déterminer les zéros de la fonction $f'(x)$. Pour avoir un maximum de généralité, nous allons considérer une fonction $g(x)$ et chercher ses zéros, soit

$$\{x \in \mathbb{R} | g(x) = 0\}. \tag{2.15}$$

Dans des cas simples (des fonctions polynomiales de degré 2 ou 3, ou des fonctions inversibles) on peut trouver analytiquement les zéros. En revanche, pour des fonctions plus complexes, ou "implicites" (on ne peut pas mettre l'équation $g(x) = 0$ sous la forme $x = ...$) la détermination des zéros est beaucoup plus difficile et nécessite l'utilisation de **méthodes itératives**. Nous allons en voir quelques unes.

2.5 Méthodes par raffinement d'intervalles

2.5.1 Méthode de la bissection

Afin de déterminer le zéro d'une fonction, une des méthodes les plus simple est la méthode de la bissection. Il s'agit de choisir deux points, a_1 et b_1, $b_1 > a_1$, tels que le signe de $g(a_1)$ et $g(b_1)$ est différent. Si cela est le cas, nous sommes assurés de l'existence d'au moins un zéro si la fonction $g(x)$ est continue (en vertu du théorème de la valeur intermédiaire). Ensuite, nous allons calculer la valeur se situant "au milieu" entre a_1 et b_1

$$c_1 = \frac{b_1 + a_1}{2}. \tag{2.16}$$

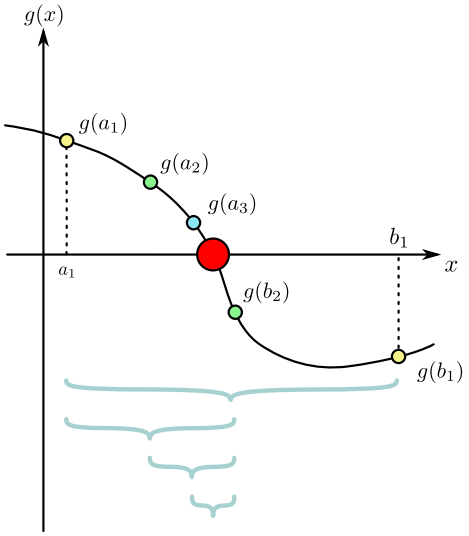

FIGURE 2.5 – Illustration de la méthode de la bissection. Source: Wikipedia https://bit.ly/2RcljBW.

Puis, nous évaluons $g(c_1)$ et si ce n'est pas un zéro, étudions son signe. Si le signe $g(c_1)$ est différent de celui de $g(a_1)$, nous remplaçons b_1 par c_1 et recommençons. Si le signe de $g(c_1)$ est différent de celui de $g(b_1)$, nous remplaçons a_1 par c_1. Nous itérons cette méthode jusqu'à ce que nous ayons atteint une valeur "suffisamment proche" (nous avons une précision acceptable pour nous) de zéro. Une façon d'exprimer "proche" est de considérer la taille de l'intervalle $b_1 - a_1$ et de le comparer avec une précision $\varepsilon > 0$ que nous aurons choisie

$$b_1 - a_1 < \varepsilon. \tag{2.17}$$

Au pire des cas, cette méthode nous rapproche de $(b_1 + a_1)/2$ du zéro à chaque itération. Après n itération, nous somme donc à une distance maximale du zéro de $(b_1 + a_1)/2^n$. On dit que cette méthode est d'ordre 1 (on divise l'intervalle de recherche par 2 et la précision par 2 à chaque itération).

Exercice 1 *(Racine de polynôme)*

Déterminer la racine du polynôme $x^4 + x^3 + x^2 - 1$ avec $a_1 = 0.5$ et $b_1 = 1$ (faire au maximum 6 itérations).

2.5.2 Méthode de la fausse position (*regula falsi*)

Une méthode un peu plus avancée est la méthode de la fausse position (voir la fig. 2.6). Dans cette méthode qui est relativement similaire à celle de la bissection, mais au lieu de diviser l'intervalle en deux parts égales à chaque itération on va choisir les point c, comme étant le point où la droite reliant $g(a_1)$ et $g(b_1)$ coupe l'axe horizontal (le zéro de la droite entre $g(a_1)$ et $g(b_1)$). Le reste de

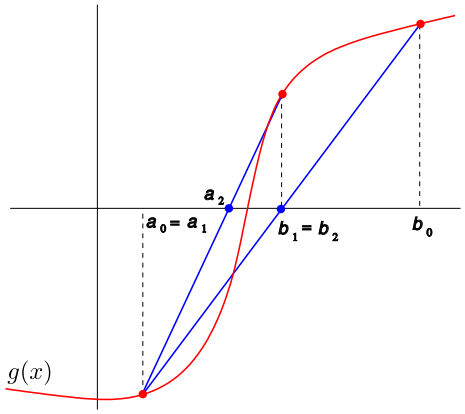

FIGURE 2.6 – Illustration de la méthode de la fausse position. Source: Wikipedia https://bit.ly/2FeHdhm.

l'algorithme reste exactement le même. On choisit deux points, a_1 et b_1, où le signe de g est différent, puis ont construit la droite passant par $g(a_1)$ et $g(b_1)$

$$y = \frac{g(b_1) - g(a_1)}{b_1 - a_1}(x - a_1) + g(a_1). \tag{2.18}$$

On cherche le point, c, où $y(c) = 0$

$$\frac{g(b_1) - g(a_1)}{b_1 - a_1}(c - a_1) + g(a_1) = 0. \tag{2.19}$$

Cette équation s'inverse aisément et on obtient

$$c_1 = a_1 - \frac{b_1 - a_1}{g(b_1) - g(a_1)} g(a_1). \tag{2.20}$$

Puis, comme pour la méthode de la bissection, on compare les signes de $g(c_1)$ avec $g(a_1)$ et $g(b_1)$ et on remplace a_1 ou b_1 par c_1 si $g(c_1)$ a un signe différent de $g(b_1)$ ou $g(a_1)$ respectivement.

Il est important de noter que si la fonction est continue, et que a_1 et b_1 sont choisis tels que $g(a_1)$ et $g(b_1)$ sont de signes opposés, alors cette méthode convergera **toujours**.

La méthode de la fausse position est plus efficace que la méthode de la bissection, elle est superlinéaire (d'ordre plus grand que un).

Exercice 2

Déterminer le zéro positif de la fonction

$$x^2 - 25 = 0, \tag{2.21}$$

à l'aide de la méthode de la fausse position.

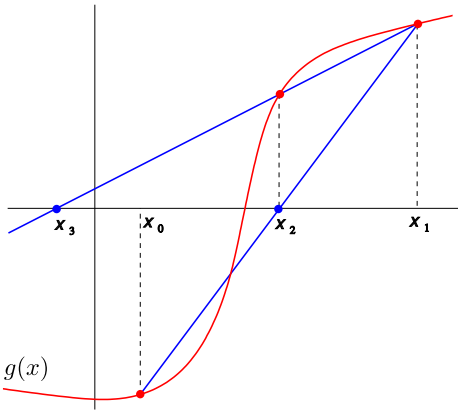

FIGURE 2.7 – Illustration de la méthode de la sécante. En partant des points x_0 et x_1, on s'approche du zéro en calculant le point x_2. Source: Wikipedia https://bit.ly/2FcAXpA.

2.5.3 Méthode de la sécante

La méthode de la sécante (voir la fig. 2.7) est très similaire à la méthode de la fausse position. La seule différence se situe dans la dernière étape de l'algorithme. Plutôt que choisir de remplacer a_1 ou b_1 par c_1, on remplace toujours la dernière valeur calculée. Ainsi après avoir choisi $a < b$, avec $g(a)$ et $g(b)$ avec des signes différents, on calcule une suite de x_i, avec $x_0 = a$, $x_1 = b$, tels que

$$x_{i+1} = x_{i-1} - \frac{x_i - x_{i-1}}{g(x_i) - g(x_{i-1})} g(x_{i-1}), \quad i \geq 2. \tag{2.22}$$

La méthode de la sécante ne garantit pas la convergence, contrairement à la méthode de la bissection et de la fausse position. En revanche elle est plus efficace, lorsque qu'elle converge, que ces deux méthodes.

Exercice 3

Déterminer le zéro positif de la fonction

$$x^2 - 25 = 0, \tag{2.23}$$

à l'aide de la méthode de la sécante.

2.5.4 Recherche de la fourchette initiale

Dans les méthodes ci-dessus, nous avons supposé que nous avions une fonction $g(x)$ continue, ainsi qu'un intervalle, $[a, b]$, avec

$$g(a) < 0, \quad g(b) > 0. \tag{2.24}$$

Mais, nous n'avons pas encore vu de méthode pour déterminer les valeur de la fourchette a, b.

Remarque 2

On peut procéder de façon très similaire pour $[a, b]$ tel que

$$g(a) > 0, \quad g(b) > 0. \tag{2.25}$$

Il suffit de prendre remplacer $g(x) \to -g(x)$.

Les méthodes pour déterminer la fourchette initiales sont également des *méthodes itératives*.

La plus simple qu'on puisse imaginer est de partir d'un point initial a (choisi aus hasard par exemple). On suppose que $g(a) < 0$ (sinon voir la remarque ci-dessus). Puis on choisir deux *hyperparamètres*: δx et k [1]. Ensuite on calcule $b = a + k \cdot \delta x$. Si $f(b) > 0$, on a terminé. Sinon on recommence avec $k \to 2 \cdot k$ et $b \to k \cdot b$.

2.6 Méthodes de descentes locales

L'idée de ce type de méthodes est, contrairement aux méthodes de la section précédente, d'utiliser des connaissances *locales* que nous pouvons avoir sur la fonction. Cette connaissance locale a en général comme effet une *convergence* plus rapide de l'algorithme de recherche de zéros.

2.6.1 Méthode de Newton (ou *Newton-Raphson*)

La méthode de Newton est également une méthode itérative, qui nécessite que la fonction $g(x)$ soit non seulement continue mais également dérivable. Revenons sur la méthode de la sécante. Il s'agissait de choisir deux points, $a < b$, et de déterminer la droite, $y(x)$, passant par $g(a)$ et $g(b)$,

$$y = \frac{g(b) - g(a)}{b - a}(x - a) + g(a).$$

Il se trouve que $g(b) - g(a)/(b - a)$ n'est autre qu'une approximation avec une formule de *différences finies* de la dérivée de g et a, $g'(a)$. Si la fonction g est dérivable, on peut simplement remplacer ce terme par $g'(a)$ et on obtient

$$y = g'(a)(x - a) + g(a). \tag{2.26}$$

Puis on détermine c, tel que $y(c) = 0$

$$0 = g'(a)(c - a) + g(a), \tag{2.27}$$

et on obtient

$$c = a - \frac{g(a)}{g'(a)}. \tag{2.28}$$

1. Leur valeur est un peu arbitraire, souvent $\delta x = 0.01$ et $k = 2$.

On peut donc généraliser l'algorithme. En partant d'un point $x_0 = a$, on construit la suite

$$x_{i+1} = x_i - \frac{g(x_i)}{g'(x_i)},\ i \geq 0. \tag{2.29}$$

On s'arrête lorsque le zéro est déterminé avec une précision suffisante, ou que la variation entre deux itérations successives est assez petite. Ce qui revient à choisir un $\varepsilon > 0$, tel que

$$|g(x_n)| < \varepsilon, \quad |x_n - x_{n-1}| < \varepsilon. \tag{2.30}$$

Lorsque qu'elle converge la méthode de Newton est la plus efficace de toutes celles que nous avons vues. On dit qu'elle est d'ordre 2. En revanche les contraintes pour sa convergence sont plus strictes que pour les méthodes vues précédemment.

Remarque 3 *(non-convergence ou convergence lente)*

Il y a un certain nombre de cas où la méthode de Newton ne converge pas.

1. S'il existe un n tel que $g'(x_n) = 0$ alors la suite diverge.
2. La suite peut entrer dans un cycle.
3. La dérivée est mal définie proche du zéro (ou sur le zéro).
4. Elle peut converger très lentement si la dérivée de la fonction est nulle sur le zéro.
5. A chaque point de départ ne correspond qu'un zéro. Si la fonction possède plusieurs zéros, il n'y a pas moyen de le savoir avec un seul point de départ. Il faut alors en essayer plusieurs.

Exercice 4

Déterminer le zéro de la fonction

$$x^2 - 25 = 0, \tag{2.31}$$

à l'aide de la méthode de Newton.

2.6.2 Résumé

A l'aide des méthodes vues ci-dessus, on peut déterminer un zéro d'une fonction (s'il existe). Ces méthodes sont également utilisables pour calculer le minimum d'une fonction comme nous l'avons discuté plus haut. Il suffit de remplacer $g(x)$ par $f'(x)$ et le tour est joué.

Exercice 5

Écrire l'algorithme de Newton pour le cas de la minimisation d'une fonction $f(x)$ quelconque, mais continûment dérivable 2 fois.

2.7 En plusieurs dimensions

Quand notre fonction de coût dépend de plusieurs arguments, on dit que c'est une fonction *multivariée*, $f(\vec{x})$, avec $\vec{x} \in \mathbb{R}^n$.

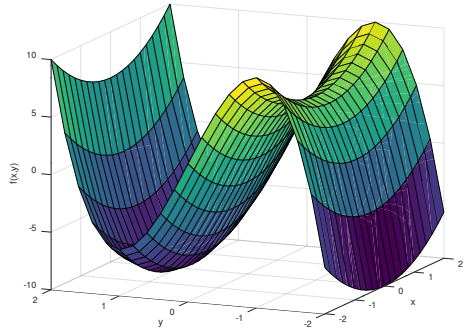

FIGURE 2.8 – La fonction $f(x, y) = x^2 - y^2$ est une fonction à plusieurs variable.

On peut également l'écrire de façon plus explicite (et aussi plus longue) comme

$$f(\vec{x}) = f(x_1, x_2, ..., x_n). \tag{2.32}$$

Bien que la fonction de coût prenne en argument plusieurs variables, elle retourne uniquement un réel

$$f : \mathbb{R}^n \to \mathbb{R}. \tag{2.33}$$

Illustration 9 *(Régression linéaire)*

Dans le cas de la régression linéaire, si la droite ne passe pas par l'origine, nous avons que la fonction de coût qui dépend de deux variables, a, et b (et plus uniquement de a)

$$f(a, b) = \frac{1}{N} \sum_{i=1}^{N} (a \cdot x_i + b - y_i)^2 . \tag{2.34}$$

2.7.1 Les dérivées en plusieurs dimensions

La dérivé d'une fonction à une seule variable peut se représenter comme

$$f'(a) = \frac{\mathrm{d}f}{\mathrm{d}x}(a) = \lim_{\mathrm{d}x \to 0} \frac{f(a + \mathrm{d}x) - f(a)}{\mathrm{d}x}. \tag{2.35}$$

La notation ici n'est pas tout à fait usuelle. L'idée est de se rappeler que ce $\mathrm{d}x$ est une toute petite variation de x, et $\mathrm{d}f$, une toute petite variation de f en a. On voit immédiatement que cette quantité est la pente de f en a. Lorsque nous étudions une fonction à plusieurs variables, nous pouvons faire le même raisonnement pour chaque variable indépendamment. Ainsi, nous calculons sa dérivée dans chacune des directions x, y, ...

Cette vision de la dérivée comme une variation de f, $\mathrm{d}f$, divisée par une petite variation de x, $\mathrm{d}x$, permet d'avoir une interprétation sur la variation locale de $f(x)$. En effet, la variation de $f(a)$ est donnée par

$$\mathrm{d}f = f'(a)\mathrm{d}x, \tag{2.36}$$

ou encore

$$f(a + \mathrm{d}x) = f(a) + f'(a)\mathrm{d}x. \tag{2.37}$$

2.7.1.1 Les dérivées partielles

Pour une fonction à deux variable, $f(x, y)$, dont le domaine de définition est

$$f : \mathbb{R}^2 \to \mathbb{R}, \tag{2.38}$$

on définit la **dérivée partielle** de f par rapport à x ou à y

$$\frac{\partial f}{\partial x}(x, y) = \lim_{h\to 0} \frac{f(x + h, y) - f(x, y)}{h}, \tag{2.39}$$

$$\frac{\partial f}{\partial y}(x, y) = \lim_{h\to 0} \frac{f(x, y + h) - f(x, y)}{h}. \tag{2.40}$$

Comme on le voit ici, pour chaque dérivée partielle, on ne fait varier qu'une seule variable, les autres sont considérées comme des constantes.

Illustration 10 *(Dérivée partielle)*

Les dérivée partielles de la fonction

$$f(x, y) = x^2 \cdot y + 3, \tag{2.41}$$

sont données par

$$\frac{\partial f}{\partial x}(x, y) = 2xy, \tag{2.42}$$

$$\frac{\partial f}{\partial y}(x, y) = x^2. \tag{2.43}$$

Pour une fonction f dépendant d'un nombre n de variables, la notation est la suivante. Soit $f(\vec{x})$ avec $\vec{x} = \{x_i\}_{i=1}^n$, ou $\vec{x} \in \mathbb{R}^n$, on définit la dérivée par rapport à la i-ème composante de $\vec{x}$ comme

$$\frac{\partial f}{\partial x_i}(x_1, x_2, ..., x_i, ..., x_n) = \lim_{h\to 0} \frac{f(x_1, x_2, ..., x_i + h, ..., x_n) - f(x_1, x_2, ..., x_i, ..., x_n)}{h}. \tag{2.44}$$

Remarque 4

Pour une fonction à une seule variable, $f(x)$, on a que

$$f'(x) = \frac{\mathrm{d}f}{\mathrm{d}x}(x) = \frac{\partial f}{\partial x}(x). \tag{2.45}$$

De façon similaire à ce qui se passe pour les fonction à une seule variables, nous pouvons définir les dérivées secondes pour les façon à une seule variable. Pour une fonction à deux variables, on a en fait quatre dérivées secondes

$$\frac{\partial}{\partial x}\frac{\partial f}{\partial x}(x,y) = \frac{\partial^2 f}{\partial x^2}(x,y), \tag{2.46}$$

$$\frac{\partial}{\partial x}\frac{\partial f}{\partial y}(x,y) = \frac{\partial^2 f}{\partial x \partial y}(x,y), \tag{2.47}$$

$$\frac{\partial}{\partial y}\frac{\partial f}{\partial x}(x,y) = \frac{\partial^2 f}{\partial y \partial x}(x,y), \tag{2.48}$$

$$\frac{\partial}{\partial y}\frac{\partial f}{\partial y}(x,y) = \frac{\partial^2 f}{\partial y^2}(x,y). \tag{2.49}$$

Remarque 5

Si f est dérivable en x et y, on a que

$$\frac{\partial^2 f}{\partial x \partial y}(x,y) = \frac{\partial^2 f}{\partial y \partial x}(x,y). \tag{2.50}$$

Illustration 11 *(Dérivées partielles deuxièmes)*

Pour la fonction $f(x,y) = x^2 - y^2$, on a

$$\frac{\partial^2 f}{\partial x^2}(x,y) = \frac{\partial(2 \cdot x)}{\partial x}(x,y) = 2, \tag{2.51}$$

$$\frac{\partial^2 f}{\partial x \partial y}(x,y) = \frac{\partial(-2 \cdot y)}{\partial x}(x,y) = 0, \tag{2.52}$$

$$\frac{\partial^2 f}{\partial y \partial x}(x,y) = \frac{\partial(2 \cdot x)}{\partial y}(x,y) = 0, \tag{2.53}$$

$$\frac{\partial^2 f}{\partial y^2}(x,y) = \frac{\partial(-2 \cdot y)}{\partial y}(x,y) = -2. \tag{2.54}$$

On peut également généraliser pour des fonction à n variables où la deuxième dérivée partielle par rapport aux variables x_i, x_j s'écrit

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x,y). \tag{2.55}$$

2.7.1.2 Le gradient

Pour une fonction à deux variables, $f(x,y)$, on a vu qu'on peut calculer ses dérivées partielles par rapport à x et y

$$\frac{\partial f}{\partial x}, \quad \frac{\partial f}{\partial y}. \tag{2.56}$$

Une construction mathématique possible est d'écrire un vecteur avec ces deux quantités

$$\text{grad} f(x,y) = \vec{\nabla} f(x,y) = \left(\frac{\partial f}{\partial x}(x,y), \frac{\partial f}{\partial y}(x,y) \right)^{\mathrm{T}}. \tag{2.57}$$

Le symbole *nabla*, $\vec{\nabla}$, est une notation un peu étrange. Il représente un vecteur contenant toutes les dérivées partielles

$$\vec{\nabla} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right)^{\mathrm{T}}. \tag{2.58}$$

Cette notation est très utile pour se souvenir de ce qu'est un gradient, car on peut l'écrire un peu comme le "produit" entre l'opérateur $\vec{\nabla}$ et f

$$\vec{\nabla} f = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right)^{\mathrm{T}} f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)^{\mathrm{T}}. \tag{2.59}$$

On peut généraliser cette notation pour n variables à

$$\vec{\nabla} = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, ..., \frac{\partial}{\partial x_n} \right)^{\mathrm{T}}. \tag{2.60}$$

et

$$\vec{\nabla} f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_n} \right)^{\mathrm{T}}. \tag{2.61}$$

Illustration 12 *(Gradient d'une fonction à deux variables)*

Pour la fonction $f(x,y) = x^2 - y^2$, le gradient est donné par

$$\vec{\nabla} f = (2x, -2y)^{\mathrm{T}}. \tag{2.62}$$

Graphiquement, ceci est un *champs de vecteur* est peut se représenter comme

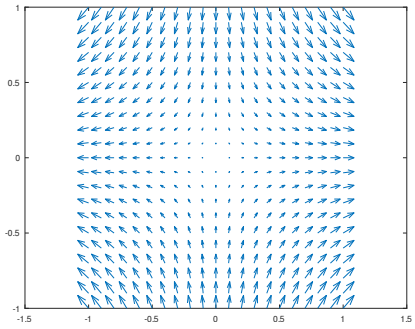

FIGURE 2.9 – Le champs de vecteur $\vec{\nabla} f(x,y) = (2x, -2y)^{\mathrm{T}}$.

Revenons à nos fonctions à deux variables. Le gradient d'une fonction a une très grande utilité pratique. En effet, il nous donne la variation de f dans chacun des direction de l'espace. On peut donc (un peu comme on avait fait pour les fonctions à une dimensions) se poser la question de la variation de f dans une direction particulière, $\vec{v}$. Comme nous connaissons le taux de variation de f dans chacune des directions, nous pouvons définir la **dérivée directionnelle** de f en un point (a, b), comme

$$(\vec{\nabla}_{\vec{v}} f)(a, b) = (\vec{\nabla} f)(a, b) \cdot \vec{v}, \tag{2.63}$$

où $\vec{v} = (v_1, v_2)^{\mathrm{T}}$. Cette grandeur représente la variation de $f(a, b)$ dans une direction particulière, $\vec{v}$. Comme pour les fonctions à une variable on peut écrire que

$$f(a + v_1, b + v_2) = f(a, b) + \vec{v} \cdot (\vec{\nabla} f(a, b)). \tag{2.64}$$

Cette dérivée directionnelle va nous permettre d'interpréter ce que représente le gradient d'une fonction.

En fait, le gradient a une interprétation très intéressante. Ce n'est rien d'autre que la direction de la pente la plus élevée sur chaque point de la fonction. C'est la direction, si vous faites de la randonnée en montagne, qui vous permettra de monter le long de la pente la plus raide en chaque point.

A l'inverse, imaginez que vous êtes un skieur et que votre montagne est décrite par la fonction $f(\vec{x})$. Le vecteur $-\vec{\nabla} f$ est la direction dans laquelle vous descendez si vous suivez tout droit la pente la plus raide.

Pour s'en convaincre essayons de prendre le problème à l'envers. On cherche la dérivée directionnelle $\vec{\nabla}_{\vec{v}} f$, telle que celle ci-soit maximale, pour tous les vecteur $\vec{v}$ de longueur 1. En d'autres termes

$$\max_{||\vec{v}||=1} \vec{v} \cdot \vec{\nabla} f. \tag{2.65}$$

Il faut à présent se rappeler que le produit scalaire de deux vecteurs peut s'écrire

$$\vec{a} \cdot \vec{b} = ||a|| \cdot ||b|| \cdot \cos\theta, \tag{2.66}$$

avec θ l'angle entre $\vec{a}$ et $\vec{b}$. De ceci, on déduit que la valeur maximale de $\vec{v} \cdot \vec{\nabla} f$ est atteinte quand $\vec{v}$ est aligné avec ∇f, ce qui ne se produit que quand $\vec{v}$ a la valeur

$$\vec{v}^* = \frac{\nabla f}{||\nabla f||}. \tag{2.67}$$

La variation maximale est donc atteinte quand on suit le vecteur pointé par ∇f. Par ailleurs, la dérivée directionnelle dans la direction de $\vec{v}^*$, on a

$$\vec{\nabla}_{\vec{v}^*} \cdot (\vec{\nabla} f) = \frac{\nabla f \cdot f}{||\vec{\nabla} f||} = ||\vec{\nabla} f||. \tag{2.68}$$

Le taux de variation maximal est donc la longueur du vecteur $\vec{\nabla} f$.

Remarque 6 *(Généralisation)*

Tout ce que nous venons d'écrire ici se généralise à un nombre arbitraire de dimensions.

2.7.1.3 Le lien avec les problème d'optimisation

Un cas qui nous intéresse particulièrement ici, est lorsque que le gradient d'une fonction est nul

$$\nabla f(x, y) = \vec{0}. \tag{2.69}$$

Cela veut dire que si nous trouvons un tel point (x, y) la variation de la fonction localement (sa "pente") sera nulle. Exactement comme pour le cas à une seule variable cela ne suffit pas pour déterminer si nous avons à faire à un minimum, un maximum, ou un point d'inflexion. Un exemple typique est la fonction

$$f(x, y) = x^2 - y^2. \tag{2.70}$$

Bien que $\nabla f(0, 0) = \vec{0}$, nous voyons sur la fig. 2.8 que bien que nous ayons un minimum dans la direction x, nous avons un maximum dans la direction y. On se retrouve dans un cas où nous avons un point-selle.

Pour pouvoir en dire plus il nous faut étudier les deuxièmes dérivées de $f(x, y)$ comme pour le cas unidimensionnel.

Prenons un exemple, où (voir fig. 2.10 pour voir à quoi elle ressemble)

$$f(x, y) = x^2 + 4y^3 - 12y - 2. \tag{2.71}$$

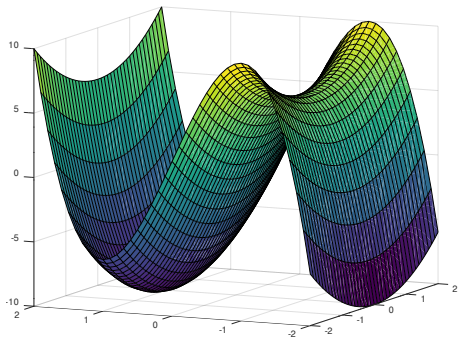

FIGURE 2.10 – La surface $f(x, y) = x^2 + 4y^3 - 12y - 2$.

Le gradient de $f(x, y)$ est donné par

$$\frac{\partial f}{\partial x} = 2x, \tag{2.72}$$

$$\frac{\partial f}{\partial y} = 12y^2 - 12. \tag{2.73}$$

Les coordonnées (x, y) où $\vec{\nabla} f = \vec{0}$ sont données par

$$2x = 0 \Leftrightarrow x = 0, \tag{2.74}$$

$$12y^2 - 12 = 0 \Leftrightarrow y_\pm = \pm 1. \tag{2.75}$$

On a donc deux points $(x, y_-) = (0, -1)$ et $(x, y_+) = 1$ qui satisfont $\vec{\nabla} f = 0$. Essayons de connaître la nature de ces points. Sont-il des maxima, minima, ou des point-selle?

Sur la fig. 2.10, on voit que le point $(0, -1)$ est un point selle, et le point $(0, 1)$ est un minimum. Nous allons à présent essayer de voir ce que cela veut dire mathématiquement sans avoir besoin de regarder le graphe de cette fonction. Inspirés par ce que nous savons des points critiques en une dimensions, nous allons étudier les deuxièmes dérivées

$$\frac{\partial^2 f}{\partial x^2} = 2, \tag{2.76}$$

$$\frac{\partial^2 f}{\partial x \partial y} = 0, \tag{2.77}$$

$$\frac{\partial^2 f}{\partial y^2} = 24y. \tag{2.78}$$

En substituant les valeur $(0, -1)$ et $(0, 1)$ dans les deuxièmes dérivées, on obtient

$$\frac{\partial^2 f}{\partial x^2}(0, 1) = \frac{\partial^2 f}{\partial x^2}(0, -1) = 2, \tag{2.79}$$

$$\frac{\partial^2 f}{\partial x \partial y}(0, 1) = \frac{\partial^2 f}{\partial x \partial y}(0, -1) = 0, \tag{2.80}$$

$$\frac{\partial^2 f}{\partial y^2}(0, 1) = 24, \quad \frac{\partial^2 f}{\partial y^2}(0, -1) = -24. \tag{2.81}$$

On voit ici, que pour les deux points $\frac{\partial^2 f}{\partial x^2} > 0$, on a donc que dans la direction x ces deux points sont des minimas. Mais cela ne suffit pas pour en faire des minimas locaux. Il faut également étudier ce qui se passe dans la direction y. Dans ce cas précis, on a qu'en $(0, 1)$ nous avons une valeur positive (c'est donc un minimum) et en $(0, -1)$ la valeur est négative (c'est donc un maximum).

Pour récapituler:

— En $(0, 1)$ c'est un minimum pour x et un minimum pour y. Et donc c'est un minimum local.
— En $(0, -1)$ c'est un minimum pour x et un maximum pour y. Et donc c'est un point-selle.

Globalement, pour avoir un min/max, il faut que les deuxièmes dérivées dans chacune des directions donnent la même interprétation pour pouvoir conclure à un minimum/maximum. Sinon c'est un point-selle.

2.7.2 La descente de gradient

Revenons à présent à l'optimisation d'une fonction de coût $f(\vec{x})$. Pour simplifier considérons la fonction

$$f(x, y) = x^2 + y^2. \tag{2.82}$$

Nous pouvons facilement nous convaincre que cette fonction possède un minimum en $(0,0)$ en la dessinant. On peut aussi aisément vérifier que $\nabla f(0,0) = \vec{0}$. En effet,

$$\nabla f(x,y) = (2x, 2y), \tag{2.83}$$

et donc

$$\nabla f(0,0) = (0,0). \tag{2.84}$$

Même si cela ne suffit pas à prouver mathématique que $\vec{0}$ est le minimum de cette fonction nous nous en satisferons.

Question 1

Avec ce qui précède, voyez-vous une façon de trouver le minimum de la fonction $f(x,y)$?

Une méthode pour trouver le minimum de $f(x,y)$ est la méthode de la *descente de gradient*. Cette méthode correspond intuitivement à la méthode que suivrait un skieur pour arriver le plus vite possible en bas d'une montagne. Pour ce faire, il suivrait toujours la pente la plus raide possible.

La méthode de la descente de gradient est une méthode itérative. Soient donnés un point de départ $\vec{x}_0$, et une fonction objectif $f(\vec{x})$, on va approximer le zéro itérativement avec une suite $\vec{x}_1$, $\vec{x}_2$, ... telle que

$$\vec{x}_1 = \vec{x}_0 - \lambda \cdot \nabla f(\vec{x}_0), \tag{2.85}$$

$$\vec{x}_2 = \vec{x}_1 - \lambda \cdot \nabla f(\vec{x}_1), \tag{2.86}$$

$$\cdots \vec{x}_{n+1} = \vec{x}_n - \lambda \cdot f(\vec{x}_n), \tag{2.87}$$

où $\lambda \in \mathbb{R}^+$ est un coefficient positif. On peut assez facilement se convaincre que si λ est suffisamment petit, alors $f(\vec{x}_{n+1}) \leq f(\vec{x}_n)$ (on ne fait que descendre la pente jusqu'à atteindre un minimum). Une illustration de ce processus peut se voir dans la fig. 2.11.

Illustration 13 *(quelques itérations)*

Prenons la fonction objectif $f(x,y)$ suivante

$$f(x,y) = x^2 + y^2, \tag{2.88}$$

et son gradient

$$\nabla f(x,y) = 2x + 2y. \tag{2.89}$$

Si on prend comme point de départ $\vec{x}_0 = (1, 0.5)$ et $\lambda = 0.25$, on a

$$\vec{x}_1 = \vec{x}_0 - \lambda \cdot \nabla f(\vec{x}_0) = (1, 0.5) - 0.25 \cdot (2 \cdot 1, 2 \cdot 0.5) = (0.5, 0.25), \tag{2.90}$$

$$\vec{x}_2 = \vec{x}_1 - \lambda \cdot \nabla f(\vec{x}_1) = (0.5, 0.25) - 0.25 \cdot (2 \cdot 0.5, 2 \cdot 0.25) = (0.25, 0.125), \tag{2.91}$$

$$\dots \tag{2.92}$$

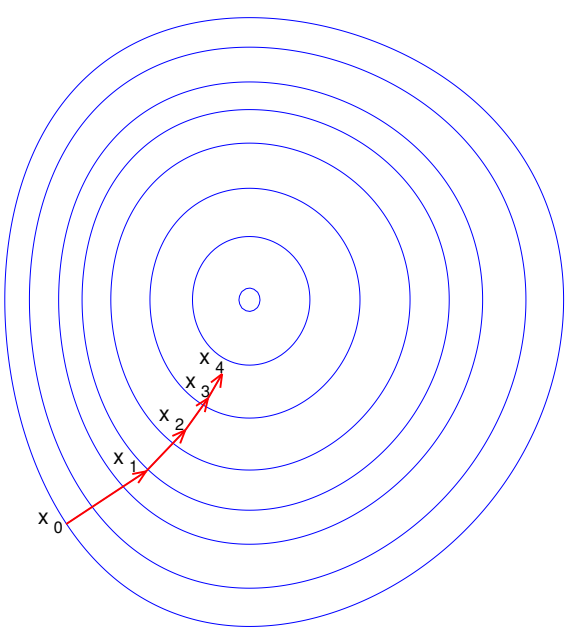

FIGURE 2.11 – Suite d'étapes pour la descente de gradient. En bleu on voit les courbes de niveaux (les courbes où $f(\vec{x})$ est constante). Source: Wikipedia https://bit.ly/2Fhvn7p

En changeant $\lambda = 0.5$, on voit qu'on arrive sur le zéro de la fonction en une itération

$$\vec{x}_1 = \vec{x}_0 - \lambda \cdot \nabla f(\vec{x}_0) = (1, 0.5) - 0.5 \cdot (2 \cdot 1, 2 \cdot 0.5) = (0, 0). \tag{2.93}$$

Comme pour les fonction à une seule variable, il est nécessaire de spécifier une condition d'arrêt pour la descente de gradient. En général, on choisit une tolérance, $\varepsilon > 0$, et la condition d'arrêt s'écrit

$$\text{Si } ||\vec{x}_{n+1} - \vec{x}_n|| < \varepsilon, \tag{2.94}$$

alors $\vec{x}_{n+1}$ est le zéro de $f(\vec{x})$.

Dépendant de la valeur de λ la *convergence* de la méthode peut varier grandement. Si λ est trop petit il faut une énorme quantité d'itérations pour atteindre le minimum. A l'inverse, en choisissant un λ trop grand, nous ne somme pas sûrs que nous convergerons un jour. En effet, on pourrait s'éloigner de plus en plus du minimum plutôt que de sen approcher. En général, on choisit $\lambda \in [0, 1)$ mais il n'y a pas de méthode générale pour en choisir une valeur "optimale". Cela signifie que pour une fonction quelconque, λ est choisi de façon empirique.

Chapitre 3

Intégrales

3.1 Interprétation géométrique

Dans ce chapitre nous nous intéressons au calcul d'aires sous une fonction f. La fonction f satisfait les hypothèses suivantes.

1. $f(x)$ est bornée dans l'intervalle $[a, b] \in \mathbb{R}$.
2. $f(x)$ est continue presque partout.

Nous définissions également l'infimum de f sur un intervalle $[x_0, x_1]$, noté

$$\inf_{[x_0,x_1]} f(x) \tag{3.1}$$

comme étant la plus grande valeur bornant par dessous toutes les valeurs prises par $f(x)$ dans l'intervalle $[x_0, x_1]$. Le suprémum sur un intervalle $[x_0, x_1]$, noté

$$\sup_{[x_0,x_1]} f(x) \tag{3.2}$$

comme étant la plus petite valeur bornant par dessus toutes les valeurs prises par $f(x)$ dans l'intervalle $[x_0, x_1]$.

Finalement nous définissons une subdivision

$$\Delta_n = \{a = x_0 < x_1 < ... < x_{n-1} < x_n = b\} \tag{3.3}$$

est une suite finie contenant $n + 1$ termes dans $[a, b]$.

On peut à présent approximer l'aire sous la fonction $f(x)$ dans l'intervalle $[a, b]$ de plusieurs façons:

1. $A^i(n) = \sum_{i=0}^{n-1} \inf_{[x_i,x_{i+1}]} f(x) \cdot (x_{i+1} - x_i)$ comme étant l'aire inférieure.
2. $A^s(n) = \sum_{i=0}^{n-1} \sup_{[x_i,x_{i+1}]} f(x) \cdot (x_{i+1} - x_i)$ comme étant l'aire supérieure.
3. $A^R(n) = \sum_{i=0}^{n-1} f(\xi_i) \cdot (x_{i+1} - x_i)$, $\xi_i \in [x_i, x_{i+1}]$

1 et 2 sont les sommes de Darboux, 3 est une somme de Riemann qui, dépendant des choix des ξ_i, peut être égale à 1 ou à 2.

L'aire de sous la fonction $f(x)$ est donnée par la limite pour $n \to \infty$ de A^i ou A^s (si elle existe). Dans ce cas $n \to \infty$ A^R (pris en sandwich entre A^i et A^n) nous donne aussi l'aire sous la fonction.

Remarque 7

1. Ces sommes peuvent être positives ou négatives en fonction du signe de f.
2. Une implémentation informatique est immédiate, en particulier pour la somme de Riemann.

Définition 10 *(Intégrabilité au sens de Riemann)*

Une fonction est dite intégrable au sens de Riemann si

$$\lim_{n\to\infty} A^i(n) = \lim_{n\to\infty} A^s(n) = \int_a^b f(x)\mathrm{d}x. \tag{3.4}$$

Dans la formule

$$\int_a^b f(x)\mathrm{d}x, \tag{3.5}$$

x est appelée variable d'intégration, a et b sont les bornes d'intégration. Pour des raisons de consistance dans les notations la variable d'intégration ne peut être désignée avec le même symbole qu'une des bornes d'intégration.

Exemple 1 *(Intégration de Riemann)*

Intégrer de $f(x) = x$ dans intervalle $[0, 1]$.

Solution 1 *(Intégration de Riemann)*

Il est élémentaire de calculer que cette aire vaut $1/2$ (c'est l'aire d'un triangle rectangle de côté 1). Néanmoins, évaluons également cette aire à l'aide de A^i et A^s. Commençons par subdiviser $[0, 1]$ en n intervalles égaux de longueur $\delta = 1/n$. Comme $f(x)$ est strictement croissante, on a que $\inf_{[x_i, x_{i+1}]} f(x) = f(x_i)$ et que $\sup_{[x_i, x_{i+1}]} f(x) = f(x_{i+1})$. On a donc que

1. $A^i(n) = \delta \sum_{i=0}^{n-1} x_i = \delta \sum_{i=0}^{n-1} \frac{i}{n} = \frac{n(n-1)}{2n^2} = \frac{n-1}{2n}$ [1]. Et donc en prenant la limite pour $n \to \infty$ il vient

$$A^i = \lim_{n\to\infty} \frac{n-1}{2n} = \frac{1}{2}. \tag{3.6}$$

1. La somme $\sum_{i=0}^{n} i = n(n+1)/2$

2. $A^s(n) = \delta \sum_{i=0}^{n-1} x_{i+1} = \delta \sum_{i=0}^{n-1} \frac{i+1}{n} = \delta \sum_{i=0}^{n} \frac{i}{n} = \frac{n(n+1)}{2n^2} = \frac{n+1}{2n}$. Et donc en prenant la limite pour $n \to \infty$ il vient

$$A^s = \lim_{n\to\infty} \frac{n+1}{2n} = \frac{1}{2}. \tag{3.7}$$

Exercice 6 *(Intégration de Riemann de x^2)*

Calculer l'aire sous la courbe de $f(x) = x^2$ dans intervalle $[0, 1]$.

Indication: $\sum_{i=0}^{n} i^2 = \frac{1}{6}n(n+1)(2n+1)$.

3.2 Interprétation physique

Supposons que nous ayons une fonction, $x(t)$, qui donne la position d'un objet pour un intervalle de temps $t \in [a, b]$. Nous pouvons aisément en déduire la vitesse $v(t)$ de l'objet, comme étant la variation de $x(t)$ quand t varie. Autrement dit $v(t) = x'(t)$.

Supposons à présent que nous ne connaissions que la vitesse $v(t)$ de notre objet. Afin de déduire sa position nous prendrions un certain nombre d'intervalles de temps $\delta t_i = t_{i+1} - t_i$ que nous multiplierions par $v(t_i)$ afin de retrouver la distance parcourue pendant l'intervalle δt_i et ainsi de suite. Afin d'améliorer l'approximation de la distance parcourue nous diminuerions la valeur de δt_i jusqu'à ce que $\delta t_i \to 0$.

Nous voyons ainsi que cette méthode, n'est autre qu'une façon "intuitive" d'intégrer la vitesse afin de trouver la position. Et que l'intégrale et la dérivée sont étroitement liées: la vitesse étant la dérivée de la position et la position étant l'intégrale de la vitesse.

3.3 Primitive

Si maintenant nous essayons de généraliser le calcul de l'intégrale d'une fonction, il s'avère que le calcul d'une intégrale est l'inverse du calcul d'une dérivée.

Définition 11 *(Primitive)*

Soit f une fonction. On dit que F est une primitive de f sur l'intervalle $D \subseteq \mathbb{R}$ si $F'(x) = f(x)\ \forall x \in D$.

Si F est une primitive de f, alors on peut définir la fonction G telle que $G(x) = F(x) + C$, $C \in \mathbb{R}$ qui est aussi une primitive de f. On voit que la primitive de f est définie à une constante additive près. En effet, si $F' = f$ on a

$$G' = F' + \underbrace{C'}_{=0} = F' = f. \tag{3.8}$$

Théorème 2 *(Unicité)*

Pour $a \in D$ et $b \in \mathbb{R}$ il existe une unique primitive F telle que $F(a) = b$.

Illustration 14 *(Unicité)*

Soit $f(x) = x$, alors l'ensemble de primitives correspondantes est $G = x^2/2 + C$. Si nous cherchons la primitive telle que $G(0) = 0$, il vient que $C = 0$ et donc la primitive est unique et vaut $F(x) = x^2/2$.

Exercices 1 *(Primitives)*

Calculez les primitives suivantes (*indication: il s'agit de trouver les fonctions* $F(x)$ *telles que* $F'(x) = f(x)$):

1. $F(x) = \int x^2 \mathrm{d}x$.
2. $F(x) = \int x^n \mathrm{d}x$, $n \in \mathbb{R}\backslash\{-1\}$.
3. $F(x) = \int \sqrt{x} \mathrm{d}x$.
4. $F(x) = \int \frac{1}{x} \mathrm{d}x$.
5. $F(x) = \int \exp(x) \mathrm{d}x$.
6. $F(x) = \int \sin(x) \mathrm{d}x$.

Maintenant que vous avez calculé toutes ces primitives de base, nous pouvons récapituler des formules qui seront importantes pour la suite:

1. $\int x^n \mathrm{d}x = \frac{1}{n+1} x^{n+1} + C$, $n \in \mathbb{R}\backslash\{-1\}$.
2. $\int \frac{1}{x} \mathrm{d}x = \ln(x) + C$.
3. $\int \exp(x) \mathrm{d}x = \exp(x) + C$.
4. $\int \sin(x) \mathrm{d}x = -\cos(x) + C$.
5. $\int \cos(x) \mathrm{d}x = \sin(x) + C$.

Théorème 3 *(Théorème fondamental du calcul intégral)*

En définissant à présent l'intégrale à l'aide de la notion de primitive, nous avons que pour $a, b \in \mathbb{R}$ et $a < b$

$$\int_a^b f(x) \mathrm{d}x = F|_a^b = F(b) - F(a). \tag{3.9}$$

On dit que x est la variable d'intégration. Elle est dite "muette" car elle disparaît après que l'intégrale ait été effectuée. On peut donc écrire l'équation ci-dessus de façon équivalente en remplaçant le symbole x par n'importe quelle autre lettre (sauf a, b, f, F).

Remarque 8

On notera que la constante additive C a disparu de cette formule. En effet, remplaçons F par $G = F + C$, il vient

$$\int_a^b f(x)\mathrm{d}x = G(b) - G(a) = F(b) + C - F(a) - C = F(b) - F(a). \tag{3.10}$$

Il suit de l'éq. 3.9 que

$$\int_a^a f(x)\mathrm{d}x = F(a) - F(a) = 0 \tag{3.11}$$

et que

$$\int_a^b f(x)\mathrm{d}x = -\int_b^a f(x)\mathrm{d}x \tag{3.12}$$

Nous pouvons à présent définir la fonction $G(x)$ telle que

$$G(x) = \int_a^x f(y)\mathrm{d}y = F(x) - F(a). \tag{3.13}$$

Il suit que $G(x)$ est la primitive de f telle que $G(a) = 0$.

Propriétés 4

Soient f et g deux fonctions intégrables sur un intervalle $D = [a, b] \subseteq \mathbb{R}$, $c \in [a, b]$, et $\alpha \in \mathbb{R}$. On a

1. La dérivée et l'intégrale "s'annulent"

$$\left(\int_a^x f(x)\mathrm{d}x\right)' = (F(x) - F(a))' = F'(x) - (F(a))' = F'(x) = f(x). \tag{3.14}$$

2. La fonction $h = f + g$ admet aussi une primitive sur D, et on a

$$\int_a^b (f(x) + g(x))\mathrm{d}x = \int_a^b f(x)\mathrm{d}x + \int_a^b g(x)\mathrm{d}x. \tag{3.15}$$

3. La fonction $h = \alpha f$ admet aussi une primitive sur D, et on a

$$\int_a^b \alpha f(x)\mathrm{d}x = \alpha \int_a^b f(x)\mathrm{d}x. \tag{3.16}$$

4. Relation de Chasles (faire la démonstration en exercice)

$$\int_a^c f(x)\mathrm{d}x = \int_a^b f(x)\mathrm{d}x + \int_b^c f(x)\mathrm{d}x. \tag{3.17}$$

   De cette relation on déduit qu'on peut calculer l'intégrale d'une fonction continue par morceaux sur $[a, b]$.

5. Si f est paire alors

$$\int_{-a}^a f(x)\mathrm{d}x = 2\int_0^a f(x)\mathrm{d}x. \tag{3.18}$$

6. Si f est impaire alors

$$\int_{-a}^a f(x)\mathrm{d}x = 0. \tag{3.19}$$

3.3.1 Intégrales impropres

Si une des bornes d'intégration ou si la fonction à intégrer admet une discontinuité à des points bien définis, nous parlons intégrales impropres.

Lorsqu'une borne d'intégration est infinie, alors nous pouvons avoir les cas de figures suivants

$$\begin{aligned}
\int_a^\infty f(x)\mathrm{d}x &= \lim_{b\to\infty}\int_a^b f(x)\mathrm{d}x,\\
\int_{-\infty}^b f(x)\mathrm{d}x &= \lim_{a\to\infty}\int_{-a}^b f(x)\mathrm{d}x,\\
\int_{-\infty}^\infty f(x)\mathrm{d}x &= \lim_{a\to\infty}\int_{-a}^a f(x)\mathrm{d}x.
\end{aligned} \tag{3.20}$$

Exemple 2 *(Intégrale impropre)*

Calculer l'intégrale suivante

$$\int_0^\infty e^{-ax}\mathrm{d}x, \quad a > 0. \tag{3.21}$$

Solution 2 *(Intégrale impropre)*

Nous pouvons réécrire l'intégrale ci-dessus comme

$$\int_0^\infty e^{-ax}\mathrm{d}x = \lim_{b\to\infty}\int_0^b e^{-ax}\mathrm{d}x = -\frac{1}{a}\lim_{b\to\infty}\left[e^{-ax}\right]_0^b = -\frac{1}{a}\left[\lim_{b\to\infty} e^{-ab} - 1\right] = \frac{1}{a}. \tag{3.22}$$

Exercice 7

Calculer l'intégrale suivante

$$\int_1^\infty \frac{1}{x^2}\mathrm{d}x. \tag{3.23}$$

Lorsque nous avons une discontinuité dans la fonction f au point $c \in [a, b]$ nous avons

$$\int_a^b f(x)\mathrm{d}x = \lim_{\varepsilon \to 0} \int_a^{c-\varepsilon} f(x)\mathrm{d}x + \int_{c+\varepsilon}^b f(x)\mathrm{d}x. \tag{3.24}$$

Exercice 8

Montrer que

$$\int_{-1}^2 \frac{1}{x}\mathrm{d}x = \ln 2. \tag{3.25}$$

Définition 12 *(Valeur moyenne)*

Soit une fonction f admettant une primitive sur $[a, b]$ avec $a < b$, alors la valeur moyenne $\bar{f}$ de cette fonction sur $[a, b]$, est définie par

$$\bar{f} = \frac{1}{b-a} \int_a^b f(x)\mathrm{d}x. \tag{3.26}$$

3.4 Méthodes d'intégration

Dans cette section, nous allons étudier différentes méthodes pour intégrer des fonctions.

3.4.1 Intégration de fonctions usuelles et cas particuliers

Le calcul d'une primitive ou d'une intégrale n'est en général pas une chose aisée. Nous connaissons les formules d'intégration pour certaines fonctions particulières.

Polynômes

Les polynômes s'intègrent terme à terme. Pour $(\{a_i\}_{i=0}^n \in \mathbb{R}$

$$\begin{aligned} &\int a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1}x^{n-1} + a_n x^n \mathrm{d}x \\ =&\int a_0 \mathrm{d}x + \int a_1 x \mathrm{d}x + \int a_2 x^2 \mathrm{d}x + \cdots + \int a_{n-1}x^{n-1}\mathrm{d}x + \int a_n x^n)\mathrm{d}x \\ =&a_0 x + \frac{a_1}{2}x^2 + \frac{a_2}{3}x^3 + \cdots + \frac{a_n}{n+1}x^{n+1} + c. \end{aligned} \tag{3.27}$$

Exercice 9

Intégrer la fonction suivante

$$\int (x+2)(x^3 + 3x^2 + 4x - 3)\mathrm{d}x. \tag{3.28}$$

Application de la règle de chaîne pour l'intégration

Une primitive d'une fonction de la forme $f(x)f'(x)$ se calcule aisément

$$\int f(x)f'(x)\mathrm{d}x = \frac{1}{2}f(x)^2 + c. \tag{3.29}$$

Nous calculons par exemple

$$\int \sin(x)\cos(x)\mathrm{d}x = \frac{1}{2}\sin^2(x) + c = -\frac{1}{2}\cos^2(x) + c'. \tag{3.30}$$

Inverse de la dérivation logarithmique

Une primitive de la forme

$$\int \frac{f'(x)}{f(x)}\mathrm{d}x = \ln(f(x)) + c. \tag{3.31}$$

Exemple 3

Calculer la primitive suivante

$$\int \frac{1}{x}\mathrm{d}x. \tag{3.32}$$

Solution 3

Le calcul de la primitive de suivante

$$\int \frac{1}{x}\mathrm{d}x = \int \frac{(x)'}{x}\mathrm{d}x = \ln(x) + c. \tag{3.33}$$

3.4.1.1 Règle de chaîne

Une des façons les plus simples de calculer une primitive est de reconnaître la règle de chaîne dans le terme à intégrer

$$\int g'(f(x))f'(x)\mathrm{d}x = \int [g(f(x))]'\mathrm{d}x = g(f(x)) + c. \tag{3.34}$$

Illustration 15

Si g est définie comme $g(x) = x^{-1}$ et $f(x) = 3x^2 + 2$, alors la primitive

$$\int \frac{f'(x)}{g'(f(x))}\mathrm{d}x = \int -\frac{6x}{(3x^2+2)^2}\mathrm{d}x = \frac{1}{3x^2+2} + c. \tag{3.35}$$

3.4.2 Intégration par parties

La dérivation d'un produit de fonctions $f \cdot g$ s'écrit

$$(f(x)g(x))' = f'(x)g(x) + f(x)g'(x). \tag{3.36}$$

En intégrant cette équation on obtient

$$f(x)g(x) = \int f'(x)g(x)\mathrm{d}x + \int f(x)g'(x)\mathrm{d}x. \tag{3.37}$$

Une primitive de la forme $\int f'(x)g(x)\mathrm{d}x$ peut ainsi se calculer de la façon suivante

$$\int f'(x)g(x)\mathrm{d}x = f(x)g(x) - \int f(x)g'(x)\mathrm{d}x. \tag{3.38}$$

De façon similaire si nous nous intéressons à une intégrale définie

$$\int_a^b f'(x)g(x)\mathrm{d}x = (f(x)g(x))|_a^b - \int_a^b f(x)g'(x)\mathrm{d}x. \tag{3.39}$$

Le choix des fonctions est complètement arbitraire. Néanmoins, le but de cette transformation est de remplacer une intégrale par une autre dont on connaîtrait la solution.

Des "règles" pour utiliser cette technique seraient que

1. g' soit facile à calculer et aurait une forme plus simple que g.
2. $\int f'\mathrm{d}x$ soit facile à calculer et aurait une forme plus simple que f'.

Exemple 4

Calculer les primitives suivantes

1. $\int xe^x\mathrm{d}x$.
2. $\int \cos(x)\sin(x)\mathrm{d}x$.

Solution 4

1. $\int xe^x\mathrm{d}x$. $g(x) = x$, $f'(x) = e^x$ et donc $g'(x) = 1$, $f(x) = e^x$. Il vient

$$\int xe^x\mathrm{d}x = xe^x - \int e^x\mathrm{d}x = xe^x - e^x + c. \tag{3.40}$$

2. $\int \cos(x)\sin(x)\mathrm{d}x$. $g = \cos(x)$, $f'(x) = \sin(x)$ et donc $g'(x) = -\sin(x)$, $f(x) = -\cos(x)$. Il vient

$$\int \cos(x)\sin(x)\mathrm{d}x = \sin^2(x) - \int \cos(x)\sin(x)\mathrm{d}x$$
$$\Rightarrow \int \cos(x)\sin(x)\mathrm{d}x = \frac{1}{2}\sin^2(x) + c.$$

On voit que le résultat de l'intégration par partie nous redonne l'intégrale de départ. Ceci nous permet d'évaluer directement la dite intégrale pour retrouver le résultat de l'éq. 3.30

Il est également possible d'enchaîner plusieurs intégrations par parties.

Exemple 5

Calculer l'intégrale de $\int x^2 e^x \mathrm{d}x$.

Solution 5

En posant $g(x) = x^2$, $f'(x) = e^x$ et donc $g'(x) = 2x$, $f(x) = e^x$. Il vient

$$\int x^2 e^x \mathrm{d}x = x^2 e^x - 2 \int x e^x \mathrm{d}x. \tag{3.41}$$

On pose de façon similaire $g(x) = x$, $f'(x) = e^x$ et donc $g'(x) = 1$, $f(x) = e^x$ et il vient

$$\int x^2 e^x \mathrm{d}x = x^2 e^x - 2 \left(x e^x - \int e^x \mathrm{d}x \right) = x^2 e^x - 2x e^x + 2e^x + c. \tag{3.42}$$

Exercice 10

Calculer les primitives suivantes

1. $\int \ln(x) \mathrm{d}x$
2. $\int x^2 \sin(x) \mathrm{d}x$
3. $\int e^x \sin(x) \mathrm{d}x$

3.4.3 Intégration par changement de variables

On observe que la dérivation de la composition de deux fonctions F et g est donnée par

$$(F \circ g)' = (f \circ g) \cdot g', \text{ ou } [F(g(y))]' = f(g(y)) \cdot g'(y), \tag{3.43}$$

où $f = F'$. Si nous intégrons cette relation on obtient

$$\int_a^b f(g(y)) g'(y) \mathrm{d}y = \int_a^b [F(g(y))]' \mathrm{d}y = F(g(y))|_a^b = F(g(b)) - F(g(a)) = \int_{g(a)}^{g(b)} f(x) \mathrm{d}x. \tag{3.44}$$

Cette relation nous mène au théorème suivant.

Théorème 4 *(Intégration par changement de variables)*

Soit f une fonction continue presque partout, et g une fonction dont la dérivée est continue presque partout sur un intervalle $[a, b]$. Soit également l'image de g contenue dans le domaine de définition de f. Alors

$$\int_a^b f(g(x))g'(x)\mathrm{d}x = \int_{g(a)}^{g(b)} f(z)\mathrm{d}z. \tag{3.45}$$

Nous utilisons ce théorème de la façon suivante. L'idée est de remplacer la fonction $g(x)$ par z. Puis il faut également remplacer $\mathrm{d}x$ par $\mathrm{d}z$ où nous avons que $\mathrm{d}x = \mathrm{d}z/g'(x)$. Finalement, il faut changer les bornes d'intégration par $a \to g(a)$ et $b \to g(b)$. Si on ne calcule pas l'intégrale mais la primitive, on ne modifie (évidemment) pas les bornes d'intégration, mais en revanche pour trouver la primitive il faut également appliquer la transformation $x = g^{-1}(z)$ sur la solution.

Exemple 6 *(Changement de variable)*

Intégrer par changement de variables $\int_1^3 6x\ln(x^2)\mathrm{d}x$.

Solution 6 *(Changement de variable)*

En définissant $z = x^2$, nous avons $\mathrm{d}x = \mathrm{d}z/(2x)$. Les bornes d'intégration deviennent $z(1) = 1^2 = 1$ et $z(3) = 3^2 = 9$. On obtient donc

$$\begin{aligned}\int_1^3 6x\ln(x^2)\mathrm{d}x &= \int_1^9 6x\ln(z)\frac{1}{2x}\mathrm{d}z = \int_1^9 \ln(z)\mathrm{d}z \\ &= 3\left[z\ln(z) - z\right]_1^9 = 3(9\ln(9) - 9 - \ln(1) + 1) = 27\ln(9) - 24.\end{aligned}$$

Exercice 11

Calculer les primitives suivantes par changement de variable

1. $\int \frac{1}{5x-7}\mathrm{d}x$
2. $\int \sin(3 - 7x)\mathrm{d}x$
3. $\int xe^{x^2}\mathrm{d}x$

3.5 Le produit de convolution

Les convolutions sont très utilisées pour le traitement du signal, le traitement d'images et les réseaux de neurones convolutifs entre autres.

3.5.1 La convolution continue

La convolution de deux fonctions intégrables, $f(t)$, et $g(t)$, notée $f * g$ se définit comme

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-t)g(t)\mathrm{d}t. \tag{3.46}$$

On constate que le membre de gauche de l'équation ci-dessus n'est rien d'autre qu'une fonction de x. Pour chaque valeur de $x = x_0$, on calcule l'intégrale,

$$\int_{-\infty}^{\infty} f(x_0-t)g(t)\mathrm{d}t. \tag{3.47}$$

Exercice 12 *(Commutativité)*

Démontrer que le produit de convolution est commutatif, soit

$$(f * g)(x) = (g * f)(x). \tag{3.48}$$

Indication: utiliser la substitution $\tau = x - t$.

Afin de pouvoir interpréter un peu ce que cela veut dire, il est intéressant de faire un calcul "simple" pour se faire une idée.

Exercice 13

Calculer la convolution du signal $f(t)$

$$f(t) = \begin{cases} 1, & \text{si } t \in [0,1] \\ 0, & \text{sinon.} \end{cases} \tag{3.49}$$

Indication: faites un dessin de ce que représente la convolution de ce f avec lui-même.

Interprétation avec les mains

Afin d'interpréter ce que représente le produit de convolution, introduisons la fonction delta de Dirac, $\delta_a(x)$. Cette fonction est un peu particulière, elle vaut zéro partout sauf en 0 (où elle est "infinie"), et son intégrale vaut 1

$$\int_{-\infty}^{\infty} \delta(x)\mathrm{d}x = 1. \tag{3.50}$$

Même si cela peut sembler étrange, on peut tenter de construire une telle fonction en prenant une suite de rectangles, centrés en 0, dont la surface vaut 1. Puis on rend ces rectangles de plus en plus fins, en imposant que la surface vaut toujours 1 et le tour est joué.

Cette fonction est intéressante, car elle a la propriété suivante lorsqu'on l'utilise pour effectuer des convolutions.

$$\int_{-\infty}^{\infty} f(y)\delta(y-x)\mathrm{d}y = f(x). \tag{3.51}$$

En d'autre termes cette intégrale est égale à la valeur de f au point où l'argument du δ est nul.

A présent, si nous considérons la convolution de $f(t)$ avec la fonction $\delta(t-a) = \delta_a$, on obtient

$$(f * \delta_a)(x) = \int_{-\infty}^{\infty} f(x-t)\delta(t-a)\mathrm{d}t = f(x-a). \tag{3.52}$$

En fait la convolution d'une fonction f avec le delta de Dirac centré en a ne fait que translater la fonction f d'une distance a.

En effectuant à présent la convolution avec une combinaison linéaire de δ de Dirac

$$(f*(\alpha\cdot\delta_a+\beta\cdot\delta_b))(x) = \int_{-\infty}^{\infty} f(x-y)(\alpha\cdot\delta(y-a)+\beta\cdot\delta(y-b))\mathrm{d}y = \alpha\cdot f(x-a)+\beta\cdot f(x-b). \tag{3.53}$$

La convolution est donc la moyenne pondérée de f translatée en a et en b par α et β respectivement.

On voit que de façon générale, qu'on peut interpréter la convolution de deux fonctions $f(t)$ et $g(t)$ comme la moyenne de $f(t)$ pondérée par la fonction $g(t)$.

Exercice 14 *(Convolution)*

Calculer la convolution de $f(x)$ avec $g(x)$, où $f(x)$ et $g(x)$ sont les fonctions

$$f(x) = \begin{cases} -1, & \text{si } -\pi \leq x \leq \pi \\ 0, & \text{sinon.} \end{cases}, \tag{3.54}$$

$$g(x) = \sin(x). \tag{3.55}$$

Le lien avec les filtres

Il se trouve que dans le cas où le filtre est linéaire (filtrer la combinaison de deux signaux est la même chose que de faire la combinaison linéaires des signaux filtrés) et indépendant du temps (les translations temporelles n'ont aucun effet sur lui) alors on peut lier la convolution et le filtrage.

Si on définit la réponse impulsionnelle d'un filtre, $h(t)$, le filtrage d'un signal $s(t)$, noté $f(s)$, n'est autre que la convolution de $h(t)$ avec $s(t)$

$$f(s) = (s * h)(x) = \int_{-\infty}^{\infty} f(x-t)g(t)\mathrm{d}t. \tag{3.56}$$

3.6 Intégration numérique

Dans certains cas, il est impossible d'évaluer analytiquement une intégrale ou alors elle est très compliquée à calculer. Dans ce cas, on va approximer l'intégrale et donc commettre une erreur.

Pour ce faire on subdivise l'espace d'intégration $[a, b]$ en N pas équidistants (pour simplifier) $\delta x = (b - a)/N$, et on approxime l'intégrale par une somme finie

$$\int_a^b f(x)\mathrm{d}x = \sum_{i=0}^{N-1} \delta x f(a + i\delta x)g_i + E(a, b, \delta x) \cong \sum_{i=0}^{N-1} \delta x f(a + i\delta x)g_i, \quad (3.57)$$

où g_i est un coefficient qui va dépendre de la méthode d'intégration que nous allons utiliser, E est l'erreur commise par l'intégration numérique et va dépendre des bornes d'intégration, de δx (du nombre de pas d'intégration), de la forme de $f(x)$ (combien est "gentille") et finalement de la méthode d'intégration.

3.6.1 Erreur d'une méthode d'intégration

D'une façon générale plus δx est petit (N est grand) plus l'erreur sera petite et donc l'intégration sera précise (et plus le calcul sera long). Néanmoins, comme la précision des machines sur lesquelles nous évaluons les intégrales est finie, si δx devient proche de la précision de la machine des erreurs d'arrondi vont dégrader dramatiquement la précision de l'intégration.

Remarque 9

De façon générale il est difficile de connaître à l'avance la valeur exacte de E. En revanche on est capable de déterminer **l'ordre** de l'erreur.

Définition 13 *(Ordre d'une méthode)*

On dit qu'une méthode d'intégration est d'ordre k, si l'erreur commise par la méthode varie proportionnellement à δx^k. On note qu'une erreur est d'ordre k par le symbole $\mathcal{O}(\delta x^k)$. Exemple: si une méthode est d'ordre deux, alors en diminuant δx d'un facteur 2, l'erreur sera elle divisée par $2^2 = 4$. Si une méthode est d'ordre 3, alors en diminuant δx d'un facteur 2, nous aurons que l'erreur est divisée par un facteur $2^3 = 8$. Etc.

Comme le calcul d'une intégrale de façon numérique ne donne en général pas un résultat exact, mais un résultat qui va dépendre d'un certain nombre de paramètres utilisés pour l'intégration, il faut définir un critère qui va nous dire si notre intégrale est calculée avec une précision suffisante.

Notons $I(N, a, b, f, g)$ l'approximation du calcul de l'intégrale entre a et b de la fonction f avec une résolution N pour la méthode d'intégration g

$$I(N, a, b, f, g) = \sum_{i=0}^{N-1} \delta x f(a + i\delta x) g_i, \tag{3.58}$$

où g_i est encore à préciser. Afin de déterminer si le nombre de points que nous avons choisi est suffisant, après avoir évalué $I(N, a, b, f, g)$, nous évaluons $I(2 \cdot N, a, b, f, g)$. En d'autres termes nous évaluons l'intégrales de la même fonction avec la même méthode mais avec un nombre de points deux fois plus élevé. Puis, nous pouvons définir $\varepsilon(N)$ comme étant l'erreur relative de notre intégration avec une résolution N et $2 \cdot N$

$$\varepsilon(N) \equiv \left| \frac{I(2N) - I(N)}{I(2N)} \right|. \tag{3.59}$$

Si à présent nous choisissons un $\varepsilon_0 > 0$ (mais plus grand que la précision machine), nous pouvons dire que le calcul numérique de notre intégrale a **convergé** (on parle de **convergence** du calcul également) pour une résolution N quand $\varepsilon(N) < \varepsilon_0$.

3.6.2 Méthode des rectangles

Pour la méthode des rectangles, nous allons calculer l'intégrale en approximant l'aire sous la fonction par une somme de rectangles, comme nous l'avons fait pour la définition de l'intégration au sens de Riemann. La différence principale est que nous ne regarderons pas les valeurs minimales ou maximales de f sur les subdivisions de l'espace, mais uniquement les valeurs sur les bornes. Cette approximation donne donc la formule suivante

$$\begin{aligned} \int_a^b f(x)\mathrm{d}x &\cong \sum_{i=0}^{N-1} \delta x f(a + i \cdot \delta x) + \mathcal{O}(\delta x), \\ &\cong \sum_{i=1}^{N} \delta x f(a + i \cdot \delta x) + \mathcal{O}(\delta x) \end{aligned} \tag{3.60}$$

Cette méthode est d'ordre 1. Une exception s'applique cependant concernant l'ordre de l'intégration. Si la fonction à intégrer est une constante $f(x) = c$, alors l'intégration est exacte.

Dans les deux cas ci-dessus on a évalué la fonction sur une des bornes. On peut améliorer la précision en utilisant le "point du milieu" pour évaluer l'aire du rectangle. L'approximation devient alors

$$\int_a^b f(x)\mathrm{d}x \cong \sum_{i=0}^{N-1} \delta x f(a + (i + 1/2) \cdot \delta x) + \mathcal{O}(\delta x^2). \tag{3.61}$$

Cette astuce permet d'améliorer la précision de la méthode à très faible coût. En effet, la précision de la méthode des rectangles est améliorée et devient d'ordre 2. Elle est exacte pour les fonctions linéaires $f(x) = c \cdot x + d$.

3.6.3 Méthode des trapèzes

Pour la méthode des trapèzes, nous allons calculer l'intégrale en approximant l'aire sous la fonction par une somme de trapèzes. Pour rappel l'aire d'un trapèze, dont les côtés parallèles sont de longueurs c et d et la hauteur h, est donnée pas

$$A = (c + d)h/2. \tag{3.62}$$

Cette approximation donne donc la formule suivante

$$\int_a^b f(x)\mathrm{d}x \cong \sum_{i=0}^{N-1} \delta x \frac{f(a + i \cdot \delta x) + f(a + (i + 1) \cdot \delta x)}{2} + \mathcal{O}(\delta x^2). \tag{3.63}$$

Cette méthode est d'ordre 2. Cette méthode d'intégration est aussi exacte pour les fonctions linéaires $f(x) = c \cdot x + d$.

3.6.4 Méthode de Simpson

Pour cette méthode, on approxime la fonction à intégrer dans un intervalle par une parabole.

Commençons par évaluer l'intégrale à l'aide d'une subdivision dans l'ensemble $[a, b]$.

L'idée est la suivante. On pose $f(x) = c \cdot x^2 + d \cdot x + e$ et il faut déterminer c, d, et e. Il faut donc choisir 3 points dans l'intervalle $[a, b]$ pour déterminer ces constantes. On choisit comme précédemment $f(a)$, $f(b)$, et le troisième point est pris comme étant le point du milieu $(f(a + b)/2)$. On se retrouve ainsi avec trois équations à trois inconnues

$$\begin{aligned} f(a) &= c \cdot a^2 + d \cdot a + e, \\ f(b) &= c \cdot b^2 + d \cdot b + e, \\ f((a + b)/2) &= \frac{c}{4} \cdot (a + b)^2 + \frac{d}{2} \cdot (a + b) + e. \end{aligned} \tag{3.64}$$

En résolvant ce système (nous n'écrivons pas la solution ici) nous pouvons à présent évaluer l'intégrale

$$\begin{aligned} I &= \int_a^b f(x)\mathrm{d}x \cong \int_a^b (cx^2 + dx + e)\mathrm{d}x, \\ &= \frac{b - a}{6}(f(a) + f(b) + 4f((a + b)/2)) + \mathcal{O}(\delta x^4). \end{aligned}$$

On peut généraliser et affiner cette formule en rajoutant des intervalles comme précédemment et en répétant cette opération pour chaque intervalle.

Il vient donc que

$$\begin{aligned} I = \frac{\delta x}{6} \sum_{i=0}^{N-1} &[f(a + i \cdot \delta x) + f(a + (i + 1) \cdot \delta x) \\ &+4f(a + (i + 1/2) \cdot \delta x)] + \mathcal{O}(\delta x^4). \end{aligned}$$

Cette méthode permet d'évaluer exactement les intégrales des polynômes d'ordre 3, $f(x) = ax^3 + bx^2 + cx + d$.

Chapitre 4

Équations différentielles ordinaires

4.1 Introduction

Pour illustrer le concept d'équations différentielles, nous allons considérer pour commencer des systèmes qui évoluent dans le temps (évolution d'une population, taux d'intérêts, circuits électriques, ...).

4.1.1 Mouvement rectiligne uniforme

Imaginons que nous connaissons la fonction décrivant le vitesse d'une particule au cours du temps et notons la $v(t)$. Nous savons également que la vitesse d'une particule est reliée à l'évolution au cours du temps de sa position. Cette dernière peut être notée, $x(t)$. En particulier, nous avons que la vitesse n'est rien d'autre que la dérivée de la position. On peut donc écrire une équation reliant la vitesse à la position

$$x'(t) = v(t). \tag{4.1}$$

Cette équation est appelée *équation différentielle*, car elle fait intervenir non seulement les fonctions $x(t)$ et $v(t)$, mais également la dérivée de la fonction $x(t)$. Si maintenant nous précisons ce que vaut la fonction $v(t)$ nous pourrons résoudre cette équation. Comme le nom de la sous-section le laisse entendre, nous nous intéressons à un mouvement rectiligne uniforme, qui décrit le mouvement d'un objet qui se déplace à vitesse constante,

$$v(t) = v. \tag{4.2}$$

Nous cherchons ainsi à résoudre l'équation différentielle

$$x'(t) = v. \tag{4.3}$$

Ou en d'autres termes, nous cherchons la fonction dont la dérivée donne une constante[1]. Vous savez sans doute que l'ensemble de fonctions satisfaisant la

1. Cette formulation devrait vous rappeler ce que nous avons vu au chapitre précédent

contrainte précédente est

$$x(t) = v \cdot t + B, \tag{4.4}$$

où B est une constante arbitraire. Cette solution générale n'est pas unique, car nous obtenons une infinité de solutions (comme quand nous avons calculé la primitive d'une fonction au chapitre précédent). Afin de trouver une solution unique, nous devons imposer une condition, typiquement une "condition initiale" à notre équation différentielle. En effet, si nous imposons la condition initiale

$$x(t_0) = x_0, \tag{4.5}$$

il vient

$$x(t_0) = x_0 = v \cdot t_0 + B \Leftrightarrow B = x_0 - v \cdot t_0. \tag{4.6}$$

Finalement, la solution du problème différentiel est donnée par

$$x(t) = v \cdot (t - t_0) + x_0. \tag{4.7}$$

Remarque 10

La solution de l'équation différentielle

$$x'(t) = v,\ x(t_0) = x_0, \tag{4.8}$$

revient à calculer

$$\begin{aligned} \int x'(t)\mathrm{d}t &= \int v\mathrm{d}t, \\ x(t) &= v \cdot t + B. \end{aligned} \tag{4.9}$$

4.1.2 Mouvement rectiligne uniformément accéléré

Dans le cas du mouvement rectiligne d'un objet dont on le connaît que l'accélération, $a(t)$, on peut également écrire une équation différentielle qui décrirait l'évolution de la position de l'objet en fonction du temps. En effet, l'accélération d'un objet est la deuxième dérivée de la position, soit

$$x''(t) = a(t), \tag{4.10}$$

ou encore la première dérivée de la vitesse.

$$\begin{aligned} v'(t) &= a(t), \\ x'(t) &= v(t). \end{aligned} \tag{4.11}$$

Par simplicité supposons que l'accélération est constante, $a(t) = a$, donc que le mouvement est uniformément accéléré. On doit résoudre [2]

$$x''(t) = a, \tag{4.12}$$

ou

$$\begin{aligned} v'(t) &= a, \\ x'(t) &= v(t). \end{aligned} \tag{4.13}$$

2. On cherche la fonction dont la deuxième dérivée est une constante, a.

Pour résoudre ce système d'équations nous résolvons la première équation pour $v(t)$ pour trouver

$$v(t) = a \cdot t + C. \tag{4.14}$$

En substituant ce résultat dans l'éq. 4.13, on a

$$x'(t) = a \cdot t + C. \tag{4.15}$$

On peut ainsi directement intégrer des deux côtés comme vu dans la sous-section précédente

$$\begin{aligned} \int x'(t)\mathrm{d}t &= \int (a \cdot t + C)\mathrm{d}t, \\ x(t) &= \frac{a}{2} \cdot t^2 + C \cdot t + D. \end{aligned}$$

On voit que la position d'un objet en mouvement rectiligne uniformément accéléré est donné par une parabole. Cette équation a néanmoins encore deux constantes indéterminées. Pour les déterminer, on doit imposer deux conditions initiales. Une possibilité est d'imposer une condition initiale par équation

$$v(t_0) = v_0, \text{ et } x(t_0) = x_0. \tag{4.16}$$

On obtient

$$v(t_0) = v_0 = a \cdot t_0 + C \Leftrightarrow C = v_0 - a \cdot t_0, \tag{4.17}$$

et

$$x(t_0) = x_0 = \frac{a}{2} \cdot t_0^2 + D \Leftrightarrow D = x_0 - \frac{a}{2} \cdot t_0^2. \tag{4.18}$$

Finalement la solution est donnée par

$$x(t) = \frac{a}{2} \cdot (t^2 - t_0^2) + v_0 \cdot (t - t_0) + x_0. \tag{4.19}$$

Remarque 11

La solution du problème différentiel peut également se calculer de la façon suivante

$$x''(t) = a, \; x(t_0) = x_0, \; v(t_0) = v_0. \tag{4.20}$$

revient à calculer

$$\begin{aligned} \int\int x'' &= \int\int a, \\ x(t) &= \frac{a}{2}t^2 + C \cdot t + D. \end{aligned} \tag{4.21}$$

4.1.3 Évolution d'une population

Imaginons une colonie de bactéries dont nous connaissons le taux de reproduction r. Nous connaissons le nombre de ces bactéries au temps t, qui est donné par $n(t)$. Nous souhaitons connaître la population au temps $t + \delta t$. On a donc

$$n(t + \delta t) = n(t) + (r\delta t) \cdot n(t) = n(t)(1 + r\delta t). \tag{4.22}$$

Imaginons que le taux de reproduction $r = 1/3600s^{-1}$, que la population à un temps donné t_0 est de $n(t_0) = 1000$, et qu'on veuille connaître la population après $\delta t = 1h = 3600s$. Il vient alors

$$n(t_0 + 3600) = (1 + 1/3600 \cdot 3600) \cdot n(t_0) = 2 \cdot 1000 = 2000. \tag{4.23}$$

Imaginons maintenant que nous voulions calculer la population après $\delta t = 2h = 7200s$. Nous avons deux façons de faire. Soit nous utilisons le résultat précédent $n(t_1) = 2000$ avec $t_1 = t_0 + 3600$ et évaluons la population après une heure supplémentaire ($\delta t_1 = 3600s$)

$$n(t_1 + 3600) = (1 + 1/3600 \cdot 3600) \cdot n(t_1) = 2 \cdot 2000 = 4000. \tag{4.24}$$

Soit nous reprenons l'équation de départ (voir l'éq. 4.22) et nous obtenons

$$n(t_0 + 7200) = (1 + 1/3600 \cdot 7200) \cdot n(t_0) = 3 \cdot 1000 = 3000. \tag{4.25}$$

On voit que ces deux résultats ne sont pas égaux. Effectuer deux itérations de notre algorithme discret avec un pas d'itération de δt, ne correspond pas à effectuer une seule itération avec un pas deux fois plus grand ($2\delta t$). Néanmoins cela devrait être le cas pour $\delta t \to 0$.

Pour nous en convaincre faisons l'exercice suivant. Reprenons l'éq. 4.24 que vous pouvons réécrire comme

$$n(t_0 + 2\delta t) = n(t_1 + \delta t) = (1 + r\delta t)n(t_1) = (1 + r\delta t)(1 + r\delta t)n(t_0) = (1 + r\delta t)^2 n(t_0). \tag{4.26}$$

Si à présent nous comparons les résultats obtenus pour $\delta t_1 = 2\delta t$ dans l'éq. 4.22 on a

$$\begin{aligned} n_1 &= (1 + r\delta t)^2 n(t_0) = (1 + 2r\delta t + (r\delta t)^2)n(t_0), \\ n_2 &= (1 + 2r\delta t)n(t_0). \end{aligned} \tag{4.27}$$

On trouve donc finalement que $n_2 - n_1 = (r\delta t)^2 n(t_0)$. On a donc que la différence tend bien vers 0 quand δt tend vers 0.

Afin de voir plus en détail ce qu'il se passe lorsque $\delta t \to 0$, reprenons l'équation de départ (l'éq. 4.22), divisons la par δt et arrangeons les termes. Il vient

$$\frac{n(t + \delta t) - n(t)}{\delta t} = r \cdot n(t). \tag{4.28}$$

En prenant la limite $\delta t \to 0$ on voit apparaître la dérivée dans le membre de gauche de l'équation ci-dessus

$$\lim_{\delta t \to 0} \frac{n(t + \delta t) - n(t)}{\delta t} = n'(t) = r \cdot n(t). \tag{4.29}$$

On voit qu'on a construit ici une équation différentielle à partir d'un système discret.

Nous pouvons à présent résoudre l'équation différentielle ci-dessus en se souvenant que la fonction dont la dérivée est proportionnelle à la fonction de départ est l'exponentielle. Il vient

$$n(t) = C \exp(rt), \tag{4.30}$$

où C est une constante. Il est en effet élémentaire de montrer que cette solution satisfait l'éq. 4.29. On voit également qu'il nous manque une condition pour avoir l'unicité de la solution ci-dessus (on ne connaît toujours pas C). La constante peut-être obtenue à l'aide d'une condition initiale (correspondant au $n(t_0)$ de tout à l'heure). Si $n(t_0) = n_0$, nous trouvons pour C

$$n(t_0) = C\exp(rt_0) = n_0 \Leftrightarrow C = n_0\exp(-rt_0). \tag{4.31}$$

substituant cette relation dans l'éq. 4.30, on obtient

$$n(t) = n_0\exp(r(t - t_0)). \tag{4.32}$$

4.1.4 Autres illustrations de l'utilisation des équations différentielles

La plupart des systèmes naturels (ou moins naturels) peuvent être décrits à l'aide d'équations différentielles. Nous allons en écrire quelques exemples ci-dessous.

4.1.4.1 Systèmes proies-prédateurs

Considérons un système où nous avons des prédateurs (des guépards) et des proies (des antilopes) [3]. Supposons que les antilopes se reproduisent exponentiellement vite et que leur seul moyen de mourir est de se faire manger par les guépards et que la chance de se faire manger est proportionnelle au nombre de guépards. Les guépards meurent exponentiellement vite de faim et se reproduisent proportionnellement au nombre d'antilopes se trouvant dans le système.

Avec ces hypothèses, on peut écrire le système d'équations suivant (a est le nombre d'antilopes, et g le nombre de guépards)

$$\begin{aligned}\frac{\mathrm{d}a}{\mathrm{d}t} &= \underbrace{k_a a(t)}_{(1)} - \underbrace{k_{g,a} g(t)a(t)}_{(2)},\\ \frac{\mathrm{d}g}{\mathrm{d}t} &= -\underbrace{k_g g(t)}_{(3)} + \underbrace{k_{a,g} a(t)g(t)}_{(4)}\end{aligned} \tag{4.33}$$

Le terme (1) représente la reproduction des antilopes avec taux k_a. Le terme (2) représente la mort des antilopes qui se font manger par les guépards avec un taux $k_{g,a}$ (la chance qu'un guépard rencontre une antilope). Le terme (3) est la mort des guépards avec un taux k_g. Finalement le terme (4) est la reproduction des guépards proportionnelle au nombre d'antilopes avec un taux $k_{a,g}$.

Nous avons à faire ici à un système d'équations différentielles. Nous n'allons pas nous intéresser aux détails de la résolution de ce système mais simplement étudier le comportement de la solution (voir la fig. 4.1 et fig. 4.2).

4.1.4.2 Circuits électriques: le circuit RC

Supposons que nous ayons le circuit RC de la Fig. fig. 4.3, où nous avons une résistance (de résistance R) branchée en série avec une capacité (de capacité

3. Ces systèmes sont dits de Lotka–Volterra.

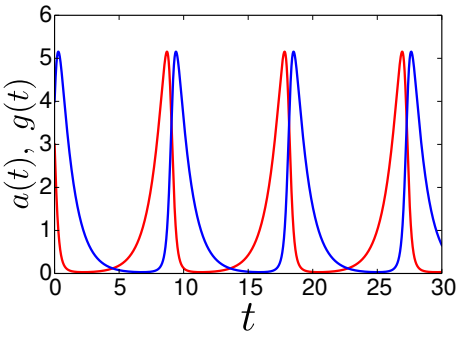

FIGURE 4.1 – L'évolution au cours du temps de la population d'antilopes et de guépards.

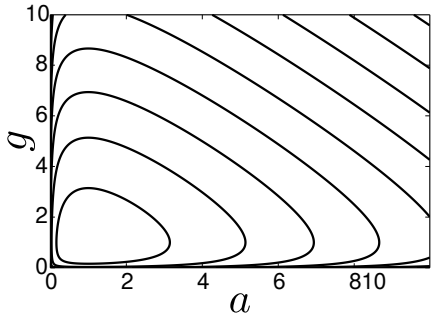

FIGURE 4.2 – Représentation paramétrique de l'évolution population d'antilopes et de guépards.

électrique C). Sur ce circuit nous avons une source qui délivre une tension U. Nous avons également un interrupteur qui quand il est en position (a) relie le circuit RC à la source, ce qui a pour effet de chargé la capacité. En position (b) la capacité se décharge et son énergie est dissipée dans la résistance.

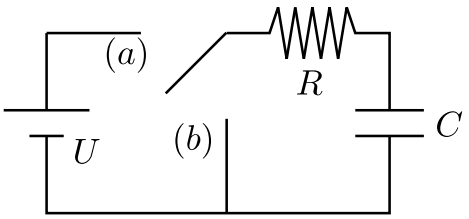

FIGURE 4.3 – Le circuit RC.

Nous souhaitons étudier la variation de la chute de tension dans la capacité U_c lorsque:

1. nous mettons l'interrupteur en position (a).
2. puis lorsque la capacité est chargée, nous mettons l'interrupteur en position (b).

Les chutes de tension dans la capacité et la résistance sont respectivement données par

$$U_C = Q/C, \quad U_R = RI, \tag{4.34}$$

où Q est la charge de la capacité et I le courant traversant la résistance. Nous avons par la loi de Kirchoff que

$$U = U_C + U_R. \tag{4.35}$$

De plus le courant traversant la résistance est donné par

$$I(t) = Q'(t). \tag{4.36}$$

En combinant ces équations, nous obtenons

$$U_C'(t) + \frac{U_C(t)}{RC} = \frac{U}{RC}. \tag{4.37}$$

Nous avons également la condition initiale $U_C(0) = 0$ (la tension au moment de la mise de l'interrupteur en position (a) est nulle).

Lors de la mise de l'interrupteur en position (b) nous avons simplement que l'éq. 4.35 devient

$$0 = U_C + U_R. \tag{4.38}$$

On a donc que l'équation différentielle pour l'évolution de la chute de tension dans la capacité devient

$$U_C'(t) + \frac{U_C(t)}{RC} = 0. \tag{4.39}$$

Et la condition initiale devient $U_C(0) = U$.

Pour cette dernière équation nous avons déjà calculé une solution très similaire et on a

$$U_C(t) = U \exp(-t/(RC)). \tag{4.40}$$

La tension dans la capacité va décroître exponentiellement vite. Pour le cas de l'interrupteur en position (a) la solution est

$$U_C(t) = U(1 - \exp(-t/(RC))). \tag{4.41}$$

La tension augmente exponentiellement au début, puis au fur et à mesure que la capacité se charge il devient de plus en plus difficile de la charger. L'augmentation de la tension se fait donc de plus en plus lentement jusqu'à devenir une asymptote horizontale en U.

4.1.4.3 Taux d'intérêts composés

Nous voulons étudier l'augmentation d'un capital $c(t)$ au cours du temps qui est soumis à un taux d'intérêt annuel r qui est composé après chaque intervalle δt. On peut également inclure des dépôts/retraits d sur l'intervalle δt. La valeur du capital après un intervalle δt est de

$$c(t + \delta t) = c(t) + (r\delta t)c(t) + d\delta t. \tag{4.42}$$

Supposons qu'on a un capital de départ 1000CHF, un taux d'intérêts annuel de 1% et un dépôt annuel de 100CHF. Après deux mois ($\delta t = 2/12 = 1/6$) le capital devient

$$c(1/6) = 1000 + 0.01/6 \cdot 1000 + 100/6 = 1018.3\text{CHF}. \tag{4.43}$$

Si maintenant, nous voulons avoir la valeur du capital à n'importe quel moment dans le temps, nous allons prendre $\delta t \to 0$. En divisant l'éq. 4.42 par δt, et en réarrangeant les termes, on obtient

$$c'(t) = rc(t) + d. \tag{4.44}$$

En supposant que $c(t = 0) = c_0$ (le capital initial), cette équation différentielle a pour solution

$$c(t) = \frac{d}{r}(e^{rt} - 1) + c_0 e^{rt}. \tag{4.45}$$

Cette solution a pour les paramètres précédents la forme suivante sur une période de 100 ans.

4.2 Définitions et théorèmes principaux

Définition 14 *(Équation différentielle ordinaire)*

Soit y une fonction dérivable n fois et dépendant d'une seule variable. Une **équation différentielle ordinaire** est un équation de la forme

$$F(x, y, y', y'', ..., y^{(n)}) = 0, \tag{4.46}$$

où F est une fonction, et y', y'', ..., $y^{(n)}$ sont les dérivées première, deuxième, ..., n-ème de y.

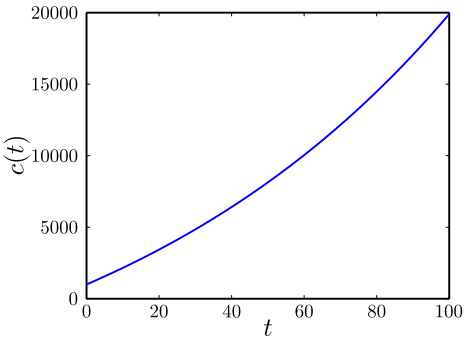

FIGURE 4.4 – L'évolution du capital c en fonction du temps sur 100 ans.

Illustration 16

L'équation suivante est une équation différentielle ordinaire

$$y'' + 4y' + 8y + 3x^2 + 9 = 0. \tag{4.47}$$

Afin de résoudre cette équation, nous cherchons une solution de la forme $y = f(x)$. On dit également que nous cherchons à intégrer l'équation différentielle.

Afin de classifier les équation différentielles, considérons les définitions suivantes

Définition 15 *(Ordre)*

L'ordre d'une équation différentielle est l'ordre le plus haut des dérivées de y qui y apparaissent. L'ordre de l'équation différentielle $F(x, y, y', y'', ..., y^{(n)}) = 0$ est de n, si $n \neq 0$.

Illustration 17

L'équation différentielle suivante est d'ordre 3

$$4y''' + x \cdot y' + 4y + 6x = 0. \tag{4.48}$$

Définition 16 *(Condition initiale)*

Une condition initiale pour une équation différentielle d'ordre n, est un ensemble de valeurs, y_0, y_1, ..., y_{n-1} donnée telles que pour une valeur x_0 donnée on a

$$y(x_0) = y_0,\ y'(x_0) = y_1,\ ...,\ y^{(n-1)}(x_0) = y_{n-1}. \tag{4.49}$$

Nous souhaitons maintenant savoir sous quelles conditions une équation différentielle admet une solution et si elle est unique. Nous n'allons pas vraiment écrire ni démontrer le théorème d'existence et d'unicité des équations différentielles ordinaires, mais simplement en donner une version approximative et la discuter

---

Théorème 5 *(Existence et unicité)*

Soit $D \subseteq \mathbb{R}$ le domaine de définition de la fonction y. Soit $y : D \to E \subseteq \mathbb{R}$ une fonction à valeur réelle continue et dérivable sur D, et $f : D \times E \to F \subseteq \mathbb{R}$ une fonction à deux variables continue sur $D \times E$. Alors, le système suivant (également appelé problème de Cauchy)

$$\begin{aligned} y' &= f(y, x), \\ y(x = x_0) &= y_0, \end{aligned} \tag{4.50}$$

admet une unique solution $y(x)$.

Ce théorème peut être étendu à une équation d'un ordre arbitraire, n, possédant $n-1$ conditions initiales. En effet, n'importe quel équation différentielle d'un ordre n peut être réécrite sous la forme de n équations différentielles d'ordre 1. Pour illustrer cette propriété considérons l'équation différentielle suivante

$$y'' + 3y' + y + 3x = 0. \tag{4.51}$$

Si nous définissons $z = y'$, nous avons le système suivant à résoudre

$$\begin{aligned} y' &= z, \\ z' + 3y' + y + 3x &= 0. \end{aligned} \tag{4.52}$$

Nous voyons que ce système est d'ordre 1, mais que nous avons augmenté le nombre d'équations à résoudre.

Cette propriété peut se généraliser de la façon suivante. Soit une équation différentielle d'ordre n

$$F(x, y, y', ..., y^{(n)}) = 0. \tag{4.53}$$

Nous pouvons définir $z_i = y^{(i-1)}$ et on aura donc que $z_{i+1} = z_i'$. On peut ainsi réécrire l'équation différentielle d'ordre n comme étant

$$\begin{aligned} z_{i+1} &= z_i', \ i = 1, ..., n-1 \\ F(x, y, y', .., y^{(n)}) = 0 \Rightarrow & G(x, z_1, z_2, ..., z_n) = 0. \end{aligned} \tag{4.54}$$

Jusqu'ici F peut être totalement arbitraire. Essayons de classifier un peu les équations différentielles en fonction des propriétés de F.

Définition 17 *(Linéarité)*

Une équation différentielle ordinaire d'ordre n est dite linéaire si on peut l'écrire sous la forme

$$a_0(x) \cdot y(x) + a_1(x) \cdot y'(x) + ... + a_n(x) \cdot y^{(n)}(x) = b(x). \tag{4.55}$$

Si les coefficients a_i ne dépendent pas de x, alors l'équation est dite à **coefficients constants**.

L'équation ci-dessus a les propriétés suivantes

1. Les a_i ne dépendent que de x (ils ne peuvent pas dépendre de y).
2. Les y et toutes leur dérivées ont un degré polynomial de 1.

Illustration 18

L'équation suivante est linéaire

$$y'' + 4x \cdot y' = e^x. \tag{4.56}$$

L'équation suivante n'est pas linéaire

$$y \cdot y'' + 4x \cdot y' = e^x. \tag{4.57}$$

Définition 18 *(Homogénéité)*

Une équation différentielle ordinaire est dite homogène si le terme dépendant uniquement de x est nul. Dans le cas où nous avons à faire à une équation différentielle linéaire, cela revient à dire que $b(x) = 0$.

Illustration 19 *(Homogénéité)*

Les équations suivantes sont homogènes

$$\begin{aligned} &y'' + 4x \cdot y \cdot y' + 3x^2 \cdot y^3 = 0, \\ &2y''' + 5x^2 \cdot y' = 0. \end{aligned} \tag{4.58}$$

Les équations suivantes ne le sont pas

$$\begin{aligned} &y'' + 4x \cdot y \cdot y' + 3x^2 \cdot y^3 = 4x + 2, \\ &2y''' + 5x^2 \cdot y' = 1. \end{aligned} \tag{4.59}$$

Exercice 15 *(Homogénéité)*

Pour chacune de ces équations différentielles ordinaires donner tous les qualificatifs possibles. Si l'équation est inhomogène donner l'équation homogène associée.

$$\begin{aligned} &y^{(4)} + 4x^2 y = 0, \\ &y' + 4x^2 y^2 = 3x + 2, \\ &\frac{1}{y+1} y'' + 4x^2 y^2 = 0, \\ &y' = y, \\ &4y'' + 4xy = 1. \end{aligned} \tag{4.60}$$

La solution des équations différentielles inhomogènes se trouve de la façon suivante.

1. Trouver la solution générale de l'équation différentielle homogène associée, notons-la $y_h(x)$.
2. Trouver une solution particulière à l'équation inhomogène, notons-la $y_0(x)$.

La solution sera donnée par la somme de ces deux solutions

$$y = y_h + y_0. \tag{4.61}$$

4.3 Techniques de résolution d'équations différentielles ordinaires d'ordre 1

Ici nous considérerons uniquement les équations différentielles ordinaires d'ordre 1. Pour certains types d'équations différentielles, il existe des techniques standard pour les résoudre. Nous allons en voir un certain nombre.

4.3.1 Équations à variables séparables

Définition 19 *(Équations à variable séparables)*

On dit qu'une équation différentielle d'ordre 1 est à variables séparables, si elle peut s'écrire sous la forme suivante

$$y' a(y) = b(x). \tag{4.62}$$

Illustration 20

L'équation suivante est à variables séparables

$$e^{x^2 + y^2(x)} y'(x) = 1. \tag{4.63}$$

Pour ce genre d'équations, la solution se trouve de la façon suivante. Nous commençons par écrire la dérivée, $y' = \mathrm{d}y/\mathrm{d}x$ et on obtient

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}x} a(y) &= b(x), \\ a(y)\mathrm{d}y &= b(x)\mathrm{d}x. \end{aligned} \tag{4.64}$$

On peut maintenant simplement intégrer des deux côtés et on obtient

$$\int a(y)\mathrm{d}y = \int b(x)\mathrm{d}x. \tag{4.65}$$

Si nous parvenons à résoudre les intégrales nous obtenons une solution pour $y(x)$ (cette solution n'est peut-être pas explicite). Il existe le cas simple où $a(y) = 1$ et il vient

$$y = \int b(x)\mathrm{d}x. \tag{4.66}$$

Exemple 7

Résoudre l'équation différentielle suivante

$$n'(t) = r \cdot n(t). \tag{4.67}$$

Solution 7

En écrivant $n' = \mathrm{d}n/\mathrm{d}t$, on réécrit l'équation différentielle sous la forme

$$\frac{1}{n}\mathrm{d}n = r\mathrm{d}t, \tag{4.68}$$

qu'on intègre

$$\begin{aligned} \int \frac{1}{n}\mathrm{d}n &= \int r\mathrm{d}t, \\ \ln(n) &= r \cdot t + C, \\ n(t) &= e^{r\cdot t + C} = A \cdot e^{r\cdot t}, \end{aligned}$$

où $A = e^C$.

Exercice 16

1. Résoudre l'équation différentielle suivante

$$c'(t) = rc(t) + d. \tag{4.69}$$

2. Résoudre l'équation différentielle suivante

$$x \cdot y(x) \cdot y'(x) = 1. \tag{4.70}$$

4.3.2 Équations linéaires

Pour une équation du type

$$y'(x) = a(x) \cdot y(x) + b(x), \tag{4.71}$$

on doit résoudre le problème en deux parties.

supposons que nous connaissons une solution "particulière" à cette équation. Notons la y_p. Si nous faisons maintenant le changement de variables $y = y_h + y_p$ et remplaçons ce changement de variables dans l'équation ci-dessus nous obtenons

$$y'_p(x) + y'_h(x) = a(x) \cdot y_p(x) + a(x) \cdot y_h(x) + b(x). \tag{4.72}$$

Comme y_p est solution de l'éq. 4.71 on a

$$y'_p(x) = a(x) \cdot y_p(x) + b(x). \tag{4.73}$$

En remplaçant cette relation dans l'éq. 4.72 il vient

$$y'_h(x) = a(x) \cdot y_h(x). \tag{4.74}$$

Cette équation différentielle n'est rien d'autre que l'équation homogène correspondant à éq. 4.71.

Nous voyons qu'une équation inhomogène se résout en trouvant la solution générale à l'équation homogène correspondante et en y ajoutant une solution particulière.

Revenons donc à la résolution de l'équation différentielle linéaire d'ordre un. La première partie de la résolution consiste à résoudre l'équation homogène associée à l'éq. 4.71

$$y'(x) = a(x) \cdot y(x). \tag{4.75}$$

Cette équation se résout par séparation des variables. La solution est donc

$$y_h(x) = Ce^{\int a(x)\mathrm{d}x}. \tag{4.76}$$

Puis nous devons chercher une solution dite particulière de l'équation inhomogène. Pour ce faire nous utilisons la méthode de la variation de la constante. Il s'agit de trouver une solution particulière qui aura la même forme que la solution de l'équation homogène, où C dépendra de x (d'où le nom de méthode de variation de la constante)

$$y_p(x) = C(x)e^{\int a(x)\mathrm{d}x}. \tag{4.77}$$

En remplaçant cette équation dans l'éq. 4.71, on obtient

$$C'(x)e^{\int a(x)\mathrm{d}x} + C(x) \cdot a(x)e^{\int a(x)\mathrm{d}x} = a(x) \cdot C(x)e^{\int a(x)\mathrm{d}x} + b(x),$$
$$C'(x) = \frac{b(x)}{e^{\int a(x)\mathrm{d}x}}.$$

Il nous reste donc à résoudre cette équation différentielle pour $C(x)$ qui est une équation à variables séparables où on aurait un $a(c) = 1$. On intègre donc directement cette équation pour obtenir

$$C(x) = \int \frac{b(x)}{e^{\int a(x)\mathrm{d}x}}\mathrm{d}x. \tag{4.78}$$

Finalement, on a que la solution de l'équation générale de l'équation inhomogène est

$$y = y_p + y_h = \left(\int \frac{b(x)}{e^{\int a(x)\mathrm{d}x}} \mathrm{d}x + C \right) e^{\int a(x)\mathrm{d}x}. \tag{4.79}$$

Exemple 8

Résoudre l'équation suivante

$$U'_C(t) + \frac{U_C(t)}{RC} = \frac{U}{RC}. \tag{4.80}$$

Solution 8

On commence par résoudre l'équation homogène

$$U'_{Ch}(t) + \frac{U_{Ch}(t)}{RC} = 0. \tag{4.81}$$

D'où on obtient

$$U_{Ch} = A \cdot e^{-\frac{1}{RC}t}. \tag{4.82}$$

Puis par variations des constantes, on essaie de déterminer la solution particulière de la forme

$$U_{Cp} = B(t) \cdot e^{-\frac{1}{RC}t}. \tag{4.83}$$

En remplaçant cette forme de solution dans l'éq. 4.80, on obtient

$$B'(t) = \frac{U}{RC} \cdot e^{\frac{1}{RC}t}. \tag{4.84}$$

Qui donne par intégration

$$B(t) = Ue^{\frac{1}{RC}t} + D. \tag{4.85}$$

Finalement, il vient que

$$U_c(t) = \left(Ue^{\frac{1}{RC}t} + D + A \right) e^{-\frac{1}{RC}t} = U + (D + A)e^{-\frac{1}{RC}t} = U + Ce^{-\frac{1}{RC}t}, \tag{4.86}$$

où $C = D + A$. Pour le cas de la charge du condensateur, on a de plus $U_c(0) = 0$. On peut donc fixer la constante $C = -U$.

Résoudre les équations différentielles suivantes

Exercice 17

1. $$y' + 2y = t^2 \tag{4.87}$$
2. $$y' + y = \frac{1}{1 + e^t}. \tag{4.88}$$

4.3.3 Équations de Bernoulli

Il existe des équations particulières qui peuvent se ramener à des équations linéaires via des changements de variables.

Une classe particulière sont les équations de Bernoulli, qui s'écrit

$$y'(x) + a(x) \cdot y(x) + b(x) \cdot y^n(x) = 0, \tag{4.89}$$

où $r \in \mathbb{R}$.

Cette équation peut être réécrite sous la forme

$$\frac{y'(x)}{y^n(x)} + \frac{a(x)}{y^{n-1}(x)} + b(x) = 0. \tag{4.90}$$

Dans ce cas là, en effectuant le changement de variable suivant

$$z = y^{1-n}, \tag{4.91}$$

on obtient (exercice)

$$z'(x) + (1-n)a(x) \cdot z(x) + (1-n)b(x) = 0. \tag{4.92}$$

On a donc ramené l'équation de Bernoulli à une équation linéaire que nous savons résoudre à l'aide de la méthode de la section sec. 4.3.2.

Exemple 9

Résoudre l'équation de Bernoulli suivante

$$y' - y - x \cdot y^6 = 0. \tag{4.93}$$

Solution 9

Avec la substitution $z = y^5$, on obtient

$$z' - 5z + 5x = 0. \tag{4.94}$$

Cette équation se résout en trouvant d'abord la solution de l'équation homogène

$$z'_h - 5z_h = 0, \tag{4.95}$$

qui est donnée par

$$z_h = Ae^{5x}. \tag{4.96}$$

En remarquant qu'une solution particulière à $z'_p - 5z_p + 5x = 0$, peut être de la forme $z_p = x + B$ (avec B une constante) on obtient

$$1 - 5(x + B) + 5x = 0,$$
$$1 - 5B = 0 \Rightarrow B = \frac{1}{5}.$$

Et finalement

$$z = z_h + z_p = Ae^{5x} + x + \frac{1}{5}. \tag{4.97}$$

Il nous reste à présent à calculer $y = z^{1/5}$ et on a

$$y = \left(Ae^{5x} + x + \frac{1}{5}\right)^{1/5}. \tag{4.98}$$

4.3.4 Équation de Riccati

L'équation de Riccati qui est de la forme

$$y'(x) + a(x) + b(x) \cdot y(x) + c(x) \cdot y^2(x) = 0, \tag{4.99}$$

et est donc quadratique en y. On notera que c'est une équation de Bernoulli (avec $n = 2$ et qui est inhomogène).

Cette équation a une propriété intéressante. Si nous connaissons une solution particulière à l'équation inhomogène, notons la y_p, alors la solution générale peut être trouvée de la façon suivante.

Faisons le changement de variable suivant $y = y_h + y_p$. L'équation ce-dessus devient donc

$$y_p' + y_h' + a(x) + b(x) \cdot y_p + b(x) \cdot y_h + c(x) \cdot (y_p^2 + 2y_p(x)y_h(x) + y_h^2) = 0. \tag{4.100}$$

En utilisant que y_p est solution de l'équation de Riccati, on a

$$y_h' + a(x) + (b(x) + 2y_p(x)c(x)) \cdot y_h + c(x) \cdot y_h^2 = 0. \tag{4.101}$$

Cette équation est une équation de Bernoulli avec $n = 2$. On sait donc comment la résoudre.

–

Exercice 18

Résoudre l'équation de Riccati suivante

$$y' + y^2 - \frac{2}{x^2} = 0. \tag{4.102}$$

Indication: la solution particulière a la forme $y = \frac{a}{x}$, avec a une constante.

–

De plus, ce genre d'équation peut-être transformée via un changement de variables en une équation linéaire d'ordre deux. Si $c(x)$ est dérivable, alors on peut faire le changement de variables suivant

$$v = y \cdot c(x), \tag{4.103}$$

et on a donc que

$$v' = y'c + yc'. \tag{4.104}$$

En insérant ces relations dans l'éq. 4.99, il vient

$$v'(x) + d(x) + e(x) \cdot v(x) + v^2(x) = 0, \tag{4.105}$$

où nous avons nommé $d(x) = a(x) \cdot c(x)$ et $e(x) = \frac{c'(x)}{c(x)} + b(x)$. Si à présent nous faisons un autre changement de variables

$$v(x) = -\frac{z'(x)}{z(x)}, \tag{4.106}$$

on obtient que l'équation ci-dessus peut se réécrire comme

$$z''(x) + e(x) \cdot z'(x) + d(x) \cdot z(x) = 0. \tag{4.107}$$

L'équation de Riccati (une équation d'ordre un non-linéaire et inhomogène) est ainsi transformée en une équation linéaire d'ordre deux.

4.4 Equations différentielles ordinaires d'ordre deux

Dans cette section, nous allons étudier des cas particuliers d'équations différentielles que nous savons intégrer. Cela sera toujours des équations linéaires.

De façon générale ces équations s'écrivent

$$a(x)y''(x) + b(x)y'(x) + c(x)y(x) = d(x), \tag{4.108}$$

où $a, b, c, d : \mathbb{R} \to \mathbb{R}$ sont des fonctions réelles. Avant de résoudre l'équation générale, nous allons considérer des plus simples.

4.4.1 EDO d'ordre deux homogène à coefficients constants

Ce genre d'équations s'écrit sous la forme

$$ay''(x) + by'(x) + cy(x) = 0. \tag{4.109}$$

Voyons maintenant comment résoudre cette équation.

Ces équations ont des propriétés intéressantes dûes à la linéarité de l'équation différentielle.

Propriétés 5

Ces propriétés (qui caractérisent le mot "linéaires") sont à démontrer en exercice.

1. Soit $f(x)$ une solution de l'éq. 4.109, alors pour $C \in \mathbb{R}$ $Cf(x)$ est également solution de éq. 4.109.
2. Soient $f(x)$ et $g(x)$ deux solutions de l'équation éq. 4.109, alors $h(x) = f(x) + g(x)$ est également solution de éq. 4.109.
3. De ces deux propriétés, on déduit la propriété suivante. Soient $f(x)$ et $g(x)$ deux solutions de l'éq. 4.109, et $C_1, C_2 \in \mathbb{R}$, $h(x) = C_1 f(x) + C_2 g(x)$ est aussi solution de l'éq. 4.109.

Afin de simplifier la discussion prenons une EDO d'ordre deux à coefficients constants particulière

$$y'' + 3y' + 2y = 0. \tag{4.110}$$

On va supposer que cette équation a pour solution une fonction de la forme $y(x) = e^{\lambda x}$. Substituons cette forme de solution dans l'équation de départ, on obtient

$$\begin{aligned} \lambda^2 e^{\lambda x} + 3\lambda e^{\lambda x} + 2\lambda^2 e^{\lambda x} &= 0, \\ \lambda^2 + 3\lambda + 2 &= 0, \end{aligned}$$

s où on a utilisé que $e^{\lambda x}$ ne peut jamais s'annuler pour le simplifier entre les deux lignes. La seconde ligne ci-dessus, s'appelle le polynôme caractéristique de notre EDO d'ordre 2.

Il nous reste à présent à déterminer λ ce qui est un simple problème d'algèbre. Le polynôme ci-dessus se factorise simplement en

$$(\lambda + 1)(\lambda + 2) = 0, \tag{4.111}$$

on a donc pour solution $\lambda = -1$, et $\lambda = -2$.

On a donc immédiatement deux solutions à notre équation différentielle

$$y_1(x) = e^{-x}, \quad y_2(x) = e^{-2x}. \tag{4.112}$$

On vérifie aisément que ces deux équations vérifient l'éq. 4.110. Précédemment, nous avons vu que la linéarité de ces équations différentielles, faisait qu'on pouvait construire des solutions plus générales. En effet, on peut montrer que la solution la plus générale à cette EDO est

$$y(x) = C_1 y_1(x) + C_2 y_2(x) = C_1 e^{-x} + C_2 e^{-2x}. \tag{4.113}$$

On constate qu'il y a deux constantes à déterminer pour avoir une solution unique. Pour ce faire il faudra donner deux conditions initiales. Une sur $y(x)$ et une sur $y'(x)$. Par exemple on pourrait avoir $y(0) = 1$ et $y'(0) = 0$ et on obtient

$$\begin{aligned} C_1 + C_2 &= 1, \\ -C_1 - 2C_2 &= 0. \end{aligned} \tag{4.114}$$

Ce système d'équations ordinaires a pour solution

$$C_1 = 2, \quad C_2 = -1. \tag{4.115}$$

On a donc finalement

$$y(x) = 2e^{-x} - e^{-2x}. \tag{4.116}$$

A présent, nous pouvons généraliser cette méthode pour l'équation éq. 4.109

$$ay''(x) + by'(x) + cy(x) = 0. \tag{4.117}$$

En faisans la même substitution que précédemment, $y = e^{\lambda x}$, on a

$$\begin{aligned} a\lambda^2 e^{\lambda x} + b\lambda e^{\lambda x} + ce^{\lambda x} &= 0, \\ a\lambda^2 + \lambda b + c &= 0. \end{aligned} \tag{4.118}$$

L'équation ci-dessus doit être résolue pour λ. Nous savons comment résoudre ce genre d'équation du second degré. La solution est donnée par

$$\lambda = \frac{-b \pm \sqrt{\Delta}}{2a}, \tag{4.119}$$

où $\Delta = b^2 - 4ac$. On a deux solutions

$$\begin{aligned} \lambda_1 &= \frac{-b - \sqrt{\Delta}}{2a}, \\ \lambda_2 &= \frac{-b + \sqrt{\Delta}}{2a}. \end{aligned} \tag{4.120}$$

Il y a trois cas possibles: $\Delta > 0$, $\Delta = 0$, $\Delta < 0$.

4.4.1.1 Le cas $\Delta > 0$

Dans ce cas, on a que $\lambda_1, \lambda_2 \in \mathbb{R}$ sont réels. La solution est donc donnée par (comme on l'a vu au paravent)

$$y(x) = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x}. \tag{4.121}$$

4.4.1.2 Le cas $\Delta = 0$

Ici, $\lambda_1 = \lambda_2 = \lambda = -b/(2a)$ et λ est réel. Dans ce cas-là les choses se compliquent un peu. Si on utilisait directement la formule ci-dessus, on aurait

$$y(x) = Ce^{\lambda x}, \tag{4.122}$$

avec $C \in \mathbb{R}$. Par contre, cette solution ne peut pas satisfaire deux conditions initiales comme nous avons vu précédemment. Il fau donc travailler un peu plus. Supposons que $y(x)$ est donné par la fonction suivante

$$y(x) = z(x)e^{\lambda x}, \tag{4.123}$$

avec $z(x)$ une fonction réelle. En substituant cela dans l'équation générale, on a

$$az'' + (2\lambda a + b)z' + (a\lambda^2 + b\lambda + c)z = 0. \tag{4.124}$$

En utilisant que $\lambda = -b/(2a)$ et $\Delta = 0$ il vient

$$z'' = 0. \tag{4.125}$$

La solution de cette équation est

$$z = C_1 + xC_2. \tag{4.126}$$

On obtient donc comme solution générale de l'équation différentielle

$$y(x) = (C_1 + C_2 x)e^{\lambda x}. \tag{4.127}$$

4.4.1.3 Le cas $\Delta < 0$

Dans ce cas-là, on a deux solutions complexes (la racine d'une nombre négatif n'est pas réelle). Les racines sont de la forme

$$\lambda_1 = \frac{-b + i\sqrt{|b^2 - 4ac|}}{2a}, \lambda_2 = \frac{-b - i\sqrt{|b^2 - 4ac|}}{2a}, \tag{4.128}$$

où i est l'unité imaginaire. En écrivant $u = -b/(2a)$ et $v = \sqrt{|b^2 - 4ac|}/(2a)$, on peut écrire $\lambda_1 = u + iv$ et $\lambda_2 = u - iv$. On a donc que λ_2 est le complexe conjugué de λ_1, ou $\lambda_1 = \bar{\lambda}_2$. En utilisant ces notations dans notre exponentielle, on a

$$\begin{aligned} y_1 &= e^{(u+iv)x} = e^{ux}e^{ivx}, \\ y_2 &= e^{(u-iv)x} = e^{ux}e^{-ivx}. \end{aligned} \tag{4.129}$$

En se rappelant de la linéarité des solutions des EDO linéaires, on peut écrire la forme générale de la solution comme ($C_1, C_2 \in \mathbb{R}$)

$$y = C_1 y_1 + C_2 y_2 = C_1 e^{ux}e^{ivx} + C_2 e^{ux}e^{-ivx} = e^{ux}(C_1 e^{ivx} + C_2 e^{-ivx}). \tag{4.130}$$

En utilisant la formule d'Euler

$$\begin{aligned} e^{ivx} &= (\cos(vx) + i\sin(vx)), \\ e^{-ivx} &= e^{ux}(\cos(vx) - i\sin(vx)), \end{aligned} \tag{4.131}$$

on peut réécrire l'éq. 4.130 comme

$$\begin{aligned} y &= e^{ux}\left(C_1(\cos(vx) + i\sin(vx)) + C_2(\cos(vx) - i\sin(vx))\right), \\ &= e^{ux}\left((C_1 + C_2)\cos(vx) + i(C_1 - C_2)\sin(vx))\right), \\ &= e^{ux}\left(C_3\cos(vx) + C_4\sin(vx))\right), \end{aligned}$$

où on a définit $C_3 \equiv C_1 + C_2$ et $C_4 \equiv i(C_1 - C_2)$.

Résoudre les EDO d'ordre 2 à coefficients constants suivantes:

1. $y'' + y' + y = 0$,
2. $y'' + 4y' + 5y = 0$, $y(0) = 1$, $y'(0) = 0$.
3. $y'' + 5y' + 6y = 0$, $y(0) = 2$, $y'(0) = 3$.
4. $2y'' - 5y' + 2y = 0$, $y(0) = 0$, $y'(0) = 1$.

4.5 Résolution numérique d'équations différentielles ordinaires

Pour la plupart des problèmes d'ingénierie classique, les solutions des équations différentielles sont trop compliquées à calculer analytiquement (si elles sont calculables). Il est donc nécessaire d'en obtenir des solutions approximées numériquement.

4.5.1 Problématique

Le problème à résoudre est une EDO avec condition initiale qui peut s'écrire de la façon suivante

$$y' = F(t, y), \quad y(t_0) = y_0, \tag{4.132}$$

où F est une fonction de y et de t, et où y_0 est la condition initiale. Nous cherchons donc à connaître l'évolution de $y(t)$ pour $t > t_0$.

4.5.2 Méthode de résolution: la méthode d'Euler

Afin de résoudre ce genre de problème numériquement il existe une grande quantité de techniques. Ici nous allons en considérer une relativement simple, afin d'illustrer la méthodologie (vous en verrez une autre dans le TP).

Nous cherchons donc à évaluer $y(t = t_0 + \delta t)$, étant donné y_0, δt et $F(t, y)$. Intégrons donc simplement notre EDO entre t_0 et $t_0 + \delta t$ dans un premier temps et on obtient

$$\int_{t_0}^{t_0+\delta t} y' \mathrm{d}t = \int_{t_0}^{t_0+\delta t} F(t, y)\mathrm{d}t. \tag{4.133}$$

Le théorème fondamental du calcul intégral nous dit que cette équation peut s'écrire

$$y(t_0 + \delta t) - y(t_0) = \int_{t_0}^{t_0+\delta t} F(t, y)\mathrm{d}t, \tag{4.134}$$

$$y(t_0 + \delta t) - y_0 = \int_{t_0}^{t_0+\delta t} F(t, y)\mathrm{d}t. \tag{4.135}$$

Ont doit donc intégrer le membre de droite de cette équation. Pour ce faire nous pouvons utiliser une des techniques vues dans le chapitre précédent. Par exemple, on peut choisir la méthode des rectangle à gauche. Cette équation devient

$$\begin{aligned} y(t_0 + \delta t) - y_0 &= \delta t F(t_0, y(t_0)), \\ y(t_0 + \delta t) &= y_0 + \delta t F(t_0, y(t_0)). \end{aligned}$$

Cette dernière équation nous permet donc d'évaluer $y(t_0 + \delta t)$ connaissant y_0. Cette méthode s'appelle "méthode d'Euler" et est une dite *explicite*, car $y(t_0+\delta t)$ ne dépend que de la valeur de y évaluée au temps t_0.

Si plutôt que d'utiliser la méthode des rectangle à gauche pour approximer l'intégrale de l'éq. 4.135, nous utilisons la méthodes des rectangles à droite on a

$$y(t_0 + \delta t) = y_0 + \delta t F(t_0 + \delta t, y(t_0 + \delta t)). \tag{4.136}$$

Dans ce cas, on voit que la valeur $y(t_0 + \delta t)$ est calculée par rapport à la valeur d'elle même. Dépendant de la forme de F on ne peut pas résoudre cette équation explicitement. On a donc à faire à une équation sous forme *implicite*. Cette façon d'approximer une EDO est dite méthode d'Euler implicite.

Sans entrer dans les détails, la différence entre une méthode explicite et une méthode implicite est une question de stabilité numérique. En effet, les méthodes explicites peuvent devenir numériquement instables (la solution numérique s'éloigne exponentiellement vite de la solution de l'EDO) si δt devient "trop grand" (la contrainte du la taille de δt s'appelle CFL, pour Courant-Friedrich-Lévy). Les méthodes implicites ne souffrent pas de ce problème de stabilité, en revanche elles sont plus coûteuses en temps de calcul et en complexité algorithmique, étant donné qu'elles requièrent la résolution d'une équation implicite.

Notre but initial était de connaître l'évolution de $y(t)$ pour $t > t_0$. Pour déterminer la valeur de $y(t_1)$ avec $t_N = t_0 + N\delta t$, il suffit donc d'effectuer N pas de la méthode d'intégration choisie (ici la méthode d'Euler explicite). On a donc que

$$y(t_0 + N\delta t) = y_0 + \delta t \sum_{i=1}^{N} F(t_i, y_i), \tag{4.137}$$

où $t_i = t_0 + i \cdot \delta t$ et $y_i = y(t_i)$. Le deuxième terme du membre de droite de cette équation est la même que la formule d'intégration en plusieurs pas pour la méthode du rectangle (voir l'équation éq. 3.60). On a vu que cette méthode a une erreur d'ordre δt. On peut en conclure que l'erreur que la précision de la méthode d'Euler est également d'ordre $\mathcal{O}(\delta t)$.

4.5.3 Méthode de résolution: la méthode de Verlet

Cette méthode d'intégration est utilisée pour l'intégration numérique d'EDO d'ordre deux avec une forme particulière qui est donnée par

$$x''(t) = a(x(t)), \tag{4.138}$$

où F est une fonction de $x(t)$. On a également les conditions initiales $x(t_0) = x_0$ et $x'(t_0) = v_0$. Cette forme d'équation différentielle est bien connue en physique sous la forme $\vec{F} = m\vec{a}$, qui peut s'écrire

$$\begin{aligned}\vec{F} &= m\vec{a}(t) = m\vec{x}''(t),\\ \frac{\vec{F}}{m} &= \vec{x}''(t),\end{aligned}$$

qui est de la forme de l'EDO de départ de l'éq. 4.138. La force peut avoir différentes forme. Cela peut être la forme de gravité $\vec{F} = m\vec{g}$, de frottement $\vec{F} = -\zeta\vec{v} = -\zeta x'(t)$, etc ou une combinaison de toutes ces forces.

Dans la section précédente, nous avons vu l'algorithme d'Euler pour résoudre des EDO. Cette méthode a pour avantage sa simplicité de codage, son faible coût de calcul, mais a pour désavantage son manque de précision. Dans un certain nombres d'applications, telles que les moteurs physiques pour les graphismes dans les jeux vidéos, ce manque de précision est inacceptable et une meilleure méthode doit être utilisée. Dans le TP vous avez vu les méthodes de Runge-Kutta. Ces méthodes améliorent la précision de façon spectaculaire, mais ont en général un coû de calcul trop élevé.

La méthode de Verlet qu'on va voir ci-dessous est augmente combine un faible coût de calcul et une amélioration notable de la précision. Elle est en effet très répandue dans l'industrie du jeu vidéo pour intégrer les équations différentielles omniprésentes dans les moteurs physiques.

La méthode de Verlet s'écrit (en utilisant les notations de la section précédente)

$$x(t_{n+1}) = x(t_n) + \delta t v(t_n) + \frac{1}{2}\delta t^2 a(x(t_n)). \tag{4.139}$$

Considérons d'abord le terme $v(t_n)$. Ce terme est approximé ici comme

$$v(t_n) = \frac{x(t_{n+1}) - x(t_{n-1})}{2\delta t}. \tag{4.140}$$

En remplaçant cette approximation dans l'équation ci-dessus, il vient

$$x(t_{n+1}) = x(t_n) + \frac{x(t_{n+1}) - x(t_{n-1})}{2} + \frac{1}{2}\delta t^2 a(x(t_n)), \tag{4.141}$$

$$2x(t_{n+1}) = 2x(t_n) + x(t_{n+1}) - x(t_{n-1}) + \delta t^2 a(x(t_n)), \tag{4.142}$$

$$x(t_{n+1}) = 2x(t_n) - x(t_{n-1}) + \delta t^2 a(x(t_n)). \tag{4.143}$$

On voit ici que cette formule est inutilisable pour évaluer $x(t_1)$ (ce qui veut dire que $n = 0$ dans le cas ce-dessus), car elle fait intervenir $x(t_{-1})$ dans le membre

de droite. Pour résoudre ce problème il suffit d'évaluer $x(t_1)$ grâce à l'éq. 4.139 où $n = 0$

$$x(t_1) = x(t_0) + \delta t v(t_0) + \frac{1}{2}\delta t^2 a(x(t_0)),$$
$$x(t_1) = x_0 + \delta t v_0 + \frac{1}{2}\delta t^2 a(x_0),$$

où x_0 et v_0 sont les conditions initiales de notre problème. Ensuite les itérations suivantes ($n > 0$) sont calculables directement avec l'éq. 4.143. Un autre avantage considérable de ce modèle est qu'il est très simple d'y inclure une force de frottement proportionnelle à la vitesse. Sans entrer dans les détails de la dérivation du schéma on a

$$x(t_{n+1}) = (2 - \delta t\zeta)x(t_n) - (1 - \delta t\zeta)x(t_{n-1}) + \delta t^2 a(x(t_n)). \tag{4.144}$$

Chapitre 5

Transformées de Fourier

5.1 Rappel sur les nombres complexes

Dans cette section, on fait un rappel sur les nombres complexes qui seront beaucoup utilisés dans la suite.

5.1.1 Les nombres réels

L'ensemble des nombres réels, noté $\mathbb{R}$, est doté d'un certain nombre de fonctions (opérateurs) tels que l'addition, la multiplication etc qui prennent un couple de nombres réels et rendent un autre nombre réel

$$\begin{aligned} + &: \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \\ \cdot &: \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \end{aligned} \tag{5.1}$$

De la définition de l'addition de deux nombres réels il vient par exemple que

$$+(7, 2) = 9. \tag{5.2}$$

On préfère la notation

$$+(7, 2) = 7 + 2 = 9. \tag{5.3}$$

Intéressons nous plus particulièrement à la multiplication et à l'addition. Ces opérations ont les propriétés d'associativité et de commutativité. Cela veut dire que

$$(a + b) + c = a + (b + c), \quad (a \cdot b) \cdot c = a \cdot (b \cdot c),$$

et

$$a + b = b + a, \qquad a \cdot b = b \cdot a.$$

5.1.2 Les couples de nombres réels

Intéressons-nous à présent à un ensemble plus grand que $\mathbb{R}$, soit $\mathbb{R}^2 \equiv \mathbb{R} \times \mathbb{R}$. Cet ensemble est l'ensemble des des couples de nombres réels. Notons les nombres $z \in \mathbb{R}^2$ comme

$$z = (a, b) \text{ tel que } a \in \mathbb{R}, \text{ et } b \in \mathbb{R}. \tag{5.4}$$

Sur ces nombres on peut définir à nouveau l'addition, la multiplication, ...

$$\begin{aligned} + &: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2, \\ \cdot &: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2. \end{aligned} \tag{5.5}$$

On peut les écrire sous la forme de leurs équivalents des nombres réels comme

$$(a, b) + (c, d) = (a + c, b + d), \tag{5.6}$$

$$(a, b) \cdot (c, d) = (a \cdot c - b \cdot d, a \cdot d + b \cdot c). \tag{5.7}$$

On voit assez facilement que l'addition sur $\mathbb{R}^2$ a une forme très similaire à celle sur $\mathbb{R}$ du point de vue de ses propriétés telles que la commutativité ou l'associativité. Cela est moins clair pour la multiplication. Il est néanmoins assez simple de vérifier la commutativité

$$\begin{aligned} (a, b) \cdot (c, d) &= (a \cdot c - b \cdot d, a \cdot d + b \cdot c) \\ &= (c \cdot a - d \cdot b, d \cdot a + c \cdot b) = (c, d) \cdot (a, b). \end{aligned}$$

Exercice 19

Vérifier l'associativité du produit sur notre ensemble $\mathbb{R}^2$.

Regardons à présent ce qui se passe si on étudie les ensemble de nombres dans $\mathbb{R}^2$ où le deuxième nombre du couple est nul tels que $(a, 0)$. Si on additionne deux tels nombres ont obtient

$$(a, 0) + (b, 0) = (a + b, 0). \tag{5.8}$$

On constate donc que ce genre de nombre se comporte exactement comme un nombre réel normal du point de vue de l'addition. Que se passe-t-il quand on multiplie deux tels nombres

$$(a, 0) \cdot (b, 0) = (a \cdot b - 0 \cdot 0, a \cdot 0 + 0 \cdot b) = (a \cdot b, 0). \tag{5.9}$$

On voit que pour la multiplication également les ensembles de nombres dont le deuxième est nul, se comporte comme un nombre réel standard.

En fait on peut montrer que ce sous-ensemble de $\mathbb{R}^2$ se comporte exactement comme $\mathbb{R}$. Il se trouve donc que $\mathbb{R}^2$ est un ensemble plus grand que $\mathbb{R}$ et qui le contient entièrement.

5.1.3 Les nombres complexes

Afin de simplifier les notations et les calculs, on peut introduire une notation différente. Introduisons donc le *nombre imaginaire* i tel que

$$(a, b) = a + i \cdot b. \tag{5.10}$$

On va maintenant définir l'ensemble des nombres complexes $z \in \mathbb{C}$ comme tout nombre qui peut s'écrire sous la forme

$$z = a + i \cdot b. \tag{5.11}$$

Avec l'addition que nous avons définie à l'éq. 5.6, nous avons avec la nouvelle notation

$$(a,b)+(c,d)=(a+c,b+d)\Leftrightarrow(a+i\cdot b)+(c+i\cdot d)=(a+c)+i(b+d). \tag{5.12}$$

On constate que les nombres multipliés par i sépare nos couples de nombres (les empêche "de se mélanger"),

Pour la multiplication nous avons de même par la définition (équation éq. 5.7)

$$(a,b)\cdot(c,d)=(ac-bd,ad+bc)\Leftrightarrow(a+i\cdot b)\cdot(c+i\cdot d)=(ac-bd)+i(ad+bc). \tag{5.13}$$

Si maintenant nous utilisons la multiplication de manière classique avec notre nouvelle notation (on distribue le produit comme pour les réels)

$$(a+i\cdot b)\cdot(c+i\cdot d)=ac+i^2\cdot bd+i(ad+bc). \tag{5.14}$$

On constate donc que pour que cette équation soit égale à l'équation éq. 5.13 on doit avoir que $i^2=-1$. Il se trouve que c'est la définition formelle du nombre imaginaire. Dans les réels i ne peut pas exister. En revanche dans l'espace plus grand des complexes i a une existence tout à fait naturelle et raisonnable. En fait le nombre i est associé au couple $(0,1)$ comme on voit par $(0,19\cdot(0,1)=(-1,0)$.

On appelle partie réelle d'un nombre complexe z, la partie pas multipliée par i (on la note $\mathrm{Re}(z)$) et partie imaginaire celle multipliée par i (on la note $\mathrm{Im}(z)$). Pour $z=a+ib$, on a donc $\mathrm{Re}(z)=a$ et $\mathrm{Im}(z)=b$.

5.1.3.1 Interprétation géométrique

Comme on l'a vu précédemment, les nombres complexes peuvent se voir comme une "notation" de $\mathbb{R}^2$. On peut ainsi les représenter sur un plan bidimensionnel (voir la fig. 5.1).

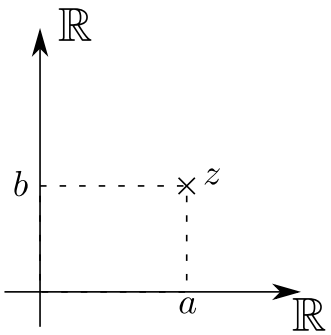

FIGURE 5.1 – Représentation du nombre complexe $z=a+ib$.

La somme de deux nombres complexes s'interprète également facilement de façon graphique. On peut le voir sur la fig. 5.2. Il s'agit en fait de simplement faire la somme des vecteurs représentant chacun des nombres complexes à sommer.

Pour la multiplication cela s'avère un peu plus difficile à interpréter. Pour cela il est plus simple de passer par une représentation via des sinus et des cosinus (en coordonnées polaires) des nombres complexes (voir la fig. 5.3.

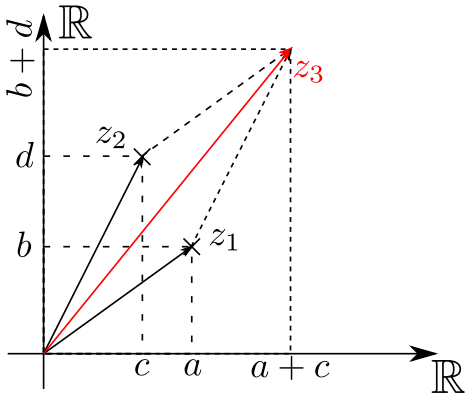

FIGURE 5.2 – Représentation de la somme de deux nombres complexes $z_1 = a+ib$ et $z_2 = c+id$. Le résultat est donné par $z_3 = a+c+i(b+d)$.

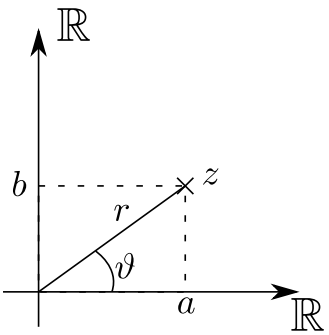

FIGURE 5.3 – Représentation du nombre complexe $z = a + ib$.

En utilisant la représentation en termes de ϑ et r, on a que $z = r(\cos\vartheta + i\sin\vartheta) = a + ib$. On a immédiatement les relations suivantes entre ces deux représentations

$$\begin{aligned} r &= \sqrt{a^2+b^2}, \\ \cos\vartheta &= \frac{a}{r}, \\ \sin\vartheta &= \frac{b}{r}. \end{aligned} \tag{5.15}$$

On dit que r est le *module* de z (aussi noté $|z|$) et que ϑ est son *argument* (aussi noté $\arg(z)$).

Si à présent on définit $z_1 = r_1(\cos\vartheta_1 + i\sin\vartheta_1)$ et $z_2 = r_2(\cos\vartheta_2 + i\sin\vartheta_2)$, on a que $z_3 = z_1 \cdot z_2$ devient

$$z_3 = r_1 r_2 \left(\cos\vartheta_1\cos\vartheta_2 - \sin\vartheta_1\sin\vartheta_2 + i\left(\cos\vartheta_1\sin\vartheta_2 + \cos\vartheta_2\sin\vartheta_1\right)\right). \tag{5.16}$$

En utilisant les relations trigonométriques suivantes

$$\begin{aligned} \cos\vartheta_1\cos\vartheta_2 - \sin\vartheta_1\sin\vartheta_2 &= \cos(\theta_1+\theta_2), \\ \cos\vartheta_1\sin\vartheta_2 + \cos\vartheta_2\sin\vartheta_1 &= \sin(\theta_1+\theta_2), \end{aligned} \tag{5.17}$$

il vient

$$z_3 = r_1 r_2 \left(\cos(\vartheta_1+\vartheta_2) + i(\sin(\vartheta_1+\vartheta_2)\right). \tag{5.18}$$

On a donc comme interprétation géométrique que le produit de deux nombres complexe donne un nombre complexe dont la longueur (module) est le produit des longueurs des nombres complexes originaux et dont l'orientation (argument) est la somme des angles des nombres complexes originaux.

Cette propriété du produit nous amène à la notation sous forme d'exponentielle des nombres complexes. L'exponentielle, possède la propriété intéressante suivante

$$e^a e^b = e^{a+b}. \tag{5.19}$$

Ou encore quand on multiplie deux nombres représentés par une exponentielle, on peut représenter le résultat par l'exponentielle de la somme de leurs arguments. Comme pour les nombre complexes en somme. Il en découle des ces considérations que

$$z = re^{i\vartheta} = r(\cos\vartheta + i\sin\vartheta). \tag{5.20}$$

On peut démontrer de façon plus rigoureuse cette relation grâce aux équations différentielles. On a vu dans le chapitre précédent que l'équation différentielle

$$f'(x) = \alpha f(x), \quad f(0) = r. \tag{5.21}$$

a pour solution $f(x) = e^{\alpha x}$ ($\alpha \in \mathbb{C}$). Si on remplace α par i, on a $f = e^{ix}$. Par ailleurs, avec $\alpha = i$, on peut également vérifier que $f(x) = r(\cos x + i\sin x)$ satisfait l'équation différentielle ci-dessus. On a donc bien que les deux formes sont égales.Remarquons que $e^{ix} = \cos(x) + i\sin(x), x \in \mathbb{R}$ est la fameuse formule d'Euler.

5.1.3.2 Quelques notations et définitions

Pour la suite de ce cours, nous allons avoir besoin d'un certain nombre de notations et de définition. En particulier, nous allons noter $\bar{z}$ le nombre complexe conjugué de z. Soit $z = a + ib$, son complexe conjugué $\bar{z}$ est donné par $\bar{z} = a - ib$. On voit que le complexe conjugué a la même partie réelle que le nombre de départ, mais une partie imaginaire opposée.

Lors de l'utilisation de la notation polaire d'un nombre complexe, nous avons que le nombre complexe conjugué est de module égal, mais d'argument opposé. En d'autres termes, si $z = re^{i\vartheta}$, alors $\bar{z} = re^{-i\vartheta}$.

On peut également écrire le module d'un nombre complexe à l'aide de la notation du complexe conjugué. Il est donné par

$$|z| = \sqrt{z\bar{z}}. \tag{5.22}$$

Finalement, on peut également exprimer les parties réelle et imaginaires d'un nombre complexe à l'aide de la notation du complexe conjugué

$$\mathrm{Re}(z) = \frac{1}{2}(z + \bar{z}), \quad \mathrm{Im}(z) = \frac{1}{2i}(z - \bar{z}). \tag{5.23}$$

Exercice 20

Démontrer ces trois relations.

Rajoutons encore la relation entre $e^{i\theta}$ et les $\cos, \sin$.

$$\begin{aligned} \cos(\theta) &= \frac{e^{i\theta} + e^{-i\theta}}{2}, \\ \sin(\theta) &= \frac{e^{i\theta} - e^{-i\theta}}{2i}. \end{aligned} \tag{5.24}$$

Exercice 21

Démontrer ces relations.

5.1.4 Espaces vectoriels

Ici nous introduisons de façon très simplifiée le concept d'espace vectoriel et certaines notions d'algèbre linéaire. Pour ce faire nous allons considérer un ensemble V muni d'une addition et d'une multiplication par un scalaire, c'est à dire par un nombre appartenant à un ensemble E. Dans notre cas E sera $\mathbb{R}$ ou $\mathbb{C}$ (l'ensemble des nombres complexes) principalement.

Définition 20

On appelle espace vectoriel sur E, un ensemble V, dont les éléments appelés vecteurs et notés v, sont sont munis des opérations $+$ (l'addition) et $\cdot$ (la multiplication par un scalaire) qui ont les propriétés suivantes

—

1. L'addition est associative et commutative. Soient $u, v, w \in V$, alors
$$u + v = v + u, \quad \text{et} \quad (u + v) + w = u + (v + w). \tag{5.25}$$
2. L'addition admet un élément neutre additif, noté 0_V, tel que
$$0_V + v = v. \tag{5.26}$$
3. Tout v admet un opposé, noté $-v$ tel que
$$v + (-v) = 0_V. \tag{5.27}$$

—

1. La multiplication par un scalaire est distributive à gauche sur l'addition (et à droite sur E). Pour $u, v \in V$ et $\alpha \in E$, on a
$$\alpha \cdot (u + v) = \alpha \cdot u + \alpha \cdot v. \tag{5.28}$$
2. La multiplication est associative par rapport à la multiplication de E. Soient $\alpha, \beta \in E$
$$(\alpha \cdot \beta) \cdot v = \alpha \cdot (\beta \cdot v). \tag{5.29}$$
3. La multiplication par un scalaire admet un élément neutre, noté 1, pour la multiplication à gauche
$$1 \cdot v = v. \tag{5.30}$$

Illustration 21 *(Espaces vectoriels)*

1. L'espace nul, $v = 0$.
2. $V = \mathbb{R}$ ou $V = \mathbb{C}$ avec $E = \mathbb{R}$.
3. Espaces de $n - uplets$. Soit V un espace vectoriel sur E.L'espace des $n-$uplets. Pour t$n > 0$, l'ensemble des $n-$uplets d'éléments de V, $v = (v_1, v_2, ..., v_n)$, $\{v_i \in E\}_1^n$, est noté V^n. Sur cet espace l'addition se définit $(u, v \in V^n)$
$$u + v = (u_1 + v_1, u_2 + v_2, ..., u_v + v_n), \tag{5.31}$$
et la multiplication par un scalaire $\alpha \in E$
$$\alpha v = (\alpha v_1, \alpha v_2, ..., \alpha v_n). \tag{5.32}$$
On a donc que l'élément neutre de l'addition est le vecteur $0_{E^n} = \underbrace{(0, 0, ..., 0)}_{n}$. L'élément opposé de v est $-v = (-v_1, -v_2, ..., -v_n)$.
Si $V = \mathbb{R}$, alors on a l'espace Euclidien. Vous avez l'habitude de l'utiliser en 2D ou 3D quand vous considérez des vecteurs. Dans ce cas $\mathbb{R}^2$ ou $\mathbb{R}^3$ avec l'addition classique et la multiplication par un réel forme un espace vectoriel.

4. Dans ce qui suit dans ce cours, nous allons utiliser encore un autre espace vectoriel un peu moins intuitif que ceux que nous avons vus jusqu'ici. Il s'agit de l'espace des fonctions, ou espace fonctionnel. Nous définissons les applications de W dans V comme un espace vectoriel dans E avec l'addition et la multiplication par un scalaire définis comme suit. Soient $f : W \to V$ et $g : W \to V$, avec $\alpha \in E$, alors

$$\begin{aligned} (f+g)(x) &= f(x) + g(x), \quad \forall x \in W, \\ (\alpha \cdot f)(x) &= \alpha \cdot f(x), \quad \forall x \in W. \end{aligned} \tag{5.33}$$

5. Espace des applications linéaires. Soit f une fonction de $f : W \to V$, avec W, V des espaces vectoriels sur E, alors une application est dite linéaire si

$$\begin{aligned} f(x+y) &= f(x) + f(y), \quad \forall x, y \in W, \\ f(\alpha \cdot x) &= \alpha \cdot f(x), \quad \forall \alpha \in E, \text{ et } x \in W. \end{aligned} \tag{5.34}$$

5.1.5 Base

Nous avons introduit la notion très générale d'espace vectoriel et nous avons présenté quelques exemples. Reprenons l'exemple de l'espace Euclidien, soit l'espace des vecteurs comme vous en avez l'habitude. Limitons nous au cas où les vecteur sont bidimensionnels, soit $v = (v_1, v_2)$ avec $v_1, v_2 \in \mathbb{R}$. D'habitude ces vecteurs sont représentés dans le système de coordonnées cartésien où on a deux vecteurs (de base) définis comme $e_1 = (1, 0)$ et $e_2 = (0, 1)$ qui sont implicites. Par exemple, si $u = (4, 5)$ cela signifie implicitement que

$$u = 4 \cdot e_1 + 5 \cdot e_2. \tag{5.35}$$

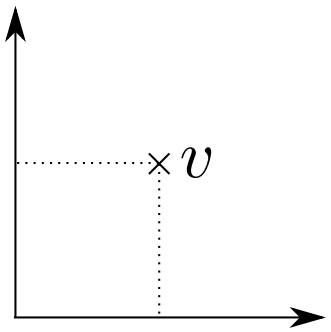

FIGURE 5.4 – Le vecteur v dans la représentation cartésienne.

De façon générale tout vecteur $v = (v_1, v_2)$ est représenté implicitement par (voir la fig. 5.4)

$$v = v_1 \cdot e_1 + v_2 \cdot e_2. \tag{5.36}$$

On dit que e_1 et e_2 forme une *base* de l'espace $\mathbb{R}^2$. En d'autres termes n'importe quel vecteur $v \in \mathbb{R}^2$ peut être exprimé comme une combinaison linéaire de e_1 et e_2.

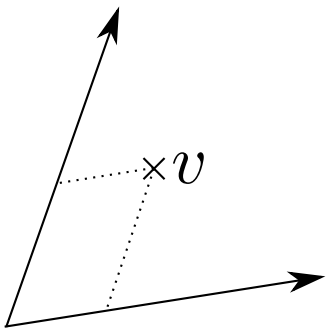

FIGURE 5.5 – Le vecteur v dans une représentation non cartésienne.

Néanmoins, le choix de la base e_1 et e_2 est totalement arbitraire. N'importe quelle autre paire de vecteurs (qui n'on pas la même direction) peut être utilisée pour représenter un vecteur quelconque dans le plan (voir la fig. 5.5).

Cette écriture en fonction de vecteurs de base, permet de faire facilement les additions de vecteurs

$$w = u + v = u_1 \cdot e_1 + u_2 \cdot e_2 + v_1 \cdot e_1 + v_2 \cdot e_2 = (u_1 + v_1) \cdot e_1 + (u_2 + v_2) \cdot e_2. \tag{5.37}$$

Illustration 22 *(Exemples de bases d'espaces vectoriels)*

1. Pour l'espace des fonctions polynomiales $f(x) = \sum_{i=0}^{N} a_i x^i$ les fonction $e_i = x^i$ forment une base.
2. Pour l'espace vectoriel des fonctions périodiques les fonctions sin et cos forment une base (voir plus de détails dans ce qui suit).

Plus formellement nous allons introduire un certain nombre de concepts mathématiques pour définir une base. Considérons toujours V un espace vectoriel sur E.

Définition 21 *(Famille libre)*

Soient $\{\alpha_i\}_{i=1}^n \in E$. On dit qu'un ensemble de vecteurs $\{v_i\}_{i=1}^n \in V$ est une famille libre si

$$\sum_{i=1}^{n} \alpha_i v_i = 0 \Rightarrow \alpha_i = 0, \ \forall i. \tag{5.38}$$

Illustration 23 *(Famille libre)*

1. $\{e_1\}$ est une famille libre de $\mathbb{R}^2$.
2. $\{e_1, e_2\}$ est une famille libre de $\mathbb{R}^2$.

3. $\{e_1, e_2, v\}$, avec $v = (1, 1)$ n'est pas une famille libre de $\mathbb{R}^2$. En effet,

$$1 \cdot e_1 + 1 \cdot e_2 - 1 \cdot v = (0, 0). \tag{5.39}$$

4. $\{\sin(x), \cos(x)\}$ est une famille libre. On ne peut pas écrire $\sin(x) = \alpha \cos(x) + \beta$. Il n'y a pas de relation linéaire qui relie les deux. La relation est non-linéaire $\sin(x) = \sqrt{1 - \cos^2(x)}$.

Définition 22 *(Famille génératrice)*

On dit qu'un ensemble de vecteurs $\{e_i\}_{i=1}^n \in V$ est une famille génératrice si

$$\forall\ v \in V, \quad \exists \{\alpha_i\}_{i=1}^n \in E, \quad \text{t.q.} \quad v = \sum_{i=1}^{n} \alpha_i \cdot e_i. \tag{5.40}$$

En d'autres termes, tout $v \in V$ peut s'exprimer comme une combinaison linéaire des vecteur e_i.

Illustration 24 *(Familles génératrices)*

1. $\{e_1\}$ n'est pas une famille génératrice de $\mathbb{R}^2$. On ne peut pas représenter les vecteurs de la forme $v = (0, v_2)$, $v_2 \neq 0$.
2. $\{e_1, e_2\}$ est une famille génératrice de $\mathbb{R}^2$.
3. $\{e_1, e_2, v\}$, avec $v = (1, 1)$ est une famille génératrice de $\mathbb{R}^2$.

Définition 23 *(Base)*

Un ensemble de vecteurs $B = \{e_i\}_{i=1}^n$ forme une base si c'est une famille génératrice et une famille libre. En d'autres termes cela signifie qu'un vecteur $v \in V$ peut se représenter comme une combinaison linéaire de $\{e_i\}_{i=1}^n$ et que cette représentation est unique

$$\forall v \in V, \quad !\exists \{\alpha_i\}_{i=1}^n \in E, \quad t.q. \quad v = \sum_{i=1}^{n} \alpha_i v_i. \tag{5.41}$$

Les α_i sont appelé les coordonnées de v dans la base B.

Illustration 25 *(Base de $\mathbb{R}^2$)*

1. $\{e_1, e_2\}$ est une base de $\mathbb{R}^2$.

2. $\{e_1, e_2, e_3\}$, avec $e_3 = (1, 1)$, n'est pas une base de $\mathbb{R}^2$, car ce n'est pas une famille libre. On a par exemple que l'élément $v = (0, 0)$ peut se représenter avec les coordonnées $\alpha = (0, 0, 0)$ et également les coordonnées $\beta = (1, 1, -1)$.

5.2 Introduction générale sur les séries de Fourier

Dans cette sous section, nous allons voir de façon très générale les concepts de la représentation de série de Fourier de fonctions.

5.2.1 Considérations historiques

Historiquement, les séries de Fourier sont apparues lorsque les mathématiciens/physiciens du 18-19ème siècles ont essayé de résoudre des équations différentielles particulières. En particulier, il y avait l'équation de la propagation d'ondes

$$\frac{\partial^2 \rho}{\partial t^2} = \alpha^2 \left(\frac{\partial^2 \rho}{\partial x^2} + \frac{\partial^2 \rho}{\partial y^2} + \frac{\partial^2 \rho}{\partial z^2} \right), \tag{5.42}$$

où ρ est l'amplitude de l'onde et α la vitesse de propagation. On a également l'équation de la chaleur

$$\frac{\partial T}{\partial t} = \kappa \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right), \tag{5.43}$$

où T est la température et κ la diffusivité thermique.

Ces équations ont une structure particulière. En effet, d'une part elles sont linéaires. Soient ρ_1 et ρ_2 deux solutions de l'équation éq. 5.42, on a que la somme $\rho_1 + \rho_2$ est également solution de éq. 5.42. Cette structure d'équation différentielle impose des contraintes assez fortes sur la forme des solutions.

Par ailleurs, le fait que les dérivées à différents ordres apparaissent dans la même équation, cela impose que les fonctions et leurs dérivées à différents ordres soient reliées entre elles. Les fonctions qu'on connaît qui ont ces propriétés sont l'exponentielle et les fonctions sinus ou cosinus. Dans le cas de propagation d'ondes, on voit qu'on a uniquement des deuxièmes dérivées, et on en déduit que les fonctions importantes seront des sinus et des cosinus.

On constate que le choix du sinus ou du cosinus pour représenter ces solutions ne tombe pas du ciel. Il est dicté par les propriétés des équations que nous tentons de résoudre. En fait, nous mettons à notre disposition des outils mathématiques appropriés pour résoudre des problèmes physiques existant et qui ont des contraintes particulières.

5.2.2 Décomposition de signaux périodiques

Nous allons considérer une fonction $f(t)$ qui est une fonction périodique, de période T, de pulsation $\omega = 2\pi/T$ et de fréquence $\nu = 1/T$. La périodicité

signifie que

$$f(t+T) = f(t), \quad \forall t. \tag{5.44}$$

Nous cherchons à décomposer f en un ensemble potentiellement infini de fonctions périodiques. Notons cet ensemble de fonctions $\{g_j\}_{j=0}^{\infty}$, où g_j est une fonction périodique. En fait on cherche une décomposition où pour un ensemble unique de $\{\alpha_j\}_{j=0}^{\infty}$

$$f(t) = \sum_{j=0}^{\infty} \alpha_j g_j(t). \tag{5.45}$$

Cette décomposition nous fait penser furieusement à une décomposition dans une base particulière, où les g_j sont les vecteurs de la base et les α_j sont les coordonnées de f dans la base des g_j.

La fonction de départ f ayant une période T, on a obligatoirement que les fonctions g_j ont une période qui doit être une fraction entière de la période, T/j. Ces fonctions $g_j(t)$ peuvent en général avoir une forme quelconque, avec l'unique contrainte qu'elles sont périodiques avec période T/j. Ça pourrait être un signal carré, triangulaire, etc. Dans les cas qui nous intéresse, on a un choix naturel qui s'impose comme fonctions périodiques: les sinus et cosinus.

Pour commencer, imaginons que nous voulions décomposer (approximer) f en une somme de $g_j \sim A_j \sin(j\omega t + \phi_j)$. On peut jouer sur deux degrés de libertés des sinus dont la période est imposée, soit l'amplitude A_j et la phase ϕ_j. On va donc écrire $f(t)$ comme

$$f(t) = \sum_{j=0}^{\infty} A_j \sin(j\omega t + \phi_j). \tag{5.46}$$

Cette forme n'est pas pratique du tout comme décomposition, en particulier à cause de la phase ϕ_j. On utilise alors la relation trigonométrique (déjà utilisée pour interpréter le produit de deux nombres complexes)

$$\sin(\theta + \phi) = \sin(\theta)\cos(\phi) + \cos(\theta)\sin(\phi). \tag{5.47}$$

Il vient

$$f(t) = \sum_{j=0}^{\infty} A_j \left(\sin(j\omega t)\cos(\phi_j) + \cos(j\omega t)\sin(\phi_j)\right). \tag{5.48}$$

En renommant

$$\begin{aligned} a_j &\equiv A_j \sin(\phi_j), \\ b_j &\equiv A_j \cos(\phi_j), \end{aligned} \tag{5.49}$$

on obtient

$$f(t) = \sum_{j=0}^{\infty} \left(a_j \cos(j\omega t) + b_j \sin(j\omega t)\right). \tag{5.50}$$

On a ainsi transformé une équation où on devait déterminer une amplitude et une phase, ce qui est plutôt compliqué, en une autre équation où on doit déterminer uniquement deux amplitude. Par ailleurs, comme cos et sin sont indépendants, on peut calculer les a_j et b_j de façon également indépendantes.

Nous voulons à présent calculer a_j et b_j pour avoir les coordonnées de f dans la base des sin et des cos. Pour ce faire, nous allons tenter de trouver les amplitudes a_j, b_j tels que les $a_j \cos(j\omega t)$ et $b_j \sin(j\omega t)$ approximent au mieux la fonction f.

Nous allons considérer les fonctions d'erreur suivantes

$$E_j^s = \int_0^T (f(t) - b_j \sin(j\omega t))^2 \mathrm{d}t, \quad E_j^c = \int_0^T (f(t) - a_j \cos(j\omega t))^2 \mathrm{d}t. \tag{5.51}$$

Puis on va déterminer a_j, b_j tels que E_j^s et E_j^c sont minimales. Pour ce faire on va utiliser les dérivées et déterminer nos coefficients en résolvant les équations

$$\frac{\mathrm{d}E_j^c}{\mathrm{d}a_j} = 0. \tag{5.52}$$

$$\frac{\mathrm{d}E_j^s}{\mathrm{d}b_j} = 0, \tag{5.53}$$

Pour l'éq. 5.52, on a

$$\begin{aligned}
\frac{\mathrm{d}E_j^c}{\mathrm{d}a_j} &= \frac{\mathrm{d}\int_0^T (f(t) - a_j \cos(j\omega t))^2 \mathrm{d}t}{\mathrm{d}a_j}, \\
&= \underbrace{\frac{\mathrm{d}(\int_0^T f^2(t)\mathrm{d}t)}{\mathrm{d}a_j}}_{=0} + \frac{\mathrm{d}(a_j^2 \int_0^T (\cos^2(j\omega t)\mathrm{d}t))}{\mathrm{d}a_j} - \frac{\mathrm{d}(2a_j \int_0^T (f(t)\cos(j\omega t)\mathrm{d}t))}{\mathrm{d}a_j}, \\
&= 2a_j \int_0^T \cos^2(j\omega t)\mathrm{d}t - 2\int_0^T f(t)\cos(j\omega t)\mathrm{d}t, \\
&= 2a_j \frac{T}{2} - 2\int_0^T \cos(j\omega t) f(t)\mathrm{d}t.
\end{aligned}$$

Finalement on obtient

$$a_j = \frac{2}{T}\int_0^T \cos(j\omega t) f(t)\mathrm{d}t. \tag{5.54}$$

Pour a_j on a de façon similaire

$$b_j = \frac{2}{T}\int_0^T \sin(j\omega t) f(t)\mathrm{d}t. \tag{5.55}$$

En particulier si $j = 0$, on a

$$b_0 = 0, \quad a_0 = \frac{2}{T}\int_0^T f(t)\mathrm{d}t. \tag{5.56}$$

On constate que $b_0/2$ correspond à la valeur moyenne de $f(t)$ dans $[0, T]$. Cela permet d'approximer des fonctions dont la valeur moyenne n'est pas nulle (les sinus et cosinus ont toujours des moyennes nulles).

Les coefficients a_j, b_j peuvent être calculés directement à partir de $f(t)$, comme nous venons de le voir. Nous pouvons obtenir le même résultat, en utilisant les

relations suivantes (exercice)

$$\begin{aligned}\int_0^T \sin(k\omega t)\sin(j\omega t)\mathrm{d}t &= \delta_{jk}\frac{T}{2},\\ \int_0^T \cos(k\omega t)\cos(j\omega t)\mathrm{d}t &= \delta_{jk}\frac{T}{2},\\ \int_0^T \sin(k\omega t)\cos(j\omega t)\mathrm{d}t &= 0,\end{aligned} \tag{5.57}$$

qui s'obtiennent en utilisant les relations trigonométriques suivantes

$$\begin{aligned}\sin\theta\sin\phi &= \frac{1}{2}\left(\cos(\theta-\phi)-\cos(\theta+\phi)\right),\\ \cos\theta\cos\phi &= \frac{1}{2}\left(\cos(\theta-\phi)+\cos(\theta+\phi)\right),\\ \sin\theta\cos\phi &= \frac{1}{2}\left(\sin(\theta+\phi)+\sin(\theta-\phi)\right).\end{aligned} \tag{5.58}$$

Cela est dû à la propriété d'orthogonalité des fonctions sinus/cosinus. En multipliant l'éq. 5.50 par $\frac{2}{T}\sin(k\omega t)$ et en intégrant entre 0 et T, on obtient

$$\frac{2}{T}\int_0^T f(t)\sin(k\omega t)\mathrm{d}t = \frac{2}{T}\sum_{j=0}^{\infty}\left(b_j\underbrace{\int_0^T \cos(j\omega t)\sin(k\omega t)\mathrm{d}t}_{=0} + a_j\underbrace{\int_0^T \sin(j\omega t)\sin(k\omega t)\mathrm{d}t}_{=\frac{T}{2}\delta_{jk}}\right),$$

$$\frac{2}{T}\int_0^T f(t)\sin(k\omega t)\mathrm{d}t = \sum_{j=0}^{\infty} a_j\delta_{jk} = a_k,$$

où δ_{jk} est le "delta de Kronecker", dont la définition est

$$\delta_{jk} = \begin{cases} 1, & \text{si } j=k\\ 0, & \text{sinon.}\end{cases} \tag{5.59}$$

En multipliant l'éq. 5.50 par $\frac{2}{T}\cos(k\omega t)$ et en intégrant entre 0 et T, on obtient

$$\frac{2}{T}\int_0^T f(t)\cos(k\omega t)\mathrm{d}t = \frac{2}{T}\sum_{j=0}^{\infty}\left(a_j\underbrace{\int_0^T \cos(j\omega t)\sin(k\omega t)\mathrm{d}t}_{=0} + b_j\underbrace{\int_0^T \cos(j\omega t)\cos(k\omega t)\mathrm{d}t}_{=\frac{T}{2}\delta_{jk}}\right),$$

$$\frac{2}{T}\int_0^T f(t)\cos(k\omega t)\mathrm{d}t = \sum_{j=0}^{\infty} b_j\delta_{jk} = b_k.$$

5.2.3 Les séries de Fourier en notations complexes

Comme on le voit dans l'éq. 5.50, on décompose $f(t)$ en une somme contenant des sinus et des cosinus. Cette écriture nous fait penser qu'il pourrait être possible

de réécrire cette somme de façon plus concise à l'aide des nombres complexes ($e^{i\theta} = \cos\theta + i \cdot \sin\theta$). Effectivement cette réécriture est possible. Pour ce faire il faut définir de nouveaux coefficients c_n,

$$c_n = \begin{cases} \frac{a_n + ib_n}{2}, & \text{si } n < 0 \\ \frac{a_0}{2}, & \text{si } n = 0 \\ \frac{a_n - ib_n}{2}, & \text{si } n > 0 \end{cases} \tag{5.60}$$

Avec cette notation, on peut réécrire l'éq. 5.50 (exercice) comme

$$f(t) = \sum_{j=-\infty}^{\infty} c_j e^{ij\omega t}. \tag{5.61}$$

En multipliant cette relation par $\frac{1}{T}e^{-ik\omega t}$ et en intégrant entre $-\frac{T}{2}$ et $\frac{T}{2}$, on obtient

$$\frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)e^{-ik\omega t}\mathrm{d}t = \frac{1}{T}\sum_{j=-\infty}^{\infty} c_j \int_{-\frac{T}{2}}^{\frac{T}{2}} e^{ij\omega t}e^{-ik\omega t}\mathrm{d}t. \tag{5.62}$$

Pour évaluer le membre de droite de cette équation nous transformons les exponentielles en sinus/cosinus. L'intégrale du membre de droite devient

$$\begin{aligned} \int_{-\frac{T}{2}}^{\frac{T}{2}} e^{ij\omega t}e^{-ik\omega t}\mathrm{d}t &= \int_{-\frac{T}{2}}^{\frac{T}{2}} \left(\cos(j\omega t) + i\sin(j\omega t)\right)\left(\cos(-k\omega t) + i\sin(-k\omega t)\right)\mathrm{d}t, \\ &= \int_{-\frac{T}{2}}^{\frac{T}{2}} \left(\cos(j\omega t)\cos(k\omega t) + \sin(j\omega t)\sin(k\omega t)\right. \\ &\qquad \left. -i(\cos(j\omega t)\sin(k\omega t) + \cos(k\omega t)\sin(j\omega t))\right)\mathrm{d}t, \\ &= T\delta_{jk}. \end{aligned}$$

En remplaçant cette relation dans l'équation ci-dessus [1], on a

$$\frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)e^{-ik\omega t}\mathrm{d}t = \sum_{j=-\infty}^{\infty} c_j\delta_{jk} = c_k. \tag{5.63}$$

Cette relation nous dit comment évaluer les coefficients c_k de la série de Fourier de $f(t)$.

On notera que pour une fonction périodique, on obtient des coefficients de la série de Fourier qui sont discrets.

5.3 La série de Fourier pour une fonction quelconque: la transformée de Fourier

Il est possible d'écrire de telles séries pour des fonctions non-périodiques. Pour ce faire, il faut prendre la limite $T \to \infty$. Pour ce faire on va écrire

$$f(t) = \sum_{j=-\infty}^{\infty} c_j e^{ij\omega t}, \tag{5.64}$$

1. Cette relation est l'équivalent des relations d'orthogonalité entre sinus et cosinus que nous avons calculées tout à l'heure.

où on remplace le coefficient c_j par l'éq. 5.63. On obtient

$$f(t) = \sum_{j=-\infty}^{\infty} \left(\frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} f(t) e^{-ij\omega t} \mathrm{d}t \right) e^{ij\omega t}. \tag{5.65}$$

En utilisant la relation

$$\frac{1}{T} = \frac{\omega}{2\pi} = \frac{\omega(j-j+1)}{2\pi} = \frac{\omega(j+1)}{2\pi} - \frac{\omega j}{2\pi}, \tag{5.66}$$

ainsi que la notation $\omega_j = j\omega$, on peut réécrire cette équation

$$\begin{aligned} f(t) &= \sum_{j=-\infty}^{\infty} \frac{1}{2\pi} (\omega_{j+1} - \omega_j) \underbrace{\left(\int_{-\frac{\pi}{\Delta\omega_j}}^{\frac{\pi}{\Delta\omega_j}} f(t) e^{-i\omega_j t} \mathrm{d}t \right)}_{\equiv \hat{f}(\omega_j)} e^{i\omega_j t}, \\ &= \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} (\Delta\omega_j) \hat{f}(\omega_j) e^{i\omega_j t}. \end{aligned}$$

Maintenant pour passer dans le cas où la fonction n'est pas périodique (la période est infinie), nous devons prendre la limite $\Delta\omega_j \to 0$ dans l'équation précédente, et on voit apparaître une somme de Riemann

$$\begin{aligned} f(t) &= \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \lim_{\Delta\omega_j \to 0} \Delta\omega_j \hat{f}(\omega_j) e^{i\omega_j t}, \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{i\omega t} \mathrm{d}\omega. \end{aligned}$$

A présent, nous avons deux opérateurs que nous allons nommer. Nous avons la transformée de Fourier

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} \mathrm{d}t, \tag{5.67}$$

et la transformée de Fourier inverse

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{i\omega t} \mathrm{d}\omega. \tag{5.68}$$

On a immédiatement qu'appliquer la transformée de Fourier et la transformée de Fourier inverse sur une fonction $f(t)$, nous donne la fonction originale $f(t)$.

La fonction $f(t)$ doit satisfaire un certain nombre de contraintes pour pouvoir calculer sa transformée de Fourier:

1. Elle doit être de carré intégrable

$$\int_{-\infty}^{\infty} |f(t)|^2 \mathrm{d}t < \infty \tag{5.69}$$

2. Elle doit avoir un nombre fini d'extrema (ne doit pas varier trop vite).
3. Elle doit avoir un nombre fini de discontinuités.

Exercice 22

Calculer les transformées de Fourier des fonctions suivantes

1. Le pulse symétrique

$$f(t) = \begin{cases} 1, & \text{si } -t_c < t < t_c \\ 0, & \text{sinon.} \end{cases} \tag{5.70}$$

2. Le pulse asymétrique

$$f(t) = \begin{cases} 1, & \text{si } 0 < t < 2t_c \\ 0, & \text{sinon.} \end{cases} \tag{5.71}$$

3. L'exponentielle décroissante

$$f(t) = \begin{cases} e^{-at}, & \text{si } t > 0 \\ 0, & \text{sinon.} \end{cases} \tag{5.72}$$

Exercice 23

Calculer les transformées de Fourier inverse de la fonction suivante

1. Le pulse symétrique

$$f(\omega) = \begin{cases} 1, & \text{si } -\omega_c < \omega < \omega_c \\ 0, & \text{sinon.} \end{cases} \tag{5.73}$$

5.4 Propriétés des transformées de Fourier

La transformée de Fourier possède plusieurs propriétés intéressantes.

Propriété 3

1. Linéarité. Soit une fonction $h(t) = af(t) + bg(t)$, alors sa transformée de Fourier est donnée par

$$\hat{h}(\omega) = a\hat{f}(\omega) + b\hat{g}(\omega). \tag{5.74}$$

2. Translation temporelle. Soit une fonction $g(t) = f(t + t_0)$, alors sa transformée de Fourier est donnée par

$$\hat{g}(\omega) = \hat{f}(\omega)e^{i\omega t_0}. \tag{5.75}$$

3. Modulation en fréquence. Soit $\omega_0 \in \mathbb{R}$ et une fonction $g(t) = e^{-i\omega_0 t}f(t)$, alors sa transformée de Fourier est donnée par

$$\hat{g}(\omega) = \hat{f}(\omega + \omega_0). \tag{5.76}$$

4. Contraction temporelle. Soit $a \in \mathbb{R}^*$ et $g(t) = f(at)$ alors sa transformée de Fourier est donnée par

$$\hat{g}(\omega) = \frac{1}{|a|}\hat{f}(\omega/a). \tag{5.77}$$

   En particulier, on a la propriété d'inversion du temps quand $a = -1$, on a $h(t) = f(-t) \Rightarrow \hat{h}(\omega) = \hat{f}(-\omega)$.

5. Spectres de fonctions paires/impaires. Soit $f(t)$ une fonction paire (impaire), alors $\hat{f}(\omega)$ sera une fonction paire (impaire).

5.5 La transformée de Fourier à temps discret (TFTD)

Nous allons maintenant plus considérer une fonction continue, mais une série de valeurs discrètes. Notons $f[n]$ une série de nombres, avec $n \in \mathbb{N}$. Nous voulons définir l'équivalent de la transformée de Fourier de l'éq. 5.67 pour ce genre de séries de points. Une façon naturelle de définir l'équivalent à temps discret de cette équation est

$$\hat{f}(\omega) = \sum_{n=-\infty}^{\infty} f[n]e^{-i\omega n}. \tag{5.78}$$

Pour les fonctions à "temps continu" et non périodiques, nous savons que la transformée de Fourier est continue et en général non périodique. Pour le cas de la transformée de Fourier à temps discret la transformée de Fourier sera périodique, soit

$$\hat{f}(\omega + 2\pi) = \hat{f}(\omega). \tag{5.79}$$

Nous démontrons cette relation par la définition de la TFTD

$$\hat{f}(\omega + 2\pi) = \sum_{n=-\infty}^{\infty} f[n]e^{-i(\omega+2\pi)n} = \underbrace{e^{-i2\pi}}_{=1} \sum_{n=-\infty}^{\infty} f[n]e^{-i\omega n} = \hat{f}(\omega). \tag{5.80}$$

D'une certaine façon nous voyons que nous avons une similarité entre la transformée de Fourier à temps discret et les séries de Fourier. Cette similarité va devenir plus claire dans ce qui suit.

Pour définir la transformée de Fourier en temps discret inverse, nous nous inspirons de la version en temps continu (voir l'équation éq. 5.68) et on a

$$f[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{f}(\omega)e^{i\omega n}\mathrm{d}\omega. \tag{5.81}$$

Pour prouver cette relation, il suffit de remplacer l'équation éq. 5.78 dans cette relation, et il vient

$$f[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(\sum_{m=-\infty}^{\infty} f[m]e^{-i\omega m}\right) e^{i\omega n}\mathrm{d}\omega. \tag{5.82}$$

En supposant que la somme converge, nous pouvons intervertir la somme et l'intégrale et on a

$$\begin{aligned} f[n] &= \frac{1}{2\pi}\left(\sum_{m=-\infty}^{\infty} f[m]\int_{-\pi}^{\pi} e^{-i\omega(m-n)}\mathrm{d}\omega\right), \\ &= \frac{1}{2\pi}\left(\sum_{m=-\infty}^{\infty} f[m]\delta_{mn}2\pi\right), \\ &= f[n]. \end{aligned}$$

Exercice 24

Calculer les transformées de Fourier (inverses quand c'est approprié) en temps discret des fonctions suivantes

1. Le pulse symétrique

$$\hat{f}(\omega) = \begin{cases} 1, & \text{si } -\omega_c < \omega < \omega_c \\ 0, & \text{sinon.} \end{cases} \tag{5.83}$$

2. Le pulse discret

$$f[n] = \begin{cases} 1, & \text{si } n = 0 \\ 0, & \text{sinon.} \end{cases} \tag{5.84}$$

Il est intéressant de noter qu'on peut représenter une suite discrète et infinie de points par une fonction continue et périodique.

5.6 La transformée de Fourier discrète

5.6.1 Motivation

Pourquoi avons-nous besoin d'encore une transformée de Fourier? Nous avons déjà vu la transformée de Fourier de fonctions périodiques, de fonctions non-périodiques, ainsi que de fonctions à temps discret. Néanmoins, même dans le cas de la transformée de Fourier à temps discret, la transformée de Fourier est une fonction continue. Cela n'est évidemment pas pratique ni même utilisable dans un ordinateur. C'est pourquoi il est nécessaire de définir une transformée de Fourier discrète qui aura les propriétés suivantes

1. Elle transformera un signal discret de longueur finie.
2. La transformée de Fourier sera discrète et de longueur finie.

5.6.2 Applications

Avant de voir en détail comment on calcule la transformée de Fourier discrète, on peut discuter quelle sont ses applications. La TFD est utilisée tout le temps en traitement du signal. En gros c'est une approximation de la transformée de Fourier à temps discret. A chaque fois qu'on désire connaître le comportement d'une fonction dans l'espace spectral, on utilisera la TFD. Un exemple typique

est l'application pour téléphones portables Shazam que vous connaissez sans doute. Le but de cette application est l'identification de chansons. Elle fonctionne de la façon suivante. Dans un premier temps elle enregistre un signal sonore. Puis avec ce signal sonore elle crée un spectrogramme (une sorte d'emprunte digitale de la chanson) qui est obtenu à l'aide de TFD. Finalement le spectrogramme est comparé avec une base de donnée de spectrogrammes et la chanson peut ainsi être identifiée. Une autre application est le filtrage de signaux. Comme vous l'avez vu (ou verrez) dans les travaux pratiques, la TFD rend très simple le filtrage de fréquences (ou de bande de fréquences). En effet, il suffit d'ôter de la TFD d'un signal les amplitudes voulues et d'effectuer la transformée de Fourier discrète inverse (TFDI) du signal filtré. Ce genre d'applications est très utilisé dans le domaine de la compression de données (jpg, mp3, ...).

5.6.3 La transformée de Fourier discrète à proprement parler

Soit $f[n]$ un séquence de N points, $n = 0..N-1$. Pour se ramener au cas de la transformée de Fourier à temps discret, on peut aussi se dire qu'on a une séquence infinie de points, mais où $f[n] = 0$, pour $n \geq N$. On dit qu'on a N échantillons de f.

Avec cette définition il est simple de calculer la transformée de Fourier à temps discret

$$\hat{f}(\omega) = \sum_{n=0}^{N-1} f[n] e^{-i\omega n}. \tag{5.85}$$

On note que la somme à présent ne se fait plus dans l'intervalle $(-\infty, \infty)$, mais uniquement entre $[0, N-1]$, car le signal est de longueur finie.

On représente donc un signal de longueur finie $f[n]$ $(n = 0, .., N-1)$ par une fonction continue de la pulsation, $\hat{f}(\omega)$. Les deux représentations sont équivalentes. On en déduit que l'information contenue dans un nombre fini de points, est la même que dans une fonction continue (et donc contenant une infinité de points). Une partie de l'information contenue dans la fonction continue doit être redondante...

L'idée à présent va être d'enlever toute l'information redondante de $\hat{f}(\omega)$ en échantillonnant $\hat{f}$ et en gardant uniquement N échantillons de $\hat{f}$. La fréquence d'échantillonnage sera de $2\pi/N$ et le domaine d'échantillonnage sera $[-\pi, \pi)$.

Nous pouvons à présent définir mathématiquement cet échantillonnage de $\hat{f}(\omega)$ comme étant une suite de points, notée $\{\hat{f}(\omega_k)\}_{k=0}^{N-1}$, où $\omega_k = 2\pi k/N$. Cette suite sera notée $\hat{f}[k]$ et appelée la *transformée de Fourier discrète* de $f[n]$.

On a donc que la transformée de Fourier discrète de $f[n]$ est donnée par

$$\hat{f}[k] = \sum_{n=0}^{N-1} f[n] e^{-i\omega_k n} = \sum_{n=0}^{N-1} f[n] e^{-\frac{2\pi i n k}{N}}. \tag{5.86}$$

En s'inspirant de définition de la transformée de Fourier inverse à temps discret de $\hat{f}(\omega)$ (voir l'équation éq. 5.81), on a que la transformée de Fourier discrète

inverse est donnée par

$$f[n] = \frac{1}{N}\sum_{k=0}^{N-1} \hat{f}[k] e^{i\omega_k n} = \frac{1}{N}\sum_{k=0}^{N-1} \hat{f}[k] e^{\frac{2\pi i k n}{N}}. \tag{5.87}$$

Montrons à présent que la transformée inverse discrète de la transformée de Fourier discrète donne bien la suite de départ

$$\begin{aligned}
f[n] &= \frac{1}{N}\sum_{k=0}^{N-1} \hat{f}[k] e^{\frac{2\pi i k n}{N}}, \\
&= \frac{1}{N}\sum_{k=0}^{N-1}\sum_{m=0}^{N-1} f[m] e^{-\frac{2\pi i k m}{N}} e^{\frac{2\pi i k n}{N}}, \\
&= \frac{1}{N}\sum_{k=0}^{N-1}\sum_{m=0}^{N-1} f[m] e^{\frac{2\pi i k (n-m)}{N}}, \\
&= \frac{1}{N}\sum_{m=0}^{N-1} f[m] \sum_{k=0}^{N-1} e^{\frac{2\pi i k (n-m)}{N}}, \\
&= \frac{1}{N}\sum_{m=0}^{N-1} f[m] N \delta_{nm}, \\
&= f[n].
\end{aligned}$$

Cette relation montre qu'on a bien la même information dans la suite de longueur finie $\hat{f}[k]$ que dans $f[n]$. On a donc enlevé avec succès toute information redondante contenue dans $\hat{f}(\omega)$.

On peut maintenant de façon simple implémenter la transformée de Fourier discrète sur un ordinateur car on a discrétisé toutes les étapes du calcul. Néanmoins les formules ci-dessus ne sont pas d'une grande efficacité. En effet, on peut montrer que la complexité de l'équation éq. 5.86 est de l'ordre N^2.

On peut écrire l'éq. 5.86 comme un produit matrice-vecteur sous la forme suivante

$$\underbrace{\begin{pmatrix} \hat{f}[0] \\ \hat{f}[1] \\ f[2] \\ \vdots \\ \hat{f}[N-1] \end{pmatrix}}_{\vec{\hat{f}}} = \underbrace{\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & w & w^2 & \cdots & w^{N-1} \\ 1 & w^2 & w^4 & \cdots & w^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & w^{N-1} & w^{2(N-1)} & \cdots & w^{(N-1)^2} \end{pmatrix}}_{\underline{\underline{W}}} \cdot \underbrace{\begin{pmatrix} f[0] \\ f[1] \\ f[2] \\ \vdots \\ f[N-1] \end{pmatrix}}_{\vec{f}}, \tag{5.88}$$

où $w = e^{-\frac{2\pi i}{N}}$. On peut donc de façon plus compacte l'écrire

$$\vec{\hat{f}} = \underline{\underline{W}} \cdot \vec{f}. \tag{5.89}$$

Les éléments de la matrice $\underline{\underline{W}}$ peuvent être précalculés et il reste donc à calculer uniquement le produit matrice vecteur $\underline{\underline{W}} \cdot \vec{f}$. Pour ce faire il faut pour chaque

ligne de $\hat{\vec{f}}$ faire le calcul de N produits et N sommes (donc une complexité N). Comme il y a N lignes à $\hat{\vec{f}}$, la complexité est $N \cdot N$.

Il existe des algorithmes beaucoup plus efficaces pour effectuer de genre de calculs que nous allons brièvement discuter maintenant. Ils réduisent la complexité algorithmique à $N\log(N)$ en général. Nous allons brièvement discuter un de ces algorithmes dans la sous-section sec. 5.6.4.

La transformée de Fourier discrète étant un échantillonnage de la transformée de Fourier à temps discret, toutes les propriétés discutées pour la transformée de Fourier à temps discret restent valides. En particulier la transformée de Fourier discrète est périodique, de période N

$$\hat{f}[k] = \hat{f}[k+N]. \tag{5.90}$$

Exercice 25

A démontrer en exercice.

5.6.4 La transformée de Fourier rapide

L'algorithme présenté ici est une version "simplifiée" de l'algorithme de Cooley-Tukey (publié en 1965). Cet algorithme a en fait été "inventé" par Gauss en 1805 quand il essayait d'interpoler la trajectoires d'astéroïdes dans le système solaire.

L'idée de l'algorithme radix-2 est d'abord de séparer le signal en deux parties. D'une part les indices pairs et d'autres part les indices impairs

$$\begin{aligned} \{f[2m]\}_{m=0}^{N/2-1} &= \{f[0], f[2], ..., f[N-2]\}, \\ \{f[2m+1]\}_{m=0}^{N/2-1} &= \{f[1], f[3], ..., f[N-1]\}. \end{aligned} \tag{5.91}$$

Puis les transformées de Fourier discrètes de chacune de ces sous-suites sont calculées et combinées pour avoir la transformée de Fourier du signal en entier. En fait on va appliquer cette décomposition de façon récursive sur chacune des deux parties. On fait donc l'hypothèse que la longueur du signal est une puissance de 2. Ce n'est en pratique pas un problème, car on peut facilement rajouter des "zéros" dans notre signal pour avoir un signal d'une longueur d'une puissance de 2.

Commençons donc par réécrire la transformée de Fourier $\hat{f}[k]$ lorsqu'on a décomposé le signal en deux sous-signaux

$$\begin{aligned} f[k] &= \sum_{m=0}^{N/2-1} f[2m] e^{-\frac{2\pi i (2m) k}{N}} + \sum_{m=0}^{N/2-1} f[2m+1] e^{-\frac{2\pi i (2m+1) k}{N}}, \\ &= \sum_{m=0}^{N/2-1} f[2m] e^{-\frac{2\pi i m k}{N/2}} + e^{-\frac{2\pi i k}{N}} \sum_{m=0}^{N/2-1} f[2m+1] e^{-\frac{2\pi i m k}{N/2}}, \\ &= \hat{p}[k] + e^{-\frac{2\pi i k}{N}} \hat{j}[k], \end{aligned}$$

où nous avons défini les transformées de Fourier discrètes des parties paires et impaires $p[k]$ et $\hat{j}[k]$

$$\begin{aligned}\hat{p}[k] &= \sum_{m=0}^{N/2-1} f[2m]e^{-\frac{2\pi imk}{N/2}},\\ \hat{j}[k] &= \sum_{m=0}^{N/2-1} f[2m+1]e^{-\frac{2\pi imk}{N/2}}.\end{aligned} \tag{5.92}$$

La transformée de Fourier discrète étant périodique (comme l'est la transformée de Fourier à temps discret), nous avons les propriétés suivantes

$$\begin{aligned}\hat{p}[k] &= \hat{p}[k+N/2],\\ \hat{j}[k] &= \hat{j}[k+N/2].\end{aligned} \tag{5.93}$$

De plus, nous avons que

$$e^{-\frac{2\pi i(k+N/2)}{N}} = e^{-\pi i}e^{-\frac{2\pi ik}{N}} = -e^{-\frac{2\pi ik}{N}}. \tag{5.94}$$

Avec ces propriétés il est aisé de réécrire

$$\hat{f}[k] = \begin{cases} \hat{p}[k] + e^{-\frac{2\pi ik}{N}}\hat{j}[k], & \text{si } 0 \leq k < N/2\\ \hat{p}[k] - e^{-\frac{2\pi ik}{N}}\hat{j}[k], & \text{si } N/2 \leq k < N\end{cases} \tag{5.95}$$

On a donc réduit le nombre de calculs nécessaires pour calculer $\hat{f}[k]$ d'un facteur 2. En continuant cette procédure jusqu'à $N = 2$ on peut montrer qu'on réduit la complexité algorithmique à $N \log N$ (mais on ne le démontrera pas dans ce cours).

5.6.5 Fréquence d'échantillonnage

Une question primordiale dans le calcul des transformée de Fourier (ou de l'analyse spectrale plus généralement) est la question de l'échantillonnage du signal que nous souhaitons analyser. Dans le monde réel un signal sonore, une image,... est considéré comme une quantité continue (il est représentée par une infinité de valeur). Lorsque nous souhaitons faire une analyse spectrale sur un ordinateur de ce signal, il est nécessaire de le digitaliser: de le rendre discret. Dès lors une question très importante est de savoir quelle est la fréquence à laquelle on va enregistrer les valeurs de notre suite temporelle afin de garder toute l'information contenue dans le signal original.

En termes mathématiques, nous avons un signal $f(t)$ que nous enregistrons entre t_0 et t_{N-1}. Nous voulons le transformer en un signal de longueur N finie, $f(t_n)$ avec $0 \leq n \leq N-1$ afin de pouvoir le représenter sur un support numérique. Pour simplifier on va supposer que l'enregistrement se fait à intervalle régulier, $\delta t = \frac{t_{N-1}-t_0}{N-1}$. On a donc que $t_n = t_0 + \delta tn$. La question qu'on se pose est quelle doit être la valeur de N pour ne pas perdre d'information sur $f(t)$ quand on échantillonne. En d'autres termes à partir de quel nombre N d'échantillons la transformée de Fourier discrète de $f[n]$ ne change plus.

Le théorème de Shannon-Nyquist nous dit que pour pouvoir représenter exactement un signal avec une fréquence maximale $F_c = 1/\delta t_c$, alors on doit l'échantillonner avec une fréquence $1/\delta t_e = F_e \geq 2F_c$. De façon similaire, si on choisit un signal et qu'on peut l'échantillonner avec une certaine précision (on détermine la fréquence maximale, F_c qu'on veut pouvoir représenter dans le signal) on a simplement besoin de choisir une fréquence d'échantillonnage $F_e \geq 2F_c$. Nous notons $F_N = 2F_c$ la fréquence de Nyquist. En prenant $F_e = F_N$ on a que $N = 1/F_e = 1/F_N$ et que l'échantillonnage permet de représenter les fréquences plus petites que $F_N/2$. Si la fréquence d'échantillonnage est plus petite que la fréquence de Nyquist de notre signal, on verra apparaître le phénomène de *repliement de spectre* (aliasing en anglais).

Chapitre 6

Probabilités et statistiques

6.1 Introduction à la statistique descriptive

En statistique, une *population* est un ensemble d'objets (d'individus) possédant un ou plusieurs *caractères* communs. L'étude des caractères d'une population a pour but de révéler des tendances au sein de la population. Ces études sont particulièrement intéressantes quand le nombre d'individus de notre population est trop élevé pour pouvoir être analysé en entier. On prélève alors un échantillon "représentatif" de notre population au hasard et on mène l'analyse statistique sur ce sous ensemble. Les éventuelles conclusions de l'étude statistique sur le sous ensemble seront ensuite appliquées à l'ensemble de la population. Grâce au calcul des probabilités nous pourrons avoir une confiance plus ou moins grande dans les conclusions tirées en fonction de la taille de l'échantillon. En effet plus celui-ci sera grand, plus la confiance dans les résultats sera élevée.

Un exemple de ce genre d'étude qui est très à la mode ces temps est le sondage (concernant le résultat d'élections ou de votations). Les sondeurs tentent en questionnant un sous-ensemble d'environ 1000 d'électeurs d'un pays (citoyens de plus de 18, moitié d'hommes et de femmes plus ou moins, ...) de prévoir les résultats d'élections ou de votations où participeront des millions d'électeurs potentiels. Il faut avouer que la tâche semble pour le moins complexe. Et la plus grande difficulté tient dans le "représentatif de la population".

6.1.1 Représentations

Il existe différentes façon de représenter les caractères d'une population selon que sa nature est *discrète* ou *continue*. Dans le cas discret d'un caractère pouvant prendre $k \in \mathbb{N}$ valeur différentes $\{x_i\}_{i=0}^{k-1}$, on représente le nombre d'individus pouvant prendre la valeur x_i par le nombre n_i. On a donc un ensemble $\{n_i\}_{i=0}^{k-1}$ d'individus pour les k valeurs des caractères de la population. Dans le cas continu le nombre d'individus d'un caractère correspondrait à une subdivision en k parties de l'ensemble des valeurs possibles pour le dit caractère.

Illustration 26

1. Cas discret: On étudie la distribution de salaires annuels dans une entreprise. Les salaires possibles sont 40′000, 50′000, 60′000 et 1′000′000 CHF.
   — Il y a 35 personnes payées 40′000 CHF.
   — Il y a 20 personnes payées 50′000 CHF.
   — Il y a 5 personnes payées 60′000 CHF.
   — Il y a 1 personne payée 1′000′000 CHF.
2. Cas continu: Lors du benchmark d'une application, A, nous effectuons plusieurs mesures (la population) du temps d'exécution (le caractère) de l'application. Les résultats obtenus sont les suivants:
   — 7 exécutions ont pris entre 50 et 51 secondes.
   — 12 exécutions ont pris entre 51 et 52 secondes.
   — 8 exécutions ont pris entre 52 et 53 secondes.
   — 23 exécutions ont pris entre 53 et 54 secondes.

Pour représenter de façon un peu plus parlante ces valeurs, deux méthodes principales existent: le tableau ou le graphique. Pour illustrer les exemples précédents sous forme de tableau on obtient pour le cas des salaires (voir Tabl. tbl. 6.1)

TABLE 6.1 – Tableau du nombre de salariés par salaire.

Salaire	Nombre de salariés
40000	35
50000	20
60000	5
1000000	1

et du benchmark de l'application (voir Tabl. tbl. 6.2)

TABLE 6.2 – Tableau du temps d'exécution et du nombre d'exécutions.

Temps d'exécution	Nombre
[50,51)	7
[51,52)	12
[52,53)	8
[53,54)	23

Sous forme de graphique on peut représenter le tableau des salaires sous la forme d'un graphique bâton (voir Fig. fig. 6.1)

ou d'un histogramme pour le temps d'exécution de l'application (voir Fig. fig. 6.2).

6.1.2 Fréquences

Plutôt que de faire apparaître le nombre d'individus d'une population possédant un caractère, il peut être plus intéressant de faire intervenir la *fréquence* ou

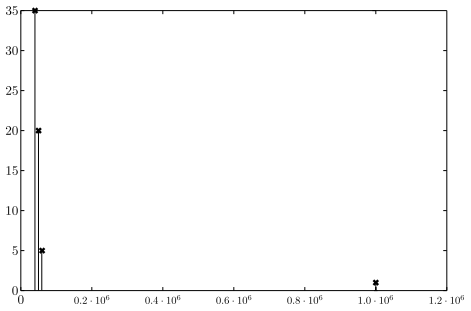

FIGURE 6.1 – Nombre salariés en fonction du salaire.

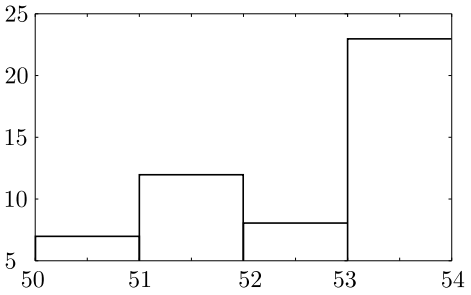

FIGURE 6.2 – Nombre d'exécutions en fonction du temps d'exécution.

le nombre relatif à la place. En effet, la fréquence donne immédiatement la proportion d'individus plutôt qu'un nombre absolu qui n'est pas forcément très interprétable tout seul.

La population totale, n, est donnée par

$$n = \sum_{i=0}^{k-1} n_i. \tag{6.1}$$

On peut donc définir la fréquence d'un caractère i, f_i comme

$$f_i = \frac{n_i}{n}. \tag{6.2}$$

Illustration 27 *(Fréquences)*

Les tableaux de fréquence des deux exemples précédents sont donnés par

1. Cas discret: la population totale est de

$$n = 35 + 20 + 5 + 1 = 61. \tag{6.3}$$

TABLE 6.3 – Tableau des salaires, du nombre de salariés et la fréquence.

Salaire	Nombre de salariés	Fréquence
40000	35	$35/61 \cong 0.573770$
50000	20	$20/61 \cong 0.327869$
60000	5	$5/61 \cong 0.081967$
1000000	1	$1/61 \cong 0.016393$

2. Cas continu: la population totale est de

$$n = 7 + 12 + 8 + 23 = 50. \tag{6.4}$$

Le tableau tbl. 6.4 affiche les différentes fréquences des temps d'exécution.

TABLE 6.4 – Tableau des temps d'exécution et la fréquence des temps d'exécution.

Temps d'exécution	Nombre	Fréquence
[50,51)	7	$7/50 = 0.14$
[51,52)	12	$12/50 = 0.24$
[52,53)	8	$8/50 = 0.16$
[53,54)	23	$23/50 = 0.46$

La fréquence possède un certain nombre de propriétés que nous retrouverons dans les sections suivantes qui sont assez intuitives

---

Propriété 4 *(Propriétés de la fréquence)*

1. Les fréquences sont toujours dans l'intervalle $[0, 1]$

$$0 \leq f_i \leq 1. \tag{6.5}$$

2. La somme de toutes les fréquences donne toujours 1

$$\sum_{i=0}^{k-1} f_i = 1. \tag{6.6}$$

Relié avec la propriété 2 ci-dessus, il peut également être intéressant d'obtenir la *fréquence cumulée*, notée $F(x)$, d'un caractère qui se définit comme la fréquence des individus qui présentent une valeur de caractère $x_i \leq x$. Les tableaux correspondants aux tableaux tbl. 6.1 et tbl. 6.2 (voir le tbl. 6.5 et le tbl. 6.6)

TABLE 6.5 – Tableau des salaires, du nombre de salariés, et la fréquence et fréquence cumulée des salaires.

Salaire	Nombre de salariés	Fréquence	Fréquence cumulée
40000	35	$35/61 \cong 0.573770$	$35/61 \cong 0.573770$
50000	20	$20/61 \cong 0.327869$	$(20+35)/61 \cong 0.90164$
60000	5	$5/61 \cong 0.081967$	$(20+35+5)/61 \cong 0.98361$
1000000	1	$1/61 \cong 0.016393$	$(20+35+5+1)/61 = 1$

TABLE 6.6 – Tableau des temps d'exécution et la fréquence et fréquences cumulées des temps d'exécution.

Temps d'exécution	Nombre	Fréquence	Fréquence cumulée
[50,51)	7	$7/50 = 0.14$	$7/50 = 0.14$
[51,52)	12	$12/50 = 0.24$	$(7+12)/50 = 0.38$
[52,53)	8	$8/50 = 0.16$	$(7+12+8)/50 = 0.54$
[53,54)	23	$23/50 = 0.46$	$(7+12+8+23)/50 = 1$

Exercice 26 *(Fréquence cumulée)*

1. Tracer les graphes de la fréquence cumulée pour les deux exemples que nous avons vus.
2. Que pouvons-nous déduire de la forme de la fonction (croissance, valeur maximale)?

6.1.3 Mesures de tendance centrale

Jusqu'ici le nombre de valeurs étudiées était limité et il est assez simple d'avoir une vue d'ensemble de la distribution des valeurs des caractères de notre population. Mais en général il est plus aisé d'utiliser une nombre de valeurs beaucoup plus restreint permettant de résumer les différents caractères et nous allons en voir deux différents qui nous donne une tendance dite centrale: la moyenne, la médiane.

La *moyenne*, notée $\bar{x}$ d'un jeu de données s'obtient par la formule suivante

$$\bar{x} = \frac{1}{n} \sum_{i=0}^{k-1} x_i \cdot n_i. \tag{6.7}$$

La moyenne peut également être calculée via les fréquences

$$\bar{x} = \sum_{i=0}^{k-1} f_i \cdot x_i. \tag{6.8}$$

Exercice 27 *(Propriétés de la moyenne)*

1. Démontrer la relation précédente.
2. Démontrer que la moyenne des écart $x_i - \bar{x}$ est nulle.

Illustration 28 *(Moyenne)*

Pour l'exemple des salaires la moyenne est donnée par

$$\bar{x}_{\text{salaire}} = \frac{35 \cdot 40000 + 20 \cdot 50000 + 5 \cdot 60000 + 1 \cdot 1000000}{61} = 60656. \tag{6.9}$$

On remarque ici que la moyenne des salaires donne une impression erronée de la situation car elle est très sensible aux valeurs extrême de la distribution. En effet, tous les salaires à l'exception d'un sont inférieurs à la moyenne. Il suffit de retirer le salaire d'un million de notre ensemble de valeurs, la moyenne de l'échantillon restant devient

$$\bar{x}_{\text{salaire}} = \frac{35 \cdot 40000 + 20 \cdot 50000 + 5 \cdot 60000}{60} = 45000. \tag{6.10}$$

La différence est de l'ordre de 25% par rapport aux 60′000 CHF obtenus avec toute la population. Il est donc nécessaire d'utiliser une autre mesure pour illustrer mieux le salaire caractéristique de notre population. De façon plus générale la moyenne est peu robuste à des valeurs extrêmes dans l'étude d'échantillons.

Une mesure qui est plus parlante est la *médiane*, notée $\tilde{x}$. La médiane se définit comme la valeur $\tilde{x}$ qui est telle que la moitié des individus de la population ont un $x_i \leq \tilde{x}$ et le reste est telle que $x_i \geq \tilde{x}$.

Pour l'exemple des salaires le salaire médian est de $40000CHF$, ce qui reflète beaucoup mieux la distribution des salaire de notre population.

Exercice 28 *(Moyenne, médiane)*

Calculer la moyenne et la médiane pour l'exemple du temps d'exécution (prendre la borne inférieure des intervalles pour chaque temps d'exécution [1]).

6.1.4 Mesures de dispersion

Nous avons vu deux mesures donnant une tendance générale des caractères d'une population. Hors ces valeurs ne nous disent absolument rien sur la manière dont ces caractères sont distribués. Sont-ils proches de la moyenne ou de la médiane? Ou en sont-ils au contraire éloignés? Nous allons voir deux mesures différentes dans cette sous-section: la variance (écart-type), et l'intervalle inter-quartile.

1. Il y a 7 temps de 50s, 12 de 51s, 8 de 52s et 23 de 53s.

Nous cherchons d'abord à calculer la moyenne des écarts à la moyenne. Hors, comme on l'a vu dans la sous-section précédente l'écart à la moyenne $x_i - \bar{x}$ est nul en moyenne. Cette grandeurs ne nous apprend rien. On peut donc s'intéresser plutôt à la moyenne de l'écart quadratique $(x_i - \bar{x})^2$ qui est une quantité toujours positive et dont la moyenne aura toujours une valeur positive ou nulle (elle sera nulle uniquement si $x_i - \bar{x} = 0, \forall i$) [2]. On définit donc la *variance*, v, comme étant la moyenne des écarts quadratiques

$$v = \frac{1}{n} \sum_{i=0}^{k-1} n_i (x_i - \bar{x})^2. \tag{6.11}$$

Si on considère la racine carrée de la variance, on obtient *l'écart-type*

$$s = \sqrt{v}. \tag{6.12}$$

Exercice 29 *(Variance, écart-type)*

Démontrer les relations suivantes

1. On peut également calculer la variance avec les fréquences

$$v = \sum_{i=0}^{k-1} f_i (x_i - \bar{x})^2. \tag{6.13}$$

2. On peut également calculer la variance à l'aide de la formule suivante

$$v = \frac{1}{n} \left(\sum_{i=0}^{k-1} n_i x_i^2 \right) - \bar{x}^2 = \bar{x^2} - \bar{x}^2 \tag{6.14}$$

Pour l'exemple du salaire on obtient pour la variance

$$\begin{aligned} v = \frac{1}{61} \big(& 35 \cdot (40000 - 60656)^2 + 20 \cdot (50000 - 60656)^2 \\ & + 5 \cdot (60000 - 60656)^2 + 1 \cdot (1000000 - 60656)^2 \big) \\ = & \, 1.4747 \cdot 10^{10}, \end{aligned}$$

et l'écart-type

$$s = \sqrt{v} = 121440. \tag{6.15}$$

Exercice 30 *(Variance, écart-type)*

Calculer la variance et l'écart type à partir des valeurs du benchmark de l'application.

---

2. on pourrait aussi étudier la moyenne de $|x_i - \bar{x}|$, mais cela est moins pratique à étudier théoriquement.

Encore une fois on constate que la valeur de l'écart-type des salaires est très dépendante de la valeur extrême de la distribution (1000000 CHF). Si on l'enlève la valeur de l'écart type est de $s = 6455$ (un facteur 20 plus petit que la valeur sur la population complète).

Comme pour la moyenne et la médiane nous pouvons définir des valeurs plus représentatives. A partir de la fréquence cumulée, F, on peut définir deux grandeurs, $Q_i \in \{x_i\}_{i=0}^{k-1}$ et $\alpha_i \in [0, 1]$ telles que

$$F(Q_i) = \alpha_i. \tag{6.16}$$

En d'autres termes Q_i est la valeur pour laquelle la fréquence cumulée vaut α_i. Q_i correspond donc au nombre d'individus dont la fréquence cumulée est de α_i. En particulier si $\alpha_i = 1/2$, alors $Q_i = \tilde{x}$ (Q_i est la médiane). Il est commun d'avoir $Q_i \in [0.25, 0.5, 0.75]$, on parle alors de quartiles. Avec $Q_1 = 0.25$ et $Q_3 = 0.75$, le nombre d'individus entre 0.25 et 0.75 est donné par

$$\frac{Q_3 - Q_1}{2}. \tag{6.17}$$

Cette valeurs est appelée l'intervalle semi-inter-quartile.

Exercice 31 *(Semi-inter quartile)*

Calculer les intervalles semi-inter-quartiles des exemples que nous avons vus plus tôt dans le cours.

6.2 Probabilités: Exemple du jeu de dé

On considère un dé à 6 faces. Le lancer de dé est une *expérience aléatoire*, car on ne peut dire quel sera le résultat avant d'avoir effectué l'expérience.

Avant de commencer à étudier les probabilités du lancer de dé, et les questions qu'on peut se poser, faisons d'abord un peu de vocabulaire qui sera utile pour la suite.

Définition 24

- L'ensemble des résultats possibles du lancer de dé est $\Omega = \{1, 2, 3, 4, 5, 6\}$ et cet ensemble est appelé l'*univers* du lancer de dé.
- Chaque résultat possible du lancer de dé (1, 2, etc), noté $\omega \in \Omega$, est appelé une *éventualité*.
- Un ensemble de résultats possibles, par exemple tous les résultats pairs du lancer de dé $A = \{2, 4, 6\} \in \Omega$, s'appelle un *événement*. Un événement composé d'une seule éventualité est appelé *événement élémentaire*.
- On dit que l'événement A est *réalisé* si on obtient 2, 4, ou 6 en lançant le dé.
- *L'événement certain* est l'univers en entier. On est certain de réaliser l'événement.

— *L'événement impossible* est l'ensemble vide, $A = \emptyset$. Il correspondrait à l'événement obtenir 7 ou plus en lançant un dé par exemple.
— Si A est un événement, on note $p(A)$ la *probabilité* que A soit réalisé.

Le calcul des *probabilités* de réalisation de certains événement est reliée à la *fréquence* que nous avons introduit dans la section précédente. Soit un univers Ω et A, B deux événements tels que $A \cap B = \emptyset$. On effectue N expériences, donc Ω est réalisé N fois. De plus on constate qu'on réalise A, K fois et B, M fois. On a donc les fréquences suivantes que A, B et Ω se réalisent

$$\begin{aligned} f(A) &= \frac{K}{N}, \\ f(B) &= \frac{M}{N}, \\ f(\Omega) &= \frac{N}{N} = 1, \\ f(A \cup B) &= \frac{M+K}{N} = f(A) + f(B). \end{aligned} \tag{6.18}$$

Les *probabilités* de réalisation des événements ci-dessus peuvent être vues comme le passage à la limite $N \to \infty$ tel que $p(A), p(B) \in \mathbb{R}$ et

$$\begin{aligned} p(A) &= \lim_{\substack{N\to\infty, \\ K/N<\infty}} \frac{K}{N}, \\ p(B) &= \lim_{\substack{N\to\infty, \\ M/N<\infty}} \frac{M}{N}, \\ p(\Omega) &= 1, \\ p(A \cup B) &= p(A) + p(B). \end{aligned} \tag{6.19}$$

Si maintenant nous voulons connaître la probabilité de tirer 6, ou encore la probabilité de réaliser $A = \{6\}$. Cela est assez intuitif pour le cas du dé. Nous avons 6 éléments dans l'univers du lancer de dé. La probabilité de réaliser $A = \{6\}$ est donc

$$p(6) = \frac{1}{6}. \tag{6.20}$$

Pour le cas du lancer de dé, on dit qu'on a un processus qui est *équiprobable*. En effet, la probabilité de réaliser chacun des événements élémentaires est la même. On a en effet la même probabilité de tirer 1, 2, 3, 4, 5, ou 6.

Si à présent, on se pose la question de la probabilité de réaliser un tirage pair, $A = \{2, 4, 6\}$, alors on trouve

$$p(\text{tirer un nombre pair}) = \frac{1}{2}. \tag{6.21}$$

De façon générale pour le lancer de dé, on a que la probabilité de réaliser

l'événement A est [3]

$$p(A) = \frac{\text{nombre d'éléments dans } A}{\text{nombre d'éléments dans } \Omega}. \tag{6.22}$$

Si maintenant, on veut savoir quelle est la probabilité de tirer n'importe quel élément dans l'univers, on a

$$p(\Omega) = \frac{\text{nombre d'éléments dans } \Omega}{\text{nombre d'éléments dans } \Omega} = 1. \tag{6.23}$$

De même la probabilité de réaliser l'événement impossible est de

$$p(\emptyset) = \frac{\text{nombre d'éléments dans } \emptyset}{\text{nombre d'éléments dans } \Omega} = 0. \tag{6.24}$$

On voit ici une propriété fondamentale des probabilités qui est que $0 \leq p(A) \leq 1$, $\forall A$.

La probabilité de ne pas tirer un 6 donc de réaliser l'événement $\bar{A} = \{1, 2, 3, 4, 5\}$ est donnée par 1 moins la probabilité de réaliser $A = \{6\}$, il vient

$$p(\bar{A}) = 1 - p(A) = \frac{5}{6}. \tag{6.25}$$

De même la probabilité de tirer un nombre impair, est donnée par 1 moins la probabilité de réaliser l'événement pair

$$p(\{1, 3, 5\}) = 1 - p(\{2, 4, 6\}) = \frac{1}{2}. \tag{6.26}$$

6.2.1 Événements disjoints

Considérons maintenant deux événements, $A = \{1, 2\}$ et $B = \{3, 4, 5\}$. Comme A et B n'ont pas d'éléments en commun, on dit que c'est deux événements *disjoints*. Les probabilités de réalisation de ces événements sont donc

$$\begin{aligned} p(A) &= \frac{2}{6} = \frac{1}{3}, \\ p(B) &= \frac{3}{6} = \frac{1}{2}. \end{aligned} \tag{6.27}$$

On va se poser deux questions à présent

1. On cherche à savoir quelle est la probabilité de réaliser A ou de réaliser B, donc de tirer un dé dont le résultat sera dans l'ensemble $C = A \cup B = \{1, 2, 3, 4, 5\}$. Le résultat est

$$p(C) = \frac{5}{6}. \tag{6.28}$$

   Une coincidence intéressante (qui n'est en fait pas une coincidence) est que

$$p(C) = p(A) + p(B) = \frac{1}{3} + \frac{1}{2} = \frac{5}{6}. \tag{6.29}$$

3. De façon générale cela n'est pas vrai. Imaginons que nous ayons un sac avec 3 boules: 2 noires et une blanche. La probabilité de réaliser A: tirer une boule noire ($p(A) = 2/3$) ou B: tirer une boule blanche ($p(B) = 1/3$) n'est pas donnné par $p(A) =$ nombre d'éléments dans A/nombre total d'éléments $= 1/2$, $p(B) =$ nombre d'éléments dans B/nombre total d'éléments $= 1/2$.

2. On cherche à savoir quelle est la probabilité de réaliser A et réaliser B en même temps, donc de tirer un dé qui sera dans l'ensemble $C = A \cap B = \emptyset$. Ici on a déjà vu que la probabilité $p(\emptyset) = 0$.

On voit donc que si des événements sont disjoints, alors la probabilité de réaliser l'un ou l'autre des événements est simplement la somme des probabilités de réaliser chacun des événements. Inversément la probabilité de réaliser les deux événements en même temps est nulle.

Nous pouvons facilement décomposer A en deux sous événements élémentaires, $A = \{1\} \cup \{2\}$. On a donc une autre façon de calculer $p(A)$

$$p(A) = p(\{1\}) + p(\{2\}) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}. \tag{6.30}$$

On a que la probabilité de réaliser un événement est la somme des événements élémentaires qui le composent.

6.2.2 Événements complémentaires

Considérons de nouveau l'événement $A = \{1, 2\}$ et cette fois l'événement $B = \Omega \backslash \{1, 2\} = \{3, 4, 5, 6\}$. L'événement B est appelé *l'événement complémentaire* de A. Il est noté $B = \bar{A}$. Les probabilité de réaliser A ou de réaliser $\bar{A}$ est la même chose que de réaliser l'événement certain, car $A \cup \bar{A} = \Omega$. On vérifie aisément dans ce cas que

$$\Omega = \{1, 2\} \cup \{3, 4, 5, 6\} \tag{6.31}$$

et

$$p(A \cup \bar{A}) = p(\Omega) = 1. \tag{6.32}$$

De plus de ce qu'on a vu précédemment, on a que

$$p(A \cup \bar{A}) = p(A) + p(\bar{A}). \tag{6.33}$$

En combinant ces deux derniers résultats, il vient que

$$p(A) + p(\bar{A}) = 1. \tag{6.34}$$

On en déduit que

$$p(A) = 1 - p(\bar{A}) = 1 - \frac{2}{3} = \frac{1}{3}. \tag{6.35}$$

Dans ce cas on peut également calculer à priori $p(B)$

$$p(B) = \frac{\text{nombre d'éléments dans } B}{\text{nombre d'éléments dans } \Omega} = \frac{4}{6} = \frac{2}{3}. \tag{6.36}$$

Ce résultat est très important car on calcule facilement $p(\bar{A})$ si on connaît $p(A)$.

6.2.3 Événements non-disjoints

Considérons de nouveau l'événement $A = \{1, 2\}$ et cette fois $B = \{2, 3, 4, 5\}$. Les probabilités de réaliser les événements respectifs sont

$$\begin{aligned} p(A) &= \frac{1}{3}, \\ p(B) &= \frac{2}{3}. \end{aligned} \tag{6.37}$$

La probabilité de réaliser A et B est maintenant la probabilité de réaliser $C = A \cap B = \{2\}$

$$p(C) = \frac{1}{6}. \tag{6.38}$$

Si on cherche à présent la probabilité de réaliser A ou B, $D = A \cup B = \{1, 2, 3, 4, 5\}$, on voit aisément que

$$p(D) = \frac{5}{6}. \tag{6.39}$$

Comme A et B ne sont pas disjoints ont constate

$$\frac{5}{6} = p(D) \neq p(A) + p(B) = 1. \tag{6.40}$$

L'inégalité est dûe au fait que dans le cas où on fait la somme $p(A) + p(B)$ on compte à double la probabilité de tirer l'éventualité 2, qui est l'intersection de A et de B. Afin de corriger donc le calcul de $p(D)$ à partir de la somme $p(A)+p(B)$ il suffit d'enlever la probabilité de tirer l'intersection C. On a donc

$$\frac{5}{6} = p(D) = p(A) + p(B) - p(C) = 1 - \frac{1}{6} = \frac{5}{6}. \tag{6.41}$$

De façon complètement générale, on a la relation suivante pour calculer la probabilité de réaliser l'union de deux événement A et B

$$p(A \cup B) = p(A) + p(B) - p(A \cap B). \tag{6.42}$$

Il en suit immédiatement que si $A \cap B = \emptyset$, alors

$$p(A \cup B) = p(A) + p(B) - p(A \cap B) = p(A) + p(B) - p(\emptyset) = p(A) + p(B). \tag{6.43}$$

6.2.4 Axiomes des probabilités

Tous ces concepts que nous avons vus précédemment peuvent être vus comme la conséquences des trois axiomes des probabilités suivants

---

Définition 25 *(Axiomes des probabilités)*

Soit Ω un univers. La probabilité de réaliser un événement $A \subseteq \Omega$ est une fonction $p(A)$ qui associe à tout événement de A un nombre réel, qui satisfait les 3 axiomes suivants

1. Une probabilité est TOUJOURS positive

$$p(A) \geq 0. \tag{6.44}$$

2. La probabilité de l'événement certain vaut 1

$$p(\Omega) = 1. \tag{6.45}$$

3. Soit $B \subseteq \Omega$. Si $A \cap B = \emptyset$, alors

$$p(A \cup B) = p(A) + p(B). \tag{6.46}$$

La probabilité de réalisation de deux événements incompatibles est égale à la somme de réalisation de chacun d'entre eux.

De ces axiomes découlent tout un tas de théorèmes

Théorème 6

Pour $A, B \subseteq \Omega$ et Ω un univers et p une probabilité.

1. $p(B \cap \bar{A}) = p(B) - p(B \cap A)$.
2. $p(\emptyset) = 0$.
3. $p(\bar{A}) = 1 - p(A)$.
4. $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.
5. $p(\bar{A} \cap \bar{B}) = 1 - p(A \cup B)$.
6. Si A et B sont disjoints, alors $p(A \cup B) = p(A) + p(B)$.
7. Si $A \subseteq B$, alors $p(B \cap \bar{A}) = p(B) - p(A)$.
8. Si $A \subseteq B$, alors $p(A) \leq p(B)$.
9. $\forall A$, $0 \leq p(A) \leq 1$.

6.2.5 Probabilités conditionnelles

Imaginons à présent que nous ayons une information supplémentaire lorsque nous lançons notre dé. Supposons par exemple que nous sachions lorsque nous lançons le dé que le résultat est pair. A partir de là la probabilité de tirer un 6 est de

$$p(6 \text{ sachant que le résultat du lancer est un nombre pair}) = 1/3, \tag{6.47}$$

alors que sans l'information sur la parité nous aurions eu $p(6) = 1/6$.

Lorsque nous rajoutons comme condition la réalisation préalable d'un événement B à la réalisation d'un événement A, nous parlons de probabilité conditionnelle, notée $P(A|B)$ (probabilité conditionnelle de A sachant que B s'est produit).

Essayons à présent de voir comment nous pouvons calculer de façon générale les probabilités conditionnelles avec notre exemple ci-dessus. Nous avons donc que nous cherchons à calculer $p(A|B) = p(6|2, 4, 6)$. Nous avons dans ce cas que $p(A) = 1/6$, $p(B) = 1/2$ et $p(A \cap B) = p(6) = 1/6$. Par ailleurs, nous pouvons remarquer que

$$p(A|B) = \frac{1}{3} = \frac{p(A \cap B)}{p(B)} = \frac{\frac{1}{6}}{\frac{1}{2}}. \tag{6.48}$$

Nous pouvons vérifier cette relation sur un exemple un peu plus compliqué. Soit $A = 1, 2, 4$ et $B = 2, 4, 6$. La probabilité conditionnelle $p(A|B)$ revient au calcul

de la probabilité de $p(A \cap B|B) = p(2, 4|2, 4, 6) = 2/3$. Avec notre formule, nous avons $p(A \cap B) = 1/3$ et $p(B) = 1/2$. Il vient donc

$$p(A|B) = \frac{p(A \cap B)}{p(B)} = \frac{2}{3}. \tag{6.49}$$

Cette formule peut en fait être vue comme la définition de la probabilité conditionnelle. Si $p(B) \neq 0$ alors on appelle probabilité conditionnelle le nombre $p(A|B)$, tel que

$$p(A|B) = \frac{p(A \cap B)}{p(B)}. \tag{6.50}$$

Exercice 32 *(Probabilités conditionnelles)*

Sur une population de 1000 hommes qui naissent, 922 atteignent l'âge de 50 ans et 665 l'âge de 70 ans.

1. Quelle est la probabilité qu'un homme qui vient de naître soit encore en vie à 50 ans?
2. Quelle est la probabilité qu'un homme qui vient de naître soit encore en vie à 70 ans?
3. Quelle est la probabilité qu'un homme de 50 ans soit encore en vie à 70?

6.2.6 Événements indépendants

Prenons maintenant le cas "pathologique" où nous cherchons la probabilité conditionnelle $p(A|B)$, mais où la réalisation de B n'a aucune influence sur la réalisation de A. On a donc

$$p(A|B) = p(A). \tag{6.51}$$

Il vient

$$p(A|B) = \frac{p(A \cap B)}{p(B)} = p(A). \tag{6.52}$$

On en déduit que

$$p(A \cap B) = p(A) \cdot p(B). \tag{6.53}$$

On calcule aussi $p(B|A)$

$$p(B|A) = \frac{p(A \cap B)}{p(A)} = \frac{p(A) \cdot p(B)}{p(A)} = p(B). \tag{6.54}$$

Donc si A ne dépend pas de B, alors la réciproque est vraie aussi. Les événements qui satisfont la propriété de l'équation éq. 6.53 sont appelés indépendants. Dans le cas contraire ils sont appelé dépendants.

Afin d'illustrer l'indépendance, prenons à nouveau le jet de dé. Supposons que nous effectuions deux tirages de suite et que l'événement A soit "tirer un 6 au premier tirage" et que l'événement B soit "tirer un 2 au deuxième tirage". On a que

$$p(A) = \frac{1}{6}, \quad p(B) = \frac{1}{6}, \quad p(A \cap B) = \frac{1}{36}. \tag{6.55}$$

On a donc bien $p(A \cap B) = p(A) \cdot p(B)$ et les événements sont indépendants. Cela semble bien naturel étant donné que le premier tirage du dé ne va en rien influencer le résultat du deuxième tirage. Tout comme un tirage de l'euromillions d'une semaine ne va pas influencer le résultat de celui de la semaine suivante.

Exercice 33 *(Événements indépendants)*

On jette une pièce de monnaie deux fois de suite. Les résultats possible pour chaque jet sont: P, ou F.

1. Écrivez l'univers des événements.
2. Calculez les probabilités des événements A "face au premier jet", B "pile au second jet".
3. Calculez la probabilité $p(A \cap B)$.
4. Est-ce que les jets sont indépendants?

6.2.7 Tirages multiples

Jusqu'ici on a lancé le dé une fois et calculé la probabilité liée à ce lancer unique. A présent, on va tirer le dé plusieurs fois et calculer les probabilités d'obtenir des séquences de réalisations. Pour notre exemple on va prendre un cas où on tire le dé deux fois successivement. Ce type de tirage est appelé *tirage successif avec remise*, car les deux tirages sont successifs et indépendants entre eux (on va tirer deux fois le même dé). L'univers de cette expérience est la combinaison de tous les résultats obtenus avec chacun des dés

$$\Omega = \{11, 12, 13, 14, 15, 16, 21, 22, 23, 24, 25, 26, ..., 61, 62, 63, 64, 65, 66\}. \quad (6.56)$$

Il y a $6 \times 6 = 6^2 = 36$ résultats possibles à ce tirage. Il faut noter ici que l'ordre dans lequel le tirage a lieu est important; le tirage 26 est différent du tirage 62. On verra par la suite des exemples où cela n'est pas le cas.

On cherche à savoir quelle est la probabilité d'obtenir l'événement $A = \{26\}$.

Comme précédemment la probabilité de réaliser l'événement A est le nombre d'éléments dans A divisé par le nombre d'éléments dans Ω. La probabilité est donc immédiatement obtenue

$$p(A) = \frac{1}{36}. \quad (6.57)$$

Une autre façon de visualiser ce genre de réalisation est de l'écrire sous forme d'arbre (voir la figure fig. 6.3).

Comme pour le cas à un tirage, tout tirage successif de dés est équiprobable et la probabilité de chaque tirage est de 1/36.

Une autre façon de calculer la probabilité d'obtenir $A = \{26\}$ est de constater que la probabilité d'obtenir ce tirage successif est la probabilité de tirer 2, puis

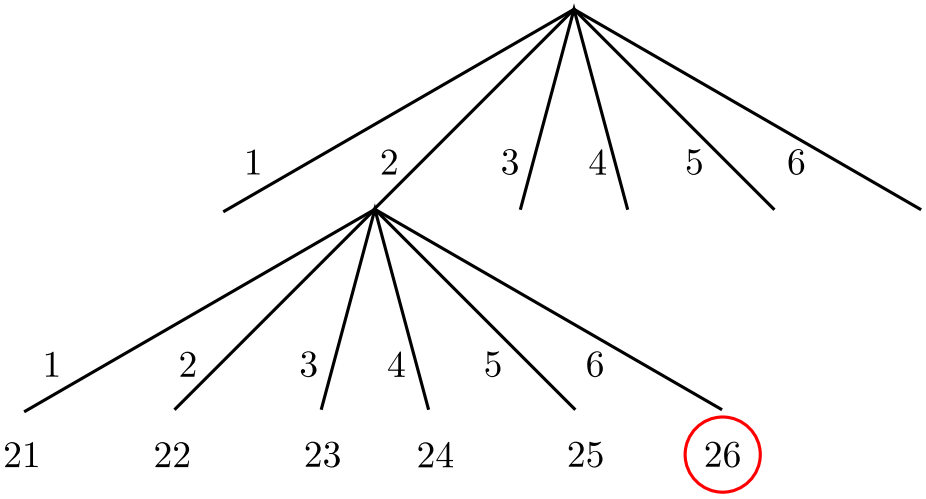

FIGURE 6.3 – Représentation du tirage 26 sous forme d'arbre.

la probabilité de tirer 6. La probabilité de cet enchaînement est obtenu en multipliant les événements élémentaires

$$p(\{26\}) = p(\{2\}) \cdot p(\{6\}) = \frac{1}{6} \cdot \frac{1}{6}. \tag{6.58}$$

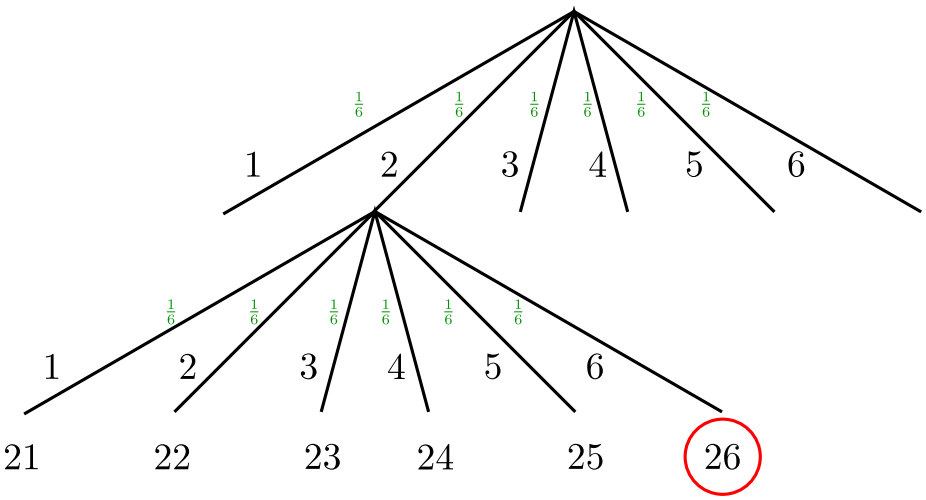

FIGURE 6.4 – Représentation du tirage 26 sous forme d'arbre avec les probabilités associées.

Afin de calculer la probabilité du tirage 26 il suffit de suivre le chemin menant de la racine à la feuille correspondante et de multiplier les probabilités inscrites sur chacune des branches.

Si à présent, nous voulons savoir quelle est la probabilité de tirer un 2 ou un 4 avec le premier dé et un nombre pair avec le second, on a trois façons de

calculer le résultat. La façon compliquée, où on compte toutes les possibilités. L'événement précédent s'écrit

$$A = \{22, 24, 26, 42, 44, 46\}. \tag{6.59}$$

On a donc que $p(A)$ est donné par

$$p(A) = \frac{\text{nombre d'éléments dans } A}{\text{nombre d'éléments dans } \Omega} = \frac{6}{36} = \frac{1}{6}. \tag{6.60}$$

L'autre façon (plus simple) est d'utiliser la propriété du produit des probabilité. Nous savons que la probabilité de tirer un 2 ou un 4 avec le premier dé est de $1/3$, puis la probabilité de tirer un nombre pair avec le deuxième est de $1/2$. On a donc finalement que

$$p(A) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}. \tag{6.61}$$

Finalement, on peut aussi utiliser la représentation sous forme d'arbre où on somme simplement les probabilités de chacun des éléments de A (voir figure fig. 6.5).

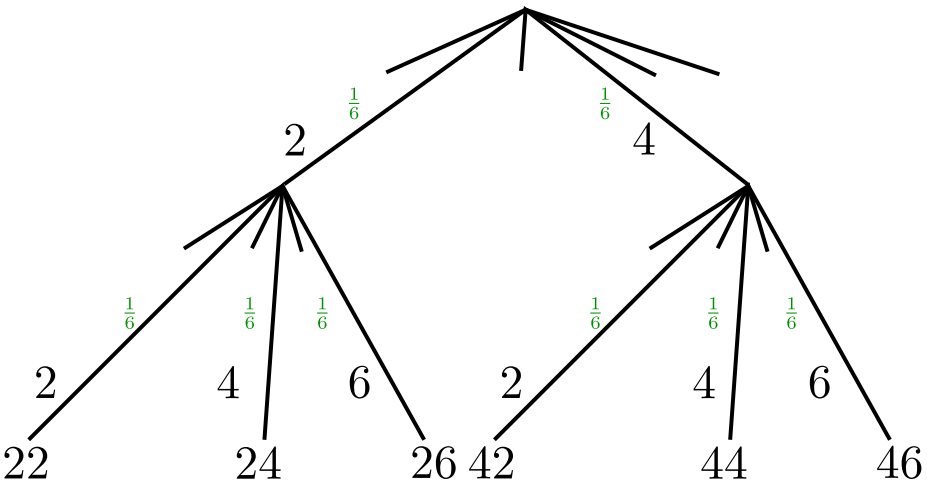

FIGURE 6.5 – Représentation de l'événement $A = \{22, 24, 26, 42, 44, 46\}$ sous forme d'arbre avec les probabilités associées. Toutes les probabilités et tirages possibles associés aux branches ne sont pas affichées pour simplifier l'affichage.

Comme vu dans la section sec. 6.2.1, il suffit de prendre la somme des probabilités des événements élémentaires

$$\begin{aligned} p(A) &= p(\{22\}) + p(\{24\}) + p(\{26\}) + p(\{42\}) + p(\{44\}) + p(\{46\}) \\ &= \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} \\ &= \frac{6}{36} = \frac{1}{6}. \end{aligned}$$

Si à présent l'ordre dans lequel les dés sont tirés n'a plus d'importance le calcul de probabilités change un peu. On désire savoir quelle est la probabilité d'obtenir

26 dans un ordre arbitraire. On peut donc obtenir cette combinaison en tirant 26 ou en tirant 62. On a donc $A = \{26, 62\}$. La probabilité de réaliser A est donc

$$p(A) = \frac{2}{36} = \frac{1}{18}. \tag{6.62}$$

On peut calculer cette probabilité de nouveau avec l'arbre ou en comptant. Une façon de nouveau plus simple dans bien des cas est d'utiliser les produits de probabilités. La probabilité de tirer 26 ou 62 est la probabilité de tirer d'abord 2 ou 6, puis de tirer le nombre restant (2 si on a d'abord tiré 6 ou 6 si on a d'abord tiré 2). La probabilité de tirer 2 ou 6 est de 1/3, puis la probabilité de tirer le nombre restant est de 1/6. On a donc que

$$p(A) = \frac{1}{3} \cdot \frac{1}{6} = \frac{1}{18}. \tag{6.63}$$

Exercice 34

1. Calculer la probabilité d'obtenir 2 comme la somme des deux nombres tirés par deux dés.
2. Calculer la probabilité d'obtenir $3, 4, 5, 6, 7, 8, 9, 10, 11, 12$ comme la somme des deux nombres tirés par deux dés.
3. Calculer la probabilité d'obtenir 7 comme la somme des deux nombres tirés par deux dés.
4. Calculer la probabilité d'obtenir 6 soit avec 1 soit avec 2 dés.
5. Déterminer le nombre de combinaisons possibles avec 3, 4, 5 dés. Pouvez vous généraliser à n dés?
6. Soit un tirage aléatoire offrant 2 possibilités (pile ou face par exemple). Quel est le nombre de combinaisons possibles si on tire n fois? Pouvez-vous généraliser pour un tirage aléatoire offrant m possibilités qu'on tire n fois?

6.2.8 La distribution multinomiale

Plus nous allons rajouter des tirages successifs plus il va être compliqué de calculer les probabilités de tirer une certaine combinaison de nombres. Il existe néanmoins une formule qui généralise les tirages successifs avec remise. Prenons le cas où nous avons un dé qui ne donne pas chaque nombre de façon équiprobable, mais avec probabilité $\{p_i\}_{i=1}^{6}$. Nous souhaitons savoir quelle est la probabilité de tirer deux fois le 1 et une fois le 2 lors de trois tirages successifs.

Dans ce tirage l'ordre dans lequel sont obtenus ces tirages ne sont pas importants. Il y a donc les tirages possibles qui sont admissibles

$$[112] = \{112, 121, 211\}. \tag{6.64}$$

On a donc que la probabilité associée est de

$$p([112]) = p(112) + p(121) + p(211). \tag{6.65}$$

Ces trois probabilités sont données par

$$\begin{aligned} p(112) &= p_1 \cdot p_1 \cdot p_2 = p_1^2 \cdot p_2, \\ p(121) &= p_1 \cdot p_2 \cdot p_1 = p_1^2 \cdot p_2, \\ p(211) &= p_2 \cdot p_1 \cdot p_1 = p_1^2 \cdot p_2. \end{aligned} \tag{6.66}$$

Les tirages étant indépendants on a que la probabilité de tirer 1 ou 2 est indépendante du moment où ils sont tirés et donc ces trois probabilités sont égales.

Finalement la probabilité de tirer deux 1 et un 2 est de

$$p([112]) = p(112) + p(121) + p(211) = 3 \cdot p_1^2 \cdot p_2. \tag{6.67}$$

A présent nous considérons la probabilité de tirer [1123] en 4 tirages. Les tirages possibles sont

$$[1123] = \{1123, 1132, 1213, 1231, 1312, 1321, 2113, 2131, 2311, 3112, 3121, 3211\}. \tag{6.68}$$

Il y a donc 12 tirages possibles pour cette combinaison. De plus les tirages étant indépendants on a que toutes ces combinaisons sont équiprobables avec probabilité

$$p(1123) = p_1^2 p_2 p_3. \tag{6.69}$$

Finalement on a

$$p([1123]) = 12 p_1^2 p_2 p_3. \tag{6.70}$$

Si nous définissons n_i le nombre de fois où on obtient le résultat i et qu'on cherche la probabilité de réaliser le tirage $[n_1, n_2, ..., n_k]$, on constate que la probabilité de réaliser le tirage est proportionnelle à $p_1^{n_1} p_2^{n_2} \cdots p_6^{n_6}$. Il nous reste à déterminer le facteur multiplicatif venant devant. Pour le cas du tirage $1, 1, 2$, nous avons $[n_1 n_2]$ avec $n_1 = 2$ et $n_2 = 1$ et le facteur devant le produit des probabilités est donné par 3. Pour le tirage $1, 1, 2, 3$ il est de 12 et nous avons $n_1 = 2$, $n_2 = 1$, $n_3 = 1$. Nous pouvons écrire

$$3 = \frac{3!}{1!2!} \text{ et } 12 = \frac{4!}{1!1!2!}. \tag{6.71}$$

En fait on peut constater que

$$\frac{n!}{n_1! n_2! \cdots n_6!}, \tag{6.72}$$

avec $n = \sum_{i=1}^{6} n_i$. On a donc que

$$p([n_1, n_2, ..., n_6]) = \frac{n!}{n_1! \cdots n_6!} p_1^{n_1} \cdots p_6^{n_6}. \tag{6.73}$$

De façon complètement générale ce genre de probabilité se calcule grâce à la *distribution multinomiale*

$$p([n_1, ..., n_k]) = \frac{n!}{n_1! \cdots n_k!} p_1^{n_1} \cdots p_k^{n_k}. \tag{6.74}$$

Exercice 35

On lance un dé parfait 10 fois. Quelle est la probabilité d'obtenir:

1. 10 fois 6?
2. 4 fois 3, 3 fois 2 et 3 fois 1?
3. 2 fois 1, 2 fois 2, 2 fois 3, 1 fois 4, 1 fois 5, et 1 fois 6?

6.3 Exemple du lotto

Dans un lotto on a dans une urne (souvent une machine spécialement conçue contenant de petites bales numérotées)
un nombre de jetons numérotés, disons pour l'exemple entre 1 et 6, qui sont tirés successivement. Une fois un jeton tiré, il ne sera pas remis dans le sac. On appelle ce genre de tirage *sans remise*. Contrairement au cas des dés vus dans la section précédente qui était '*avec remise*. On tire un nombre fixé de jetons, disons 3. On souhaite déterminer la probabilité d'obtenir une suite donnée de 2 numéros, disons 25. Disons aussi que pour cet exemple l'ordre du tirage a de l'importance (ce qui n'est pas le cas du lotto).

Afin de calculer cette probabilité le fait qu'on effectue un tirage avec remise est primordial. En effet considérons le cas initial illustré dans la fig. 6.6.

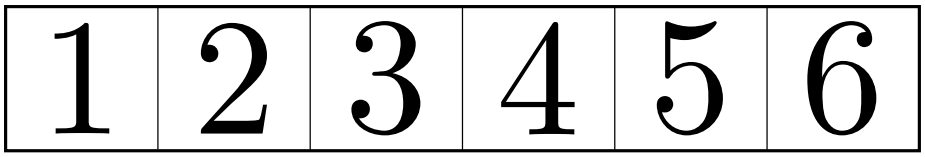

FIGURE 6.6 – Les six numéros présents initialement dans le sac.

Pendant le premier tirage, nous tirons le numéro 2 (voir figure fig. 6.7). Notons que le tirage du 2 a une probabilité $\frac{1}{6}$.

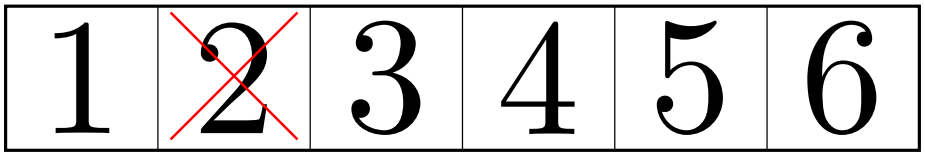

FIGURE 6.7 – Le numéro 2 est tiré lors du premier tirage.

Il est donc enlevé du sac et il nous reste uniquement 5 chiffres parmi lesquels choisir (les chiffres 1, 3, 4, 5, et 6, comme dans la fig. 6.8).

Comme il ne nous reste que 5 chiffres, la probabilité de tirer un des nombres restant, disons le 5, est de $\frac{1}{5}$ (voir la figure fig. 6.9).

Le 5 sera lui aussi retiré et il ne restera que 4 numéros dans le sac et ainsi de suite.

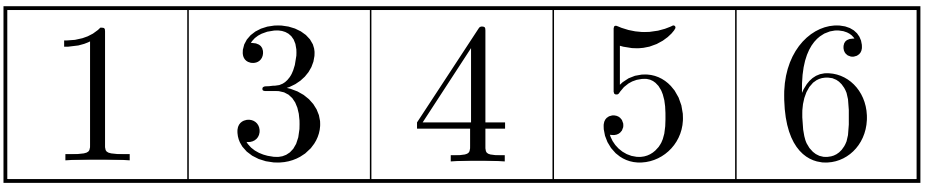

FIGURE 6.8 – Il ne reste que 5 chiffres dans le sac.

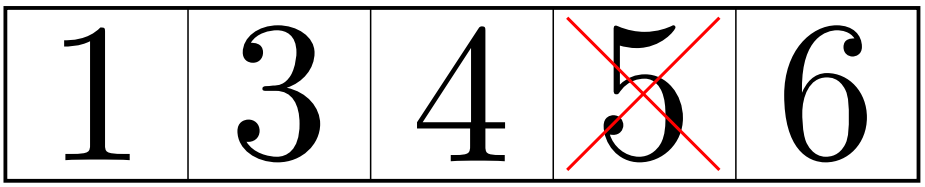

FIGURE 6.9 – Il ne reste que 5 chiffres dans le sac et nous tirons le 5.

On voit donc que la probabilité de tirer la suite ordonnée 25 est de

$$p(\{25\}) = p(\{2\}) \cdot p(\{5\}) = \frac{1}{6} \cdot \frac{1}{5} = \frac{1}{30}. \tag{6.75}$$

A présent, si nous considérons que l'ordre n'a pas d'importance, on a comme dans la section précédente que l'événement qui nous intéresse est $A = \{25, 52\}$. On peut donc décomposer ce cas en 2 et dire qu'on a dans un premier temps la probabilité de tirer 2 ou 5 parmi 6 nombres, puis on a la probabilité de tirer le 5 ou le 2 (respectivement si on a tiré 2 ou 5) parmi 5. Les deux probabilités sont donc données respectivement par $p(\{2, 5\}) = \frac{2}{6}$ puis par $p(\{5, 2\} \backslash \{2 \text{ ou } 5) = \frac{1}{5}$ pour trouver la probabilité $\frac{1}{15}$.

Exercice 36

1. Le jeu Euromillions consiste en un tirage de 5 numéros parmi 50 possible, puis par le tirage de 2 "étoiles" parmi 11 possibles. Déterminez la probabilité de trouver la bonne combinaison à un tirage.
2. Le jeu du swiss lotto, consiste au tirage de 6 numéros parmi 42 possibles, puis au tirage d'un numéros parmi 6. Calculez la probabilité de gagner au swiss lotto.

6.4 Quelques exercices

Afin de continuer avec ces concepts de tirages aléatoires avec ou sans remise de suites ordonnées ou non, nous allons faire quelques exercices. Il peut se révéler utile de dessiner un arbre pour ces exercices.

1. Dans une urne se trouvent 2 boules blanches et 3 boules noires. On tire successivement deux boules sans remise. Calculer et comparer les probabilités des deux événements suivants

— Tirer deux boules de même couleur.
— Tirer deux boules de couleurs différentes.

2. Une bille, lâchée en O tombe dans l'une des trois boîtes A, B, ou C. A chaque bifurcation, la bille tombe à gauche avec la probabilité de 0.25 et à droite avec la probabilité de 0.75 (voir fig. 6.10)

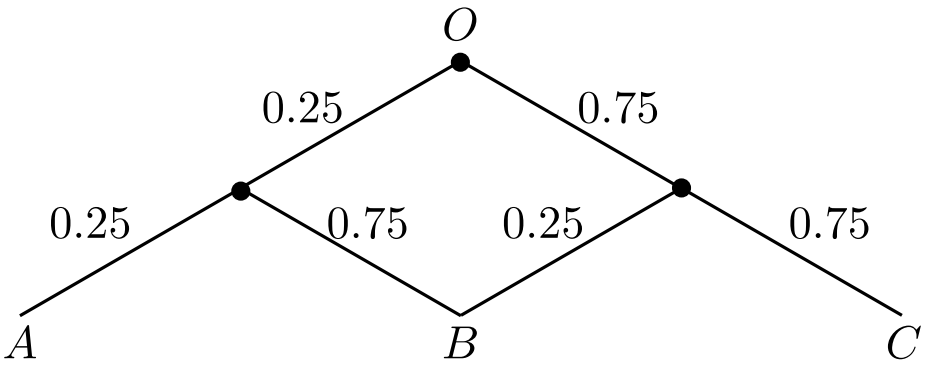

FIGURE 6.10 – Une bille lâchée en O tombe dans la boîte A, B, ou C.

— Calculer les probabilités $p(A)$, $p(B)$, $p(C)$ pour qu'une bille lâchée de O tombe respectivement dans la boîte A, B ou C.
— On lâche deux billes en O. Calculer la probabilité pour que les deux billes tombent dans la même boîte.
— On lâche trois billes en O. Calculer la probabilité d'avoir une bille dans chaque boîte.
— On lâche dix billes en O. Calculer la probabilité d'avoir au moins trois billes dans la boîte B.

3. A la naissance, la probabilité qu'un enfant soit un garçon est de $p(G) = 0.514$.
— Calculer et la probabilité qu'un enfant soit une fille.
— On considère la naissance de deux enfants. Calculer et la probabilité que les deux enfants soient de même sexe.
— On considère la naissance de deux enfants. Calculer et la probabilité que les deux enfants soient de sexes différents.

6.5 Variables aléatoires

Lors d'une expérience aléatoire, il est assez commun de relier chaque événement de l'univers, $A \in \Omega$, à un nombre réel, $X(A) \in \mathbb{R}$. Cette relation est définie par une fonction qui porte le nom de variable aléatoire et peut s'écrire mathématiquement sous la forme

$$X : \Omega \to \mathbb{R}. \tag{6.76}$$

Afin de mieux comprendre ce concept voyons quelques exemples

1. Lors d'un jet de dé unique l'univers est défini par $\Omega = \{1, 2, 3, 4, 5, 6\}$. On peut de façon assez naturelle définir notre variable aléatoire comme

$$X : i \to i. \tag{6.77}$$

2. Si nous lançons une pièce de monnaie les deux issues possibles sont pile p, ou face f ($\Omega = p, f$). Nous pouvons définir la variable aléatoire X comme

$$X : \begin{cases} p \to 0 \\ f \to 1 \end{cases} \tag{6.78}$$

3. Si nous lançons une pièce de monnaie à deux reprises, les issues possibles sont (p, p), (p, f), (f, p), (f, f). Nous pouvons définir la variable aléatoire X comme

$$X : \begin{cases} (p, p) \to 0 \\ (p, f) \to 1 \\ (f, p) \to 1 \\ (f, f) \to 2 \end{cases} \tag{6.79}$$

Comme nous nous sommes posés la question de connaître la probabilité d'obtenir un certain résultat lors d'une expérience aléatoire, il en va de même avec la probabilité que la variable aléatoire X prenne une valeur donnée, $\alpha \in \mathbb{R}$ ou prenne une valeur incluse dans un intervalle $I \subseteq \mathbb{R}$.

Pour illustrer ce qui se passe, intéressons-nous au dernier exemple ci-dessus avec le double pile ou face. On se pose les questions suivantes

1. Quelle est la probabilité que X prenne la valeur 1?
2. Quelle est la probabilité que X prenne une valeur incluse dans $I = [0.6, 3]$?
3. Quelle est la probabilité que X prenne une valeur inférieure à 2?

Prenons ces trois questions une par une

1. Les deux façons d'obtenir $X = 1$ est d'avoir les tirages (p, f) ou (f, p), soit $A = \{(p, f), (f, p)\}$. Les probabilités de chacun des événements de l'univers étant équiprobables on a

$$p(X = 1) = p(A) = 1/2. \tag{6.80}$$

2. Le seul événement donnant un X qui n'est pas dans l'intervalle $J = [0.6, 3]$ est $B = (p, p)$ ($X(B) = 0$). On a donc que

$$p(0.6 \leq X \leq 3) = p(\bar{B}) = 1 - p(B) = \frac{3}{4}. \tag{6.81}$$

3. De façon similaire les trois événements donnant $X < 2$ sont dans $C = \{(p, p), (p, f), (f, p)\}$. On a donc

$$p(X < 2) = p(C) = \frac{3}{4}. \tag{6.82}$$

On constate au travers de ces trois exemples que la probabilité que la variable aléatoire X prenne une valeur particulière α ou soit dans un intervalle I est reliée à la probabilité d'obtenir un événement D qui serait la préimage de α ou d'un intervalle I. On peut noter dans le cas général qu'on a $D = X^{-1}(I)$.

Définition 26 *(Variable aléatoire)*

On dit que la fonction $X : \Omega \to \mathbb{R}$ est une *variable aléatoire* si la préimage de X sur tout intervalle, $I \subseteq \mathbb{R}$, est un événement $A \in \Omega$. La probabilité que X prenne une valeur dans l'intervalle I est égale à la probabilité de réaliser l'événement A

$$p(X \in I) = p(A). \tag{6.83}$$

Définition 27 *(Fonction de répartition)*

On dit que la fonction $F : \mathbb{R} \to \mathbb{R}$ est une *fonction de répartition* si $F(x) = p(X \leq x)$ pour tout $x \in \mathbb{R}$.

Nous distinguons deux sortes de variables aléatoires: les variables aléatoires discrètes et continues. Nous les discuterons brièvement dans les deux sous-sections suivantes.

6.6 Nombres aléatoires

Les nombres aléatoires, bien que pas directement reliés aux probabilités, sont utilisés dans un certain nombre de domaines qui vont de la cryptographie aux simulations physiques. Nous allons voir une introduction simplifiée à la génération de nombres aléatoires sur un ordinateur et les différentes problématiques reliées à leur génération.

Une très bonne référence concernant les nombre aléatoires est le site `http://www.random.org`.

6.6.1 Générateurs algorithmiques: une introduction (très) générale

Le but des générateurs de nombres aléatoires est de produire une suite de nombres entiers, $(n \in \mathbb{N})$

$$\{X_0, X_1, ..., X_n\}, \tag{6.84}$$

avec $X_i \in A$, où $A = [0, m]$, avec $m \in \mathbb{N}$ (dans le cas de la fonction `rand()` de C, M est donné par la constante prédéfinie `RAND_MAX` qui and certains cas est $2^{31} - 1$). La probabilité de tirer chacun des nombres dans l'intervalle A est égale. On dit que la distribution des nombres est uniforme. De plus, les nombres tirés ne doivent pas dépendre de l'histoire des nombres tirés précédemment et on dit que les nombres sont indépendants.

Si on veut maintenant plutôt tirer des nombres réels uniformément distribués entre $[0, 1]$, il suffit de diviser les nombres X_i par m après chaque tirage. De façon similaire, si nous voulons tirer des nombres dans l'intervalle $[\alpha, \beta]$, on utilise la formule de remise à l'échelle suivante

$$N_i = \alpha + (\beta - \alpha)X_i/m. \tag{6.85}$$

Il faut remarquer que pour que cette formule puisse est utilisée il est nécessaire que $(\beta - \alpha) < M$.

Les transformations que je donne ici ne sont pas toujours celles implémentées. En effet, il existe des transformations beaucoup plus efficaces d'un point de vue computationnel pour changer l'intervalle des nombres aléatoires.

Sans entrer dans les détails, la génération de nombres aléatoires n'ayant pas une distribution uniforme s'obtient en effectuant une transformation un peu plus complexe que celle ci-dessus en partant toujours de la suite de nombres aléatoires entiers.

Les nombres aléatoires produits de façon algorithmique (donc avec un ordinateur) ne peuvent pas être vraiment aléatoires, car ils sont obtenus avec une machine déterministe (les opérations faites à l'aide d'un ordinateur sont par définition reproductibles avec une chance d'erreur quasiment nulle). On parle donc de nombre pseudo-aléatoires.

Néanmoins, bien que ces chiffres ne soient pas vraiment aléatoires, ils peuvent posséder des propriétés qui les rendent satisfaisants pour la plupart des applications. Cette suite de nombres doit avoir des propriétés particulières quand $m \to \infty$. Sans entrer pour le moment trop dans les détails, on veut par exemple que la moyenne des nombres tirés soit $m/2$, que la corrélation entre des sous-suites de nombres soit nulle, ou encore qu'il n'existe pas de séquence qui se répète (ou au moins que la période de répétition soit très très longue). Néanmoins, il est assez compliqué de définir des tests très robustes pour évaluer la qualité des nombres aléatoires algorithmiques.

6.6.2 Les générateurs congruenciels linéaires

Pendant très longtemps, les générateurs de nombres aléatoires algorithmiques ont été des générateurs congruenciels linéaires, dont la génération est donné par la formule suivante. Soit X_i un nombre aléatoire, alors le prochain nombre de la série est donné par

$$X_{i+1} = (aX_i + c) \mod m, \tag{6.86}$$

où a, c et m sont des paramètres de notre générateur. On constate que la seule partie éventuellement aléatoire de n'importe quelle séquence est la valeur initiale de notre séquence X_0 (aussi appelée *graine*). Tous les autres nombres obtenus sont déterministes. Pour chaque valeur de graine, on aura toujours la même séquence de nombre tirés.

Il est très important de noter que la qualité des nombres aléatoires obtenus sont extrêmement dépendants des valeurs de a, c et m choisies (et des relations entre elles). Si par exemple, on choisit $a = 1$, $c = 1$, $m = 10$ et $X_0 = 0$, on va avoir comme suite de nombre aléatoire

$$\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 0, 1, 2, 3, ...\}, \tag{6.87}$$

ce qui n'est pas très aléatoire vous en conviendrez... Il est donc très important de tenter d'optimiser les valeurs a, c et m pour avoir des séquences aussi "aléatoires" que possible.

Une première chose à remarquer c'est que m sera la valeur maximale de la période de notre générateur de nombre aléatoire (la période est le nombre de tirages qu'il faudra effectuer pour que la série se répète exactement).

Quelques paramètres utilisés dans des générateurs connus sont par exemple

— la fonction `rand()` du langage C

$$a = 1103515245, \quad c = 12345, \quad m = 2^{32}. \tag{6.88}$$

— la fonction `drand()` du langage C

$$a = 25214903917, \quad c = 11, \quad m = 2^{48}. \tag{6.89}$$

— le générateur `RANDU` des ordinateurs IBM des années 1960

$$a = 65539, \quad c = 0, \quad m = 2^{32}. \tag{6.90}$$

Ce genre de générateur de nombres aléatoires est très efficace d'un point de vue computationnel mais la qualité des nombres aléatoires est en général insuffisante. Plusieurs améliorations ont été proposées. Par exemple, pour chaque étape, on peut générer k nombres aléatoires avec un générateur congruentiel linéaire et combiner les nombres.

La méthode probablement la plus populaire consiste à utiliser des récurrences matricielles sur la représentation binaire des nombres. Soit $\tilde{X}_i$ la représentation sur k bits de X_i, alors $\tilde{X}_{i+1}$ est donné par

$$\tilde{X}_{i+1} = A\tilde{X}_i \mod 2, \tag{6.91}$$

où A est une matrice $k \times k$. Ce genre de générateur a l'énorme avantage d'être extrêmement efficace. Ils sont à la base de l'algorithme Mersenne Twister. Ces générateurs ont généralement une période extrêmement longue (qui a la particularité d'être un nombre premier de type Mersenne dont la forme est $m = 2^l - 1$, avec $l \in \mathbb{N}$).

Bien que ne soyant pas parfaits ces générateurs ont aussi le grand avantage d'être très rapides et peu gourmands en ressources de calcul. La facilité de description et d'utilisation de tels générateurs, permet des tests très poussés quant à leur qualités et leurs limites par la communauté scientifique. Finalement, les besoins de débogage de codes, la reproductibilité d'une série de nombres aléatoires peut être d'un grand secours.

6.6.3 Les générateurs physiques

Une autre façon de générer des nombres aléatoires, serait d'utiliser des phénomènes physiques qui contiennent de façon inhérente des processus aléatoires. On peut imaginer lancer un dé "à la main", mesurer les émissions radioactives d'atomes (mesurer leur spin), etc... Ou encore effectuer des lancer de jeux aussi peu biaisés que possibles (roulette, dé, etc).

Néanmoins, cette façon de faire a un certain nombre de désavantages. Le premier est que l'acquisition des données "en temps réel" de ces processus est en général plusieurs ordres de grandeurs trop lente par rapport aux besoins pratiques. Par

rapport à un générateur algorithmique très peu coûteux, un dispositif “physique” peut être très coûteux en espèces sonnantes et trébuchantes.

Il a néanmoins été envisagé de stocker de très grandes quantités de nombres aléatoires sur un support quelconque et de les fournir à l'utilisateur quand cela s'avère nécessaire. Le problème principal qui a été révélé par cette façon de faire est que le processus de mesure des différents processus est loin d'être parfait et engendre des biais importants dans la qualité des nombres obtenus ce qui les rend souvent en pratique moins bons que les nombres obtenus avec des générateurs de nombres pseudo-aléatoires...

6.6.4 Comment décider si une suite de nombres pseudo-aléatoires peut être considérée comme aléatoire

Cette question est extrêmement compliquée. Pour simplifier considérons le tirage de nombres entiers $X_i \in \{0, 1\}$. Les tirages aléatoires sont uniformément distribués, on a donc que $p(0) = p(1) = 1/2$. Supposons qu'on obtient une suite de 10 nombres avec deux générateurs différents

$$\begin{aligned} X &= \{0, 0, 1, 1, 1, 0, 1, 0, 1, 0\}, \\ Y &= \{0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}. \end{aligned} \tag{6.92}$$

On voit que la suite Y semble beaucoup moins aléatoire que la suite X. En effet, la probabilité de tirer 10 fois 0 en 10 tirages est de $p(Y) = 1/2^{10} = 1/1024$, alors que la probabilité d'avoir autant de 0 que de 1 est de $1/2$. De façon générale on aimerait que la répartition soit 35%-65% avec une probabilité de 90%.

Néanmoins, ce critère n'est pas suffisant. En effet la suite

$$Z = \{0, 1, 0, 1, 0, 1, 0, 1, 0, 1\}, \tag{6.93}$$

satisfait bien le critère ci-dessus. En revanche la probabilité de n'avoir pas deux tirages 0 ou 1 de suite est très faible (moins de 5%).

De ces constatations on peut dire qu'un générateur de nombres pseudo-aléatoires est de bonne qualité si les tirages qui sont effectués vérifient les propriétés du tirage avec une forte probabilité. On constate que cette définition est vague. En particulier la définition de “forte” est pas très précise. Il faut cependant noter que souvent nous sommes intéressés à des suites qui ont une longueur n. Donc pour $n \to \infty$ on va vouloir que les probabilités vont toutes tendre vers 1.

Néanmoins, il est certain qu'aucun générateur ne peut être parfait. En effet, les nombres étant toujours représentés avec une précision finie, il est impossible d'être capable de représenter exactement toutes les propriétés d'une série de nombres vraiment aléatoires avec un générateur pseudo-aléatoire. On va donc plutôt considérer une autre définition pour la qualité d'un générateur algorithmique.

Considérons une simulation nécessitant la génération de nombres aléatoires. Un “bon” générateur de nombres pseudo-aléatoire produit une série de nombres qui peut être utilisée en lieu et place de vrai nombres aléatoires sans que la simulation n'en soit affectée. Par exemple, le calcul du nombre π vu dans les exercices doit être trouvé avec la précision désirée avec le générateur de nombre pseudo-aléatoires pour que celui-ci soit considéré comme bon.

6.6.5 Quelques règles générales

La règle précédente bien que satisfaisante, n'est pas forcément simple à tester. En effet, il ne permet pas de prévoir la qualité d'un générateur a priori. Il nous faut donc quelques qualités minimales pour les générateurs de nombres aléatoires.

6.6.5.1 La périodicité

Tout générateur de nombres pseudo-aléatoires va à un moment ou un autre devenir périodique (la séquence de nombres générés vont se répéter à l'infini). Notons la période du générateur aléatoire T. Il est évident que dès qu'on atteint un nombre de tirages équivalent à la période ($\text{card}(X) \sim T$), on va avoir des nombres pseudo-aléatoires qui ne sont plus du tout satisfaisants. En fait on peut montrer que des problèmes apparaissent dès que le nombre de tirages atteint un nombre équivalent à $T^{1/3}$. Une condition primordiale pour avoir un "bon" générateur de nombres pseudo-aléatoire est donc une période élevée. Pour des générateurs aléatoires modernes, un période $T < 2^{100}$ n'est pas considérée comme satisfaisante pour la plupart des applications.

Évidemment il est impossible de tester la périodicité de tels générateurs de façon expérimentale ($2^{100} \sim 10^{30}$). Cela ne peut se faire que par des études analytiques approfondies. Comme expliqué dans la sec. 6.6.2 la période maximale d'un générateur congruentiel linéaire est m. Dans les 3 exemples donnés la période est respectivement de 2^{32}, 2^{48}, ou 2^{32}. Ils ne devraient donc plus être utilisés dans des applications modernes. A titre de comparaison le générateur Mersenne Twister possède une période de $2^{19937} - 1$.

Il est évident que la période à elle seule ne suffit pas à déterminer si un générateur de nombres pseudo-aléatoires est bon. En particulier on peut prendre un générateur congruentiel, où

$$X_{i+1} = (X_i + 1) \mod m, \tag{6.94}$$

avec m aussi grand qu'on veut (disons $m = 2^{2000}$ par exemple) mais la séquence de nombres générés ne sera absolument pas aléatoire, étant donné qu'on aura

$$X = \{0, 1, 2, 3, 4, 5, 6, ..., 2^{2000} - 1, 0, 1, 2, ...\}, \tag{6.95}$$

si $X_0 = 0$. Cela pourrait ne pas être problématique en soi, si la séquence avec une graine $X_0 = 1$ n'était pas si similaire

$$X = \{1, 2, 3, 4, 5, 6, ..., 2^{2000} - 1, 0, 1, 2, ...\}. \tag{6.96}$$

Il est donc nécessaire d'avoir d'autres critères que la seule période. C'est le sujet de la sous-section suivante.

6.6.5.2 La discrépance

Afin d'éliminer les générateurs de nombres pseudo-aléatoires comme l'exemple qu'on vient de citer, il faut étudier la répartition des nombres. Sans tomber dans le cas pathologique de la section précédente, on peut imaginer des nombres qui ont l'air aléatoires, mais qui ont un biais. Reprenons l'exemple du tirage entre $[0, 1]$. Nous pouvons imaginer une suite très longue sans période avec des tirages

aléatoires, mais avec beaucoup plus de 0 que de 1, ce qui évidemment serait problématique.

On doit donc trouver un moyen de tester la répartition des nombres de façon plus quantitative. Une façon de le faire est de considérer l'ensemble des $k-$uplets de nombres définis par

$$X^k = \{X_1, X_2, ..., X_k\}, \tag{6.97}$$

où X_0 est supposé tiré uniformément dans l'ensemble de départ (ici supposons que c'est $[0, 1]$ à titre d'exemple). En prenant toutes les graines existantes, on attend d'un bon générateur qu'il recouvre tout l'espace des résultats possibles pour les $k-$uplets formés avec des nombres aléatoires dans $[0, 1]^k$. En d'autres termes, il faut que des graines différentes génèrent des $k-$uplets différents pour toutes valeurs de k.

De nouveau ce genre de tests est très compliqué à tester expérimentalement pour k de l'ordre de la période du générateur de nombres aléatoires. Des analyses théoriques sont dès lors primordiales, mais bien en dehors du champs de ce cours...

Il existe beaucoup d'autres possibilités (il y a des recommandations sur le site `http://www.random.org`) pour tester des nombres aléatoires.

Chapitre 7

Remerciements

Je voudrais remercier (par ordre alphabétique) les étudiants du cours qui ont contribué à améliorer ce polycopié. En espérant que cette liste continuera à s'allonger avec les années. Merci à Messieurs Borel, Cirilli, El Kharroubi, Gay-Balmaz, Ibanez, Lovino, N'Hairi, Perret, Pin, Rod, Seemüller, Sousa, et Sutter. Je voudrais également remercier A. Malaspinas pour sa relecture et ses corrections.