MATHÉMATIQUES DE BASE (MIS 101, cours 2007-2008)

Alain Yger

30 mai 2012

Table des matières

Chapitre 1

Bases de logique et théorie des ensembles

1.1 Opérations logiques

1.1.1 Objets, assertions, relations

En mathématiques, on travaille sur des *objets* (on dit aussi *êtres*) entre lesquels on vérifie des *relations*.

On commence par définir les objets, en indiquant avec des illustrations :

- les nombres : $\mathbb{N}$ (quand on a du apprendre à compter), $\mathbb{Z}$ (quand il a fallu apprendre à calculer des gains, des pertes et des dettes depuis la fin du Moyen-âge) $\mathbb{Q}$, puis $\mathbb{R}$ (le nombre $\sqrt{2}$ sortant du cadre des fractions), puis $\mathbb{C}$ (pour résoudre $x^2 + 1 = 0$ et répondre aux exigences de la Physique du XIX-ème siècle), puis ensuite pour enrichir la boîte à outils nécessaire à l'algèbre, à la physique et à l'informatique, les quaternions, les octaves de Cayley...
- les objets géométriques : la sphère, la chambre à air (peut on dénouer un noeud sur une sphère, peut on le dénouer sur une chambre à air ?), le ruban ou le ruban de Mœbius (peut on s'orienter sur un ruban, le peut on sur un ruban de Mœbius ?), où la notion de forme (on dit *topologie*) joue un rôle majeur ;
- les transformations physiques (comme la transformation de Fourier, incarnation mathématique de la diffraction en optique) peuvent aussi être considérées comme des objets mathématiques.

Une *relation* est une assertion concernant divers objets mathématiques, assertion qui se doit d'être VRAIE ou FAUSSE. Par exemple :

$$n^2 \geq n \text{ quand } n \geq 1$$

est une assertion VRAIE, tandis que

$$n^2 < n \text{ quand } n > 1$$

est une assertion fausse (n est ici nombre entier positif). Par contre

$$i \leq 2i + 1$$

(i étant le nombre complexe correspondant à $(0, 1)$) n'est pas une relation car il n'y a pas d'ordre dans $\mathbb{C}$; ce n'est ni vrai, ni faux, mais simplement vide de sens.

1.1.2 Vrai et faux

La distinction VRAI/FAUX a beaucoup d'incarnations ; en voici deux :
- en informatique VRAI=1, FAUX=0, on associe donc à la relation un *bit*, élément de $\{0, 1\}$;
- dans les modélisations par des circuits électriques : VRAI= le courant passe (l'interrupteur est fermé), 0=le courant ne passe pas (l'interrupteur est ouvert).

Une assertion VRAIE est une assertion qui peut se déduire d'un petit nombre d'axiomes traduisant les propriétés invariantes des objets.

Les axiomes, eux, ne se démontrent pas : ils fondent le cadre dans lequel on se place ; changer le système d'axiomes, c'est changer la règle du jeu.

Exemple d'axiome : Le postulat d'Euclide, qu'Euclide formulait ainsi :

"Et si une droite tombant sur deux droites fait les angles intérieurs du même côté plus petits que deux droits, ces deux droites, prolongées à l'infini, se rencontreront du côté où les angles sont plus petits que deux droits"

(à vous de le transcrire dans le plan en faisant un dessin sur une feuille) ne se démontre pas ; il fonde la géométrie Euclidienne. Par contre, on peut montrer (faites l'exercice) que l'énoncer revient à dire :

"Par un point extérieur à une droite passe une unique droite qui lui est parallèle"

(Playfair a prouvé en 1795 que c'était bien dire la même chose, ce que l'on pourra s'entraîner à vérifier).

On peut refuser le postulat d'Euclide et faire par exemple dans le plan de la *géométrie sphérique* ; le plan est pensé comme le globe terrestre privé du pôle Nord *via* la correspondance qui associe sur la figure ci-dessous le point M du globe au point m du plan équatorial.

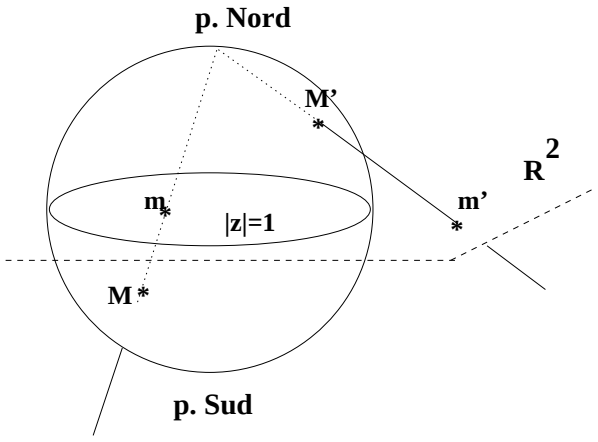

FIGURE 1.1 – Le globe terrestre et la projection stéréographique depuis le pôle Nord

Les droites du plan correspondent aux cercles tracés sur le globe et passant par le pôle Nord : il n'y a plus de droites parallèles !

Établir à partir d'un jeu d'axiomes qu'une assertion est VRAIE, c'est prouver un *théorème*. Si ceci est fait de manière intermédiaire dans le but de prouver ultérieurement qu'une autre assertion est VRAIE, on parle plutôt de *lemme*. Lorsque l'on déduit d'un théorème qu'une assertion est VRAIE, on prouve un *corollaire* de ce théorème.

1.1.3 Quelques opérations entre assertions

On va définir des opérations entre deux assertions R et S en donnant les *tables de vérité* de ces nouvelles assertions, c'est-à-dire les tables qui permettent de décider, suivant que R et S sont vraies ou fausses, si la nouvelle assertion est vraie ou fausse.

Définition 1.1.

1. La *disjonction* **[ou]** logique $R \vee S$.

R	S	R v S
0	0	**0**
0	1	**1**
1	0	**1**
1	1	**1**

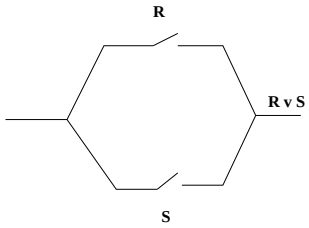

FIGURE 1.2 – La disjonction

La disjonction correspond au montage en parallèle dans les circuits électriques (R et S étant pensés comme des interrupteurs).

1. La *conjonction* **[et]** logique $R \wedge S$.

R	S	R ∧ S
0	0	**0**
0	1	**0**
1	0	**0**
1	1	**1**

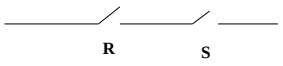

FIGURE 1.3 – La conjonction

La conjonction correspond au montage en série dans les circuits électriques (R et S étant pensés comme des interrupteurs).

3. L'*implication* R implique S (notée $R \Longrightarrow S$) :

R	S	R ⇒ S
0	0	**1**
0	1	**1**
1	0	**0**
1	1	**1**

FIGURE 1.4 – L'implication

Ce qui est un peu inattendu dans la définition de l'implication est que $R \Longrightarrow S$ est toujours vraie dans tous les cas où R est fausse ; on considère (ce qui n'est pas absurde) qu'une assertion fausse implique

n'importe quoi ! Par exemple les assertions

$$1 = 2 \Longrightarrow 2 = 3 \qquad 1 = 2 \Longrightarrow 2 = 2$$

sont toutes les deux vraies.

4. L'équivalence $R \Longleftrightarrow S$ est la conjonction $(R \Longrightarrow S) \wedge (S \Longrightarrow R)$; en voici la table de vérité :

R	S	R $\Longleftrightarrow$ S
0	0	**1**
0	1	**0**
1	0	**0**
1	1	**1**

FIGURE 1.5 – L'équivalence

5. La *négation* non R consiste à inverser les bits 0 et 1. Voici la table :

R	nonR
0	**1**
1	**0**

FIGURE 1.6 – La négation

1.1.4 Règles de logique

Étant données des assertions R, S, T,..., on peut former de nouvelles assertions en les utilisant couplées avec les opérations précédemment introduites (la disjonction, la conjonction, l'implication, l'équivalence, la négation). On obtient ainsi de nouvelles assertions, expressions parfois compliquées mêlant les assertions $R, S, T, ...$ et les opérations, par exemple

$$(R \Longrightarrow S) \Longrightarrow \Big((R \vee T) \Longrightarrow (S \vee T)\Big) ;$$

attention ! la position des parenthèses est très importante pour la lecture des expressions (on doit toujours veiller à ne faire une opération logique qu'entre deux assertions).

La table de vérité d'une telle assertion composite $\mathcal{R}(R, S, T, ...)$ s'écrit en étudiant cas par cas les diverses possibilités (R vraie ou fausse, S vraie ou fausse, T vraie ou fausse, *etc.*) et en utilisant les tables de vérité des diverses opérations. Si par hasard la table de vérité de $\mathcal{R}(R, S, T, ...)$ s'écrit comme une colonne de 1, on dit que l'assertion $\mathcal{R}(R, S, T, ...)$ est *inconditionnellement vraie* (ou encore que c'est une *évidence*), c'est-à-dire qu'elle est vraie quelque soient les valeurs (0 ou 1) prises par les assertions R, S, T,..., et on appelle alors $\mathcal{R}(R, S, T, ...)$ *règle logique*. Ces règles logiques seront utilisées dans les raisonnements conduisant à des lemmes, théorèmes, corollaires.

Exemples. On vérifiera en exercice que les assertions composites suivantes sont inconditionnellement vraies et deviennent donc des règles logiques :

$$\begin{array}{rcl}
(R \vee R) & \Longrightarrow & R \\
R & \Longrightarrow & R \\
R & \vee & (\text{non}\, R) \\
R & \Longleftrightarrow & (\text{non}\, (\text{non}\, R)) \\
R & \Longrightarrow & (R \vee S) \\
(S \vee S) & \Longrightarrow & (S \vee R) \\
(R \vee S) & \Longrightarrow & (S \vee R) \\
(R \wedge S) & \Longrightarrow & (S \wedge R) \\
(R \vee S) & \Longleftrightarrow & (S \vee R) \\
(R \wedge S) & \Longleftrightarrow & (S \wedge R) \\
(R \Longrightarrow S) & \Longrightarrow & \Big((R \vee T) \Longrightarrow (S \vee T)\Big)
\end{array}$$

Certaines de ces assertions inconditionnellement vraies jouent un rôle très important dans les raisonnements logiques ; en voici deux exemples :

1. La règle de *contraposition* que l'on retrouvera plus loin comme moteur de certains raisonnements :

$$(R \Longrightarrow S) \Longleftrightarrow \Big((\text{non}\, S) \Longrightarrow (\text{non}\, R)\Big)$$

(à vérifier avec les tables de vérité).

2. La règle de *transitivité*, elle aussi importante, car elle permet le cheminement logique :

$$\Big[\Big((R \Longrightarrow S) \text{ VRAIE}\Big) \wedge \Big((S \Longrightarrow T) \text{ VRAIE}\Big)\Big] \Longrightarrow \Big((R \Longrightarrow T) \text{ VRAIE}\Big)\,.$$

1.2 Ensembles et parties d'un ensemble ; quantificateurs

Definition 1.2. Un *ensemble* E est par définition une collection d'objets, dits *éléments* de E.

Par exemple

$$\mathbb{N} := \{0, 1, 2, ...\} \qquad \mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$$

sont des ensembles (ils ont d'ailleurs tous les deux la particularité que l'on peut en numéroter les éléments). C'est aussi le cas de $\mathbb{Q}$ (on le verra plus tard). Les nombres réels forment aussi un ensemble (noté $\mathbb{R}$), mais dont on ne peut cette fois numéroter les éléments.

On s'aidera pour raisonner sur les ensembles de schémas où ces ensembles seront représentés sous forme de patatoïdes.

On notera $x \in E$ l'assertion

$$<< x \text{ est un élément de } E >>$$

Supposons maintenant que R soit une assertion dans laquelle figure un symbole matérialisé par une lettre x (par exemple $x \geq 3$ ou $x \in \mathbb{R}$, ou encore $(x \in [1, +\infty[) \Longrightarrow (x \geq 0)$). On note (pour tenir compte de la présence de ce caractère x) l'assertion R en l'écrivant plutôt $R\{x\}$.

Si E est un ensemble et $R\{x\}$ une telle assertion, on définit deux nouvelles assertions que l'on note ainsi :

$$\begin{array}{ll} \forall\, x \in E & R\{x\} \\ \exists\, x \in E & R\{x\} \end{array}$$

La première se lit :

$$<< \text{ pour tout élément } x \text{ de } E,\ R\{x\} >> ;$$

elle est vraie si pour tout élément x de E, l'assertion $R\{x\}$ est vraie ; elle est fausse sinon.
La seconde se lit :

$$<< \text{ il existe un élément } x \text{ de } E\,,\ R\{x\} >> ;$$

elle est vraie s'il existe au moins un élément x de E tel que l'assertion $R\{x\}$ soit vraie ; elle est fausse sinon.

On a donc la règle (que l'on vérifie en introduisant par exemple le sous-ensemble A de E constitué des éléments x pour lesquels $R(x)$ est vraie et son complémentaire)

$$\Big(\text{non}\Big(\forall\, x \in E\,,\ R\{x\}\Big)\Big) \Longleftrightarrow \Big(\exists\, x \in E\,,\ \text{non}\,(R\{x\})\Big)\,.$$

On peut également introduire des assertions dans lesquelles figurent plusieurs symboles mathématiques, par exemple deux symboles x et y ; on notera une telle assertion $R\{x, y\}$ (par exemple "$x + y \geq 1$" ou "x divise y"). Si E et F sont deux ensembles, on peut alors introduire les six assertions :

$$\begin{array}{ll} \forall x \in E\,, \forall y \in F\,, & R\{x, y\} \\ \forall x \in E\,, \exists y \in F\,, & R\{x, y\} \\ \exists x \in E\,, \forall y \in F\,, & R\{x, y\} \\ \exists x \in E\,, \exists y \in F\,, & R\{x, y\} \\ \forall y \in F\,, \exists x \in E\,, & R\{x, y\} \\ \exists y \in F\,, \forall x \in E\,, & R\{x, y\} \end{array}$$

On vérifiera facilement les règles suivantes :

$$\begin{array}{rcl} \text{non}\Big(\forall x \in E\,, \forall y \in F\,,\ R\{x, y\}\Big) & \Longleftrightarrow & \Big(\exists x \in E\,, \exists y \in F\,,\ \text{non}\, R\{x, y\}\Big) \\ \text{non}\Big(\forall x \in E\,, \exists y \in F\,,\ R\{x, y\}\Big) & \Longleftrightarrow & \Big(\exists x \in E\,, \forall y \in F\,,\ \text{non}\, R\{x, y\}\Big) \\ \text{non}\,\Big(\exists x \in E\,, \forall y \in F\,,\ R\{x, y\}\Big) & \Longleftrightarrow & \Big(\forall x \in E\,, \exists y \in F\,,\ \text{non}\, R\{x, y\}\Big) \\ \text{non}\,\Big(\exists x \in E\,, \exists y \in F\,,\ R\{x, y\}\Big) & \Longleftrightarrow & \Big(\forall x \in E\,, \forall y \in F\,,\ \text{non}\, R\{x, y\}\Big) \\ \Big(\exists x \in E\,, \forall y \in F\,,\ R\{x, y\}\Big) & \Longrightarrow & \Big(\forall y \in F\,, \exists x \in E\,,\ R\{x, y\}\Big) \end{array}$$

Les symboles $\forall$ et $\exists$ (dont on remarque qu'ils doivent être échangés lorsque l'on prend la négation d'une assertion) sont appelés *quantificateurs*. L'ordre dans lequel on les écrit est important : par exemple "$\forall x \in E,\ ,\ \exists\, y \in F\,, ...$" signifie littéralement : "pour tout x de E, il existe y dans F, dépendant *a priori* de x,..." alors que " $\exists y \in F\,,\ \forall x \in E\,, ...$" signifie "il existe y dans F tel que, pour tout x dans E, ...".

Définition 1.2. Si E est un ensemble, on appelle *sous-ensemble* de E toute sous-famille composée d'éléments de E (donc extraite de E) ; on dit aussi qu'un sous-ensemble de E est une *partie* de E.

On convient que la partie n'ayant aucun élément, que l'on appelle l'*ensemble vide* et que l'on note $\emptyset$, est un sous-ensemble particulier de E. C'est donc un sous-ensemble de tous les ensembles.

Définition 1.3. Si E est un ensemble et A et B deux parties de E, on note $A \subset B$ (A *inclus* dans B) l'assertion "$\forall x \in A\,, x \in B$".

Pour toute partie A de E, les assertions $A \subset E$ et $\emptyset \subset A$ sont donc vraies.

Deux parties A et B sont dites *égales* (on note $A = B$) si $A \subset B$ et $B \subset A$, c'est-à-dire si A et B ont exactement les mêmes éléments.

On définit maintenant deux opérations importantes entre les sous-ensembles d'un même ensemble E.

Définition 1.4. Si A et B sont deux parties de E, on appelle *union* de A et B et on note $A \cup B$ la partie de E dont les éléments appartiennent à A ou à B. On appelle *intersection* de A et B (et on note $A \cap B$) la partie de E dont les éléments appartiennent à A et à B.

Deux parties A et B sont dites *disjointes* si $A \cap B = \emptyset$, c'est-à-dire si A et B n'ont aucun élément en commun. Si une partie A s'écrit $A = A_1 \cup A_2$ avec $A_1 \cap A_2 = \emptyset$, on dit que A_1 et A_2 réalisent une *partition* de A.

On a $A \cup B = B \cup A$ et $A \cap B = B \cap A$. Les opérations $\cup$ et $\cap$ sont dites *commutatives*. De plus $A \cup \emptyset = \emptyset \cup A$ pour toute partie A de E, ce qui fait dire que $\emptyset$ est *élément neutre* pour l'opération $\cup$. Enfin $A \cap \emptyset = \emptyset$ pour toute partie A de E, ce qui fait dire que $\emptyset$ est cette fois *élément absorbant* pour l'opération $\cap$.

Définition 1.5. Si A est une partie de E, on appelle *complémentaire* de A dans E et on note $C_E A$ ou encore $E \setminus A$ (ou aussi A^c lorsqu'il est implicite que l'on travaille dans un ensemble E) le sous-ensemble constitué des éléments de E n'appartenant pas à A.

Par exemple, le complémentaire de $\{1, 2, 3, 4, 5\}$ dans $\mathbb{N}$ est $\{n \in \mathbb{N} \text{ t.q. } n \geq 6\}$.

On vérifie les deux égalités :

$$\begin{aligned} (A \cap B)^c &= A^c \cup B^c \\ (A \cup B)^c &= A^c \cap B^c\,. \end{aligned}$$

Définition 1.6. Si A et B sont deux parties de E, on note $A \setminus B$ l'ensemble des éléments de A qui n'appartiennent pas à B et $A \Delta B$ (*différence symétrique* de A et B) l'ensemble

$$A \Delta B := (A \setminus B) \cup (B \setminus A)\,.$$

On vérifiera que

$$A \cup B = (A \cap B) \cup (A \Delta B)$$

et que les deux parties $A \cap B$ et $A \Delta B$ réalisent une partition de $A \cup B$.

La preuve de ces diverses assertions sera illustrée par des diagrammes où les parties seront représentées par des patatoïdes ; faites les dessins !

Dernières règles importantes à retenir :

$$(A \cup B) \cup C = A \cup (B \cup C) \qquad (A \cap B) \cap C = A \cap (B \cap C)$$

(on dit que les deux opérations $\cup$ et $\cap$ sont *associatives*) et

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

(on dit que l'opération $\cap$ est *distributive* par rapport à l'opération $\cup$). Les deux opérations $\cup$ et $\cap$ semblent se comporter respectivement comme l'addition et la multiplication des nombres réels ; d'ailleurs $\emptyset$ joue exactement le même rôle que 0 (neutre pour la première opération, absorbant pour la seconde). Cette analogie sera intéressante plus tard.

1.3 Quelques mots de l'axiomatique de la théorie des ensembles

L'univers peut être symbolisé par un nuage de "bulles", chaque "bulle" représentant un ensemble, ces "bulles" étant reliés par des flèches suivant la règle suivante : il y a une flèche allant de la "bulle" E à la "bulle" F si et seulement si E est un élément de F (attention ! il ne faut pas confondre avec "E est inclus dans F" !)

La théorie des ensembles (formalisée entre 1920-1925 par les mathématiciens allemands F. Zermelo, 1871-1953 et A. Frænkel, 1891-1965) présuppose quatre axiomes majeurs :

1. Deux ensembles de l'univers ayant les mêmes éléments sont égaux (*axiome d'extensionnalité*)

2. Étant donnés deux ensembles de l'univers, il en existe un troisième (à exactement deux éléments) qui a pour uniques éléments les deux ensembles précédents (*axiome de la paire*).

3. À tout ensemble E de l'univers, on peut associer un autre ensemble dont les éléments sont exactement les éléments des éléments de E ; ce nouvel ensemble F est donc la version "explosée" de la "bulle" E, toutes les "bulles" éléments de E ayant éclaté pour libérer tous leurs éléments dans F(*axiome de la somme*)

4. Enfin, c'est l'axiome le plus important pour nous : à tout ensemble E de l'univers, on peut associer un autre ensemble de l'univers dont les éléments sont exactement les parties de E ; ce nouvel ensemble est noté $\mathcal{P}(E)$ (*axiome des parties*).

On dit qu'un ensemble E est fini s'il est composé d'un nombre fini d'éléments, que l'on appelle le *cardinal* de E. On dit qu'un ensemble E est *dénombrable* si on peut en numéroter les éléments

$$E = \{x_0, x_1, x_2, ...\},$$

les x_j étant des éléments distincts ; l'ensemble des nombres entiers positifs $\mathbb{N}$ est le prototype d'ensemble dénombrable ; de même, l'ensemble $\mathbb{Z}$ est dénombrable (voici un moyen de numéroter les éléments de $\mathbb{Z}$: $x_0 = 0$, $x_1 = 1$, $x_2 = -1$, $x_3 = 2$, $x_4 = -2$, $x_5 = 3$, $x_6 = -3$, *etc.*). Il est un peu plus difficile de montrer que $\mathbb{Q}$ est dénombrable (on le montrera plus tard dans ce cours) mais l'on verra plus loin que $\mathbb{R}$, lui, ne l'est pas (ni $\mathbb{C}$, bien sûr).

Si l'on a un ensemble fini E, il est évident que l'on dispose d'une stratégie pour choisir un élément de E ; c'est aussi vrai si E est dénombrable (il suffit de prendre x_0).

En revanche, pareille opération de *choix* pose problème en théorie des ensembles. Au quatre axiomes, s'en ajoute un (qui ne s'en déduit pas) qui est aussi fondamental, *l'axiome du choix*, axiome que l'on peut énoncer ainsi :

«*Étant donnée une collection d'ensembles non vides de l'univers n'ayant deux à deux aucun élément commun, on peut construire un nouvel ensemble en prenant un élément dans chacun des ensembles de la collection*»

Admettre l'axiome du choix conduit à des paradoxes comme celui de Banach-Tarski (apparu en 1923) : on peut faire un puzzle avec une boule et reconstituer avec les morceaux du puzzle deux boules de même volume que la boule initiale !

On convient cependant communément aujourd'hui de raisonner en admettant l'axiome du choix.

Il s'avère aussi que les quatre axiomes de la théorie des ensembles proposés ci-dessus (avant l'axiome du choix) n'excluent nullement que des ensembles puissent s'auto-appartenir, situation qui ne correspond pas à l'idée intuitive d'appartenance à un ensemble. On évite ce phénomène en ajoutant aux quatre axiomes proposés plus haut (en fait cinq avec l'axiome du choix) un sixième axiome, dit *axiome de fondation* :

« *Tout ensemble non vide contient un élément avec lequel il n'a aucun élément en commun* »

On verra un peu plus loin pourquoi cet axiome exclut qu'il existe un ensemble E qui s'auto-appartienne, donc tel que E soit élément de E (attention ! "élément de" et non "partie de" !). C'est donc cet axiome qui exclut que l'ensemble de tous les ensembles puisse être un ensemble.

1.4 Produit de deux ensembles

Supposons maintenant que E et F soient deux ensembles.

Définition 1.6. On définit un nouvel ensemble noté $E \times F$ (produit de E par F) en considérant l'ensemble des couples (x, y) avec $x \in E$ et $y \in F$.

Attention à ne pas confondre le couple (x, y) (il y a un ordre, on prend l'élément x d'abord, puis y en second) et l'ensemble à deux éléments $\{x, y\}$ (où les deux éléments sont mis dans un même sac sans que l'un soit le premier, l'autre le second) !

Les ensembles $E \times F$ et $F \times E$ sont deux ensembles distincts si E et F n'ont pas les mêmes éléments (ceci résulte de l'axiome d'extensionnalité). Ce sont aussi deux ensembles distincts de l'ensemble

$$H := \{\{x, y\}\,;\, x \in E\,, y \in F\}$$

des paires (indifférenciées) constituées d'un élément de E et d'un élément de F pris "en vrac".

On peut refaire le produit de $E \times F$ avec un troisième ensemble G, *etc.* On définit ainsi le plan $\mathbb{R}^2$, l'espace $\mathbb{R}^3$, l'espace-temps $\mathbb{R}^4$, *etc.*

Si l'on a trois ensembles E, F, G et si l'on dispose de trois éléments $x \in E$, $y \in F$, $z \in G$, on décide d'identifier le triplet (x, y, z) et la paire $((x, y), z)$. Avec cette identification, on convient que

$$(E \times F) \times G = E \times F \times G = E \times (F \times G)\,.$$

L'opération de produit d'ensembles est associative.

1.5 Union et intersection d'une famille de parties d'un même ensemble

Définition 1.7. Si E est un ensemble et $\mathcal{A}$ une partie de $\mathcal{P}(E)$, on appelle *union des éléments de* $\mathcal{A}$ et on note

$$\bigcup_{A \in \mathcal{A}} A$$

l'ensemble des points de E qui appartiennent au moins à l'une des parties A élément de $\mathcal{A}$.

Définition 1.8. Si E est un ensemble et $\mathcal{A}$ une partie de $\mathcal{P}(E)$, on appelle *intersection des éléments de* $\mathcal{A}$ et on note

$$\bigcap_{A \in \mathcal{A}} A$$

l'ensemble des points de E qui appartiennent à toutes les parties A éléments de $\mathcal{A}$.

1.6 Apprendre à raisonner : le raisonnement par l'absurde

Il y a deux trois modèles de raisonnement fréquemment utilisés dans une déduction logique : le raisonnement *par l'absurde*, celui *récurrence*, celui enfin *par contraposition* (que l'on utilise parfois alors que ce n'est pas indispensable, il faut se méfier des "faux" raisonnements par contraposition qui sont en fait des raisonnements directs déguisés !)

On présente d'abord celui qui demande le moins de matériel, le raisonnement par l'absurde.

LE PRINCIPE : *Supposons que l'on travaille sous une certaine axiomatique et que l'on veuille prouver qu'une certaine assertion R est VRAIE. On suppose R fausse puis on exhibe (en utilisant notre système d'axiomes et les règles de déduction logique) une nouvelle assertion S dont on peut montrer à la fois qu'elle est VRAIE et FAUSSE (sous l'hypothèse que R est FAUSSE). Notre raisonnement fondé sur l'hypothèse «R fausse» conduit donc à une contradiction ; on en conclut que l'hypothèse faite sur R était fausse et par conséquent que l'assertion R est VRAIE.*

Nous allons donner un exemple emprunté à la théorie des ensembles.

Supposons que nous travaillons avec les quatre axiomes de Zermelo-Frænkel (dégagés dans la sectin 1.3) mais que nous rajoutions à notre axiomatique l'*axiome de fondation* (vu en fin de section 1.3).

Un exemple de raisonnement par l'absurde. Sous ces six axiomes de la théorie des ensembles, montrons par l'absurde qu'il n'existe aucun ensemble qui puisse s'auto-appartenir ($E \in E$). Raisonnons par l'absurde et supposons qu'un ensemble E s'auto-appartienne, c'est-à-dire que $E \in E$. Prenons $F = \{E\}$ (singleton constitué du seul élément E). L'axiome de fondation assure que $\{E\}$ et E n'ont aucun élément commun (E étant le seul élément de $\{E\}$). Or, $E \in E$ et $E \in \{E\}$, ce qui prouve que E est un élément de E et de $\{E\}$. L'assertion «E et $\{E\}$ n'ont aucun élément commun» est donc VRAIE et FAUSSE, ce qui est contradictoire. On prouve ainsi par l'absurde qu'aucun ensemble ne s'auto-appartient.

L'ensemble de tous les ensembles (c'est-à-dire l'univers) ne peut donc être un ensemble (car il s'auto-appartiendrait) !

1.7 Apprendre à raisonner : le principe de contraposition

PRINCIPE : *Soient R et S deux assertions. Pour prouver que l'assertion $R \Longrightarrow S$ est VRAIE, on utilise l'évidence de contraposition*

$$(R \Longrightarrow S) \Longleftrightarrow (\text{non}\, S \Longrightarrow \text{non}\, R)$$

qui nous assure que $R \Longrightarrow S$ est VRAIE si est seulement si

$$\text{non}\, S \Longrightarrow \text{non}\, R$$

est vraie. Prouver que $R \Longrightarrow S$ est VRAIE revient donc à prouver que $(\text{non}\, S) \Longrightarrow (\text{non}\, R)$ *est VRAIE.*

Exemple : Soit $x = p/q$ un nombre réel rationnel non nul ($pq \neq 0$) ; alors

$$(y \notin \mathbb{Q}) \Longrightarrow (xy \notin \mathbb{Q})$$

est VRAIE ; en effet, si $xy = a/b \in \mathbb{Q}$ (avec $b \neq 0$), on a $y = q/p \times a/b \in \mathbb{Q}$, ce qui prouve que $(xy \in \mathbb{Q}) \Longrightarrow (y \in \mathbb{Q})$.

1.8 Compter, calculer, ordonner, raisonner par récurrence

On connaît l'idée du raisonnement par récurrence : pour prouver par exemple que l'assertion

$$R\{n\} \ : \ 0+1+\cdots+n^2=\sum_{k=1}^{n}k^2=\frac{n(n+1)(2n+1)}{6}$$

pour tout $n\in\mathbb{N}$, il suffit de montrer que $R\{0\}$ est vraie (ce qui est immédiat ici car $0=0$) puis de montrer que, si $R\{n\}$ est vraie, $R\{n+1\}$ l'est aussi ; or, ici

$$\begin{aligned}\sum_{k=0}^{n+1}k^2 &= \Big(\sum_{k=0}^{n}k^2\Big)+(n+1)^2=\frac{n(n+1)(2n+1)}{6}+(n+1)^2\\ &= (n+1)\Big(\frac{n(2n+1)}{6}+(n+1)\Big)=\frac{(n+1)(n+2)(2(n+1)+1)}{6}\end{aligned}$$

et l'on voit donc que ($R\{n\}$ vraie) $\Longrightarrow$ ($R\{n+1\}$ vraie), ce qui prouve bien que $R\{n\}$ est vraie pour ou entier $n\in\mathbb{N}$. Le principe sur lequel est fondé le raisonnement par récurrence nous oblige à nous ressourcer aux axiomatiques qui régissent la construction de l'ensemble des nombres entiers positifs $\mathbb{N}$, ensemble essentiel aux mathématiques discrètes (celles dont se nourrit l'informatique). C'est ce qui justifie ici dans ce cours la digression qui suit sur l'ensemble $\mathbb{N}$ des nombres entiers positifs, cadre de l'*arithmétique*.

1.8.1 L'axiomatique de $\mathbb{N}$

Le logicien et mathématicien italien Guiseppe Peano (1858-1932) proposa 5 axiomes régissant la construction des entiers naturels positifs (l'ensemble $\mathbb{N}$) ; de fait, le mathématicien allemand Dedekind fut le premier à les formuler en 1888.

1. $\mathbb{N}$ contient au moins un élément noté 0 ;
2. Tout élément de $\mathbb{N}$ admet un successeur $S(n)$;
3. Si deux éléments de $\mathbb{N}$ ont des successeurs égaux, ils sont égaux ;
4. 0 n'est successeur d'aucun élément ;
5. Si A est un sous-ensemble de $\mathbb{N}$ tel que $0\in A$ et que $(\forall x\in A\,,S(x)\in A)$ soit vraie, alors $A=\mathbb{N}$ (*principe de récurrence*).

C'est sur l'axiome 5 que se fonde le principe du *raisonnement par récurrence* sur lequel nous allons revenir.

1.8.2 Deux opérations sur $\mathbb{N}$

Les axiomes de Peano permettent de définir une règle de dénombrement ; les entiers sont numérotés à partir de 0 suivant la règle de succession : 0, $1=S(0)$, $2=S(1)$, $3=S(2)$,.... (le principe de récurrence nous assurant que l'on liste ainsi tous les entiers) et l'on peut donc, si k est un entier donner un sens à la phrase :

« On peut répéter k-fois une instruction »

On convient que «répéter l'instruction 0 fois» revient à ne pas l'exécuter.

Ceci nous permet de définir deux opérations de manière algorithmique, l'*addition* et la *multiplication* des entiers.

Voici la suite d'instructions permettant de définir l'addition des entiers a et b :

```
somme=a;
repeter b fois
    somme=S(somme);
fin
```

L'addition de deux entiers a et b est notée $a+b$; on remarque que $a+1=S(a)$.

On définit de même la multiplication de a par b suivant la suite d'instructions suivante :

```
produit=0;
repeter a fois
   produit=produit + b;
fin
```

On vérifiera que l'addition et la multiplication sont deux opérations sur $\mathbb{N}$ partageant certaines propriétés qu'ont (isolément ou entre elles) l'union et l'intersection sur l'ensemble des parties d'un ensemble E, à savoir :

- les deux opérations sont commutatives ;
- les deux opérations sont associatives (on réécrira ce que cela veut dire) ;
- 0 est élément neutre pour l'addition (comme $\emptyset$ pour l'union) ;
- 0 est élément absorbant pour la multiplication (comme $\emptyset$ pour l'intersection) ;
- la multiplication est distributive par rapport à l'addition (comme l'opération d'intersection l'est par rapport à l'union) ;

On dit que l'ensemble $\mathbb{N}$ équipé de l'addition est un *monoïde*.

1.8.3 Un ordre total sur $\mathbb{N}$

On définit un *ordre* sur $\mathbb{N}$ de la manière suivante :

$$(a \leq b) \Longleftrightarrow (\exists x \in \mathbb{N} \text{ tel que } b = a + x)$$

Si $a \leq b$, on a $a+c \leq b+c$ et $ac \leq bc$ pour tout c dans $\mathbb{N}$ et l'on dit que l'ordre ainsi défini est *compatible* avec l'addition et la multiplication. On lit $a \leq b$ aussi «a est plus petit que b».

On a les trois propriétés très importantes suivantes :

A1. Toute partie non vide A de $\mathbb{N}$ possède un plus petit élément : en effet, l'ensemble A des nombres plus petits que tous les éléments de A contient 0 et, d'après le principe de récurrence, contient un nombre k dont le successeur cesse d'être plus petit que tous les éléments de A. Ce nombre k est bien le plus petit de tous les éléments de A (montrer pourquoi, c'est un bon exercice de petit raisonnement logique pour vous entraîner). C'est aussi le plus grand de tous les nombres qui sont plus petits que tous les éléments de A (on dit que c'est *le plus grand minorant* où encore la *borne inférieure* de A, il se trouve que c'est en fait un élément de A).

A2. Toute partie non vide A de $\mathbb{N}$ majorée par un nombre M (*i.e* $\forall x \in A\,, x \leq M$) admet un plus gr--and élément : il suffit de prendre le plus petit élément de l'ensemble des nombres qui majorent A. Ce nombre est le *plus petit majorant* de A (on dit aussi la *borne supérieure* de A,il se trouve que c'est, ici encore, un élément de A).

A3. $\mathbb{N}$ n'admet pas de majorant

Les trois propriétés A1,A2,A3 équivalent aux axiomes de Peano et l'on peut donc dire que $\mathbb{N}$ est le seul ensemble qui puisse être équipé d'un ordre de manière à ce que les trois clauses A1,A2,A3 soient remplies.

La clause A1 nous dit que l'ordre est *total* : si a et b sont deux éléments de $\mathbb{N}$, on a toujours $a \leq b$ ou $b \leq a$.

Plus généralement, on appelle *ordre* sur un ensemble une relation $\leq$ ayant les trois propriétés suivantes :
- $a \leq a$ (*réflexivité*)
- $((a \leq b) \wedge (b \leq a)) \Longrightarrow (a = b)$ (*antisymétrie*)
- $((a \leq b) \wedge (b \leq c)) \Longrightarrow (a \leq c)$ (*transitivité*).

L'ordre est dit *total* si étant donné deux éléments a et b de l'ensemble, on a toujours $a \leq b$ ou $b \leq a$. Sinon, l'ordre est dit *partiel*. Par exemple, la relation «a divise b» dans $\mathbb{N}$ est une autre relation d'ordre sur $\mathbb{N}$, non totale. L'inclusion $A \subset B$ est une relation d'ordre (encore non total) sur l'ensemble $\mathcal{P}(E)$ des parties d'un ensemble E.

1.8.4 Le principe du raisonnement par récurrence

PRINCIPE 1 : *Soit $R\{n\}$ une assertion impliquant le caractère n et n_0 un nombre entier positif fixé. Alors, l'assertion*

$$\Big(R\{n_0\} \wedge \Big(\forall n \geq n_0\,,\ R\{n\} \Longrightarrow R\{n+1\}\Big)\Big) \Longrightarrow \Big(\forall n \geq n_0\,,\ R\{n\}\Big)$$

est une évidence dans l'axiomatique de Peano.

Preuve. Si $a \in \mathbb{N}$, on note $a + n_0$ le n_0-ème successeur de a. Il suffit d'appeler A l'ensemble des entiers tels que $R\{a + n_0\}$ est VRAIE ; on a $0 \in A$ et $a \in A \Longrightarrow S(a) \in A$; on peut donc appliquer l'axiome 5.

On a une version bis :

PRINCIPE 2 : *Soit $R\{n\}$ une assertion impliquant le caractère n et n_0 un nombre entier positif fixé. Alors, l'assertion*

$$\Big(R\{n_0\} \wedge \Big(\forall n \geq n_0\,,\ \Big((\forall k \in \{n_0, ..., n\}\,,\ R\{k\}) \Longrightarrow R\{n+1\}\Big)\Big)\Big) \Longrightarrow \Big(\forall n \geq n_0\,,\ R\{n\}\Big)$$

est une évidence dans l'axiomatique de Peano.

On donne des exemples empruntés à l'arithmétique et à la relation de division.

Définition 1.9. Un nombre $p \geq 2$ de $\mathbb{N}$ est dit premier si et seulement si ses seuls diviseurs dans $\mathbb{N}$ sont 1 et p.

Par exemple, $2, 3, 5, 7$, sont premiers, 4,6, 8, 9 ne le sont pas.

Tout nombre $n \geq 2$ admet au moins un diviseur premier. C'est un exemple d'application du principe 2. Cette assertion (notée $R\{n\}$) est vraie si $n = 2$. On la suppose vraie pour tout $k \leq n$ et l'on examine $n + 1$. De deux choses l'une :
- soit $n + 1$ est premier et alors $n + 1$ a un diviseur premier, $R\{n + 1\}$ est vraie ;
- soit $n + 1$ admet un diviseur p différent de 1 et $n + 1$, donc tel que $p \in \{2, ..., n\}$. On a $n + 1 = pq$ et p admet un diviseur premier p' puisque $p \leq n$ (hypothèse de récurrence) ; ce diviseur p' divise p, qui lui divise $n + 1$, donc p' divise $n + 1$ et $R\{n + 1\}$ est vraie.

L'assertion

$$\Big((\forall k \in \{2, ..., n\}\,,\ R\{k\}) \text{ vraie}$$

implique donc ($R\{n + 1\}$ vraie et l'assertion ($\forall n \geq 2, R\{n\}$ VRAIE) résulte du principe 2. □

Un exemple de raisonnement par l'absurde : il y a une infinité de nombres premiers. Supposons le contraire et appelons $p_1, ..., p_N$ les nombres premiers (en nombre fini). Le nombre

$$P := p_1 \cdots p_N + 1$$

admet d'après ce qui précède un diviseur premier p, donc dans la liste $p_1, ..., p_N$. Mais p divise $p_1 \cdots p_N$ et $p_1 \cdots p_N + 1$, donc divise 1, ce qui est impossible (car $p \geq 2$). L'hypothèse faite ("il existe un nombre fini de nombres premiers") conduit à une contradiction. L'assertion contraire ("il y a une infinité de nombres premiers") est donc vraie. □

Un autre exemple de preuve par récurrence : le théorème d'Euclide. *Soient a et b deux entiers positifs avec $b \neq 0$; il existe un unique couple (q, r) d'éléments de $\mathbb{N}$ avec $a = bq + r$ et $r \in \{0, ..., b-1\}$. Le nombre r est dit reste dans la division euclidienne de a par b.*

Preuve. On prouve l'assertion

$$\mathrm{R}\{a\} : \text{"il existe } (q, r) \in \mathbb{N} \times \{0, ..., b-1\}\,,\ a = bq + r\text{"}$$

par récurrence sur a en utilisant le principe 1. L'assertion $R\{0\}$ est vraie car $0 = b \times 0 + 0$. Supposons $R\{a\}$ vraie ; on a

$$a + 1 = (bq + r) + 1 = bq + (r + 1)$$

avec $r \in \{0, ..., b-1\}$. Si $r \leq b - 2$, on a $r + 1 \leq b - 1$ et $R\{a+1\}$ est vraie. Si $r = b - 1$, alors

$$a + 1 = bq + (b-1) + 1 = b(q+1) = b(q+1) + 0$$

et l'assertion $R\{a+1\}$ est vraie avec $r = 0$. L'assertion $R\{a\}$ implique donc toujours $R\{a+1\}$ et le principe 1 s'applique.

L'unicité de (q, r) vient du fait que si $a = q_1 b + r_1 = q_2 b + r_2$, on a, si $r_1 \leq r_2$

$$r_2 - r_1 = (q_1 - q_2)b \in \{0, ..., b-1\}\,,$$

ce qui implique $q_1 - q_2 = 0$ car sinon $(q_1 - q_2)b \geq b$; on a donc aussi $r_1 = r_2$ et l'unicité du couple (q, r) est prouvée. □

1.8.5 La division dans $\mathbb{N}$ et l'algorithme d'Euclide

Si a est un entier non nul, tout diviseur de a est nécessairement plus petit que a. Par conséquent, si a et b sont deux entiers positifs non tous les deux nuls, l'ensemble des diviseurs communs de a et de b est un sous-ensemble majoré de $\mathbb{N}$ qui admet un plus grand élément.

Définition 1.10. Si a et b sont deux entiers positifs non tous les deux nuls, on appelle *plus grand diviseur commun* (en abrégé PGCD) de a et b le plus grand de tous les nombres entiers positifs divisant à la fois a et b. Si $b = 0$, le PGCD de a et b vaut donc a. On dit que a et b sont *premiers entre eux* si leur PGCD est égal à 1.

Le calcul du PGCD de deux nombres entiers positifs a et b avec b non nul fait apparaître une démarche mathématique constructive (donc implémentable sur une machine) que l'on appelle un *algorithme*. Ce terme vient du surnom Al-Khwarizmi du mathématicien ouzbek Abu Ja'far Mohammed Ben Musa, 780-850, (dont le début du titre d'un des ouvrages fournit d'ailleurs aussi le mot *algèbre*).

Notons (comme sous la syntaxe du logiciel de calcul MATLAB que nous utilisons ici)

```
[q,r] = div(a,b)
```

l'instruction qui calcule, étant donnés deux nombres entiers tels que $b \neq 0$, l'unique couple (q, r) avec $r \in \{0, ..., b-1\}$ tel que $a = bq + r$ donné par la division euclidienne de a par b. On pourrait tout aussi bien utiliser les commandes de MAPLE10. Considérons la suite d'instructions (on les a traduites ici en français pour les expliciter) qui conduit au calcul du PGCD de deux entiers a et b ; le nombre à calculer est noté PGCD dans la suite d'instructions ci-dessous.

```
fonction PGCD=PGCD(a,b);

x=a;
y=b;
tant que y est non nul, faire
   [q,r] = div(x,y);
   si r=0
     PGCD = y;
     y=0;
   sinon
     [q1,r1]= div(y,r);
     x=r;
     PGCD = x;
     y=r1;
fin
```

ce qui se lit comme la suite de calculs

$$\begin{aligned} a &= bq_0 + r_0 \\ b &= r_0 q_1 + r_1 \\ r_0 &= r_1 q_2 + r_2 \\ \vdots &= \vdots \\ r_{N-2} &= q_N r_{N-1} + r_N \\ r_{N-1} &= r_N q_{N+1} + 0 \end{aligned}$$

(comme les restes successifs $r_0, r_1, ...$ décroissent strictement et qu'il y a un nombre fini d'entiers entre 0 et b, il vient forcément un moment où r_N divise r_{N-1}, r_N étant le dernier reste obtenu non nul). Le PGCD de a et b est aussi celui de b et r_0, de r_0 et r_1, et ainsi, en cascade, de r_N et 0 ; il vaut donc r_N.

CONCLUSION : *Le PGCD de a et b est donc égal au dernier reste non nul r_N dans ce très célèbre algorithme de division dit algorithme d'Euclide.*

Par exemple, pour $a = 13$ et $b = 4$,

$$\begin{aligned} 13 &= 4 \times 3 + 1 \\ 4 &= 4 \times 1 + 0 \end{aligned}$$

le dernier reste non nul étant ici $r_0 = 1$. On a donc $\text{PGCD}\,(a, b) = 1$.

Plus loin dans ce cours, on apprendra à "remonter" ces calculs une fois que l'on saura manier les nombres entiers négatifs.

1.8.6 Un autre algorithme issu de la division euclidienne : écrire en base b

Soit b un nombre entier supérieur où égal à 2 et a un nombre entier quelconque. On peut exécuter la suite d'instructions

```
fonction X=newbase(a,b);

X=[];
x=a;
tant que x est non nul, faire
   (q,r) = div(x,b);
   x= q ;
   X=[r,X];
fin
```

Cet algorithme s'arrête forcément (on va voir tout de suite pourquoi) et fournit une liste

$$X = [d_{N-1}, d_{N-2}, ..., d_0]$$

d'entiers entre 0 et $b-1$ (d_{N-1} étant non nul) tels que

$$a = d_0 + d_1 b + d_2 b^2 + \cdots + d_{N-1} b^{N-1}\,; \qquad (**)$$

cette liste est d'ailleurs l'unique liste de nombres entre 0 et $b-1$ telle qu'il soit possible d'écrire a sous la forme (**) pour un certain entier N. Cette suite X est appelée *écriture en base b* de l'entier a. Le cas $b = 2$ est intéressant car l'écriture en base 2 se fait avec 2 chiffres. Par exemple, en base 2, on vérifiera, comme

$$7 = 1 + 1 \times 2 + 1 \times 2^2,$$

que l'écriture de 7 est 111.

La raison pour laquelle l'algorithme s'arrête (on obtient un quotient nul au bout d'un certain dans les divisions) est que, comme $b \geq 2$, b^n finit par dépasser a si n est assez grand.

La décomposition en base 10 est la plus connue (les chiffres d_k étant les chiffres $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Cette idée présidera au chapitre suivant à la construction des nombres décimaux.

1.8.7 Le développement d'une fraction en fraction continue

En utilisant l'algorithme de division euclidienne (encore lui !), on peut construire l'algorithme qui permet d'écrire toute fraction a/b (avec $a \in \mathbb{N}$ et $b \in \mathbb{N} \setminus \{0\}$ sous la forme

$$a/b = p_0 + \frac{1}{p_1 + \frac{1}{p_2 + \frac{1}{p_3 + \cdots}}}\,, \qquad (\dagger)$$

avec $p_0 \in \mathbb{N}$ et $p_k \in \mathbb{N} \setminus \{0\}$ pour $k \geq 1$, la suite $p_0, p_1, p_2, ...$ étant finie ; par exemple

$$\begin{aligned} \frac{26}{15} = 1 + \frac{11}{15} &= 1 + \frac{1}{1 + \frac{4}{11}} \\ &= 1 + \frac{1}{1 + \frac{1}{2 + \frac{1}{1 + \frac{1}{3}}}} \end{aligned}$$

Une telle écriture (unique) est dite *développement en fraction continue* de la fraction a/b. Voici (vous pouvez vous entraîner à le vérifier, puis éventuellement à le tester sur des exemples comme nous l'avons fait en cours) la syntaxe de l'algorithme qui, étant donnés deux nombres entiers positifs a et b (avec $b \neq 0$) retourne la suite $p_0, p_1, ..., p_N$ des nombres impliqués dans le développement en fraction continue (†) de la fraction a/b.

```
fonction DVLP=fraccont(a,b);

DVLP=[];
x=a;
y=b;
tant que y > 0
    [q,r] = div(x,y);
    DVLP=[DVLP,q];
    x = y;
    y = r;
fin
```

Les développements en fraction continue infinis comme

$$1+\frac{1}{1+\frac{1}{1+\frac{1}{1+\dots}}}$$

donnent naissance à des nombres irrationnels, comme le *nombre d'or*

$$x=1+\frac{1}{1+\frac{1}{1+\frac{1}{1+\dots}}}=1+\frac{1}{x}$$

que l'on calculera en résolvant l'équation du second degré $x^2-x-1=0$ (faites ce petit calcul pour vous rafraîchir les idées sur la résolution des équations du second degré). On verra ainsi que

$$x=\frac{\sqrt{5}+1}{2}\,,$$

ce qui laisse pressentir que $\sqrt{5}$ ne saurait être un nombre rationnel (on y reviendra au chapitre 2, lorsque nous passerons de l'étude des nombres rationnels à celle des nombres réels). Cette idée (du développement en fraction continue "illimité") sera précisément une de celles nous permettant d'introduire la notion de *nombre réel via* les approximations rationnelles.

1.9 Notion de fonction ; éléments de combinatoire

1.9.1 Fonction d'un ensemble E dans un ensemble F ; exemples

Définition 1.11. Soient E et F deux ensembles non vides ; on appelle *fonction* (ou encore *application* f de l'ensemble E dans l'ensemble F toute partie G_f du produit $E\times F$ telle que, pour tout x dans E, il existe un unique élément y tel que $(x,y)\in G_f$; on dit que G est le *graphe* de l'application f et l'on note $y=f(x)$ ou encore $f\ : x\mapsto f(x)$. L'ensemble E est dit *domaine de définition* de la fonction f.

Exemples : Les sous-ensembles

$$G:=\{(x^2,x)\,;\,x\in\mathbb{R}\}\qquad G:=\{(\sin x,x)\,;\,x\in\mathbb{R}\}$$

ne sont pas des fonctions, tandis que les sous-ensembles

$$G:=\{(x,x^2)\,;\,x\in\mathbb{R}\}\qquad G:=\{(x,\sin x)\,;\,x\in\mathbb{R}\}$$

en sont (ce sont respectivement les fonctions $x\mapsto x^2$ et $x\mapsto\sin x$).

L'ensemble des fonctions de E dans F est donc une partie de l'ensemble produit $X\times Y$ (dès que E contient plus d'un élément, ce n'est pas tout l'ensemble $E\times F$). On note en général F^E l'ensemble des applications de E dans F. Ceci peut sembler artificiel mais on en verra une justification un peu plus loin (dans la section 1.9.6, lorsque l'on se restreindra au cadre des ensembles finis).

1.9.2 Injection, surjection, bijection

Définition 1.12. Soient E et F deux ensembles non vides ; une fonction f de E dans F est dite *injective* (on dit que c'est une *injection*) si et seulement si l'assertion

$$\forall x_1 \in E\,,\ \forall x_2 \in E\,,\ (f(x_1) = f(x_2)) \Longrightarrow (x_1 = x_2)$$

est vraie (on utilise aussi l'assertion contraposée qui est équivalente et que l'on écrira).

Exemple : L'application $x \mapsto x^3$ de $\mathbb{R}$ dans $\mathbb{R}$ est injective, tandis que $x \mapsto x^2$ (toujours de $\mathbb{R}$ dans $\mathbb{R}$) ne l'est pas.

Définition 1.13. Soient E et F deux ensembles non vides ; une fonction f de E dans F est dite *surjective* (on dit que c'est une *surjection*) si et seulement l'assertion

$$\forall y \in F\,,\ \exists x \in E\,, y = f(x)$$

est vraie.

Exemple. L'application $x \mapsto x^3$ de $\mathbb{R}$ dans $\mathbb{R}$ est surjective, tandis que $x \mapsto x^2$ (toujours de $\mathbb{R}$ dans $\mathbb{R}$) ne l'est pas.

On appelle *bijection* entre deux ensembles E et F une fonction de E dans F qui est à la fois injective et surjective. Cela se lit encore

$$\forall y \in F\,,\ \exists\,!\, x \in E\,, y = f(x)$$

si le symbole $\exists\,!\,x$ se lit "il existe un x unique ...".

Exemple important. Si E est un ensemble, on peut associer une fonction importante χ_A à toute partie A de E ; la fonction χ_A (dite *fonction caractéristique* de A) est la fonction de E dans l'ensemble $\{0,1\}$ définie par

$$\chi_A(x) := \begin{cases} 1 \text{ si } x \in A \\ 0 \text{ sinon} \end{cases}$$

On vérifiera en exercice que si A et B sont deux parties de E, alors, pour tout $x \in E$,

$$\begin{aligned} \chi_{A\cup B}(x) &= \max(\chi_A(x), \chi_B(x)) \\ \chi_{A\cap B}(x) &= \chi_A(x) \times \chi_B(x) \\ \chi_{E\setminus A}(x) &= 1 - \chi_A(x)\,. \end{aligned}$$

Ceci étant posé, on constate que l'application de $\mathcal{P}(E)$ dans l'ensemble des fonctions de E dans $\{0,1\}$ qui à A associe χ_A est une bijection. Elle met donc en bijection l'ensemble $\mathcal{P}(E)$ des parties de E avec l'ensemble des applications de E dans $\{0,1\}$ que l'on note $\{0,1\}^E$.

1.9.3 Image directe, image réciproque

Soient E et F deux ensembles non vides et f une fonction de E dans F.

Pour toute partie A de E, on appelle *image directe* de A et on note $f(A)$ le sous-ensemble de F défini par

$$f(A) := \{y \in F\,;\, \exists x \in A \text{ tel que } y = f(x)\}\,.$$

Pour toute partie B de F, on appelle *image inverse* de B et on note $f^{-1}(B)$ le sous-ensemble de E défini par

$$f^{-1}(B) := \{x \in E\,;\, f(x) \in B\}\,.$$

Exemple. Si f est l'application de $\mathbb{R}$ dans $\mathbb{R}$ définie par $f(x) = x^2$ (elle n'est ni injective, ni surjective), on vérifiera pour s'entraîner que $f([1,2]) = [1,4]$ et que $f^{-1}([1,4]) = [-2,-1] \cup [1,2]$; remarquons sur cet exemple que $f^{-1}(f(A))$ peut être beaucoup plus gros que A !

On vérifie tout de suite que ces deux opérations de prise d'image directe et de prise d'image inverse sont liées en fait par les relations

$$\forall A \in \mathcal{P}(E)\,, \qquad A \subset f^{-1}(f(A))$$
$$\forall B \in \mathcal{P}(F)\,, \qquad f(f^{-1}(B)) \subset B\,.$$

Proposition 1.1. *Si f est injective, pour toute partie A de E, on a $f^{-1}(f(A)) = A$; si f est surjective, pour toute partie B de F, $f(f^{-1}(B)) = B$.*

Preuve. Prouvons le premier point ; on connaît déjà une inclusion et il suffit donc de montrer que $f^{-1}(f(A)) \subset A$. Prenons x dans $f^{-1}(f(A))$, ce qui signifie qu'il existe x' dans A tel que $f(x) = f(x')$; comme f est injective, on a $x = x'$ et donc $x \in A$, c'est gagné. L'autre point (pour f surjective) se montre encore plus facilement : si $y \in B$, y s'écrivant $f(x)$, on a nécessairement $x \in f^{-1}(B)$ (par définition) et donc $y \in f(f^{-1}(B))$, d'où l'inclusion $B \subset f(f^{-1}(B))$. □

1.9.4 Composition des applications, inverses à gauche et à droite

Si E, F, G sont trois ensembles non vides, f une application de E dans F, g une application de F dans G, on définit *l'application composée* $g \circ f$ par

$$\forall x \in E\,,\ g \circ f(x) := g(f(x))\,;$$

c'est une application de E dans G.

Exemple. Si $E = F = G = \mathbb{R}$ et si f désigne l'application définie par $f(x) = x^2 + 1$, l'application $f \circ f$ est l'application

$$x \in \mathbb{R} \longmapsto X = x^2 + 1 \longmapsto X^2 + 1 = (x^2+1)^2 + 1 = x^4 + 2x^2 + 2\,;$$

ces applications du type $f \circ f$, $f \circ f \circ f$, ..., dites *itérées* d'une application donnée f, jouent un rôle important dans un domaine aujourd'hui en plein essor des mathématiques (aux confins de la physique), la *dynamique* ; les structures fractales dérivent par exemple de cette idée d'itération d'application.

On dit qu'une application f de E dans F est *inversible à gauche* si et seulement si il existe une application g de F dans E telle que

$$\forall x \in E\,,\ (g \circ f)(x) = x\,.$$

On a alors la proposition :

Proposition 1.2. *Soient E et F deux ensembles non vides ; une application f de E dans F est injective si et seulement si elle est inversible à gauche.*

Preuve. Si f est inversible à gauche et que $f(x_1) = f(x_2)$ on a $g(f(x_1)) = x_1 = g(f(x_2) = x_2$, donc $x_1 = x_2$, ce qui prouve que f est injective. Si maintenant f est injective, on construit un inverse à gauche en posant $g(y) = x$ si $y = f(x)$ (il n'y a qu'un seul $x \in E$ qui convienne) et $g(y) = x_0$ où x_0 est n'importe quel élément de E si $y \notin f(E)$. □

On dit qu'une application f de E dans F est *inversible à droite* si et seulement si il existe une application g de F dans E telle que

$$\forall y \in F\,,\ (f \circ g)(y) = y\,.$$

On a alors la proposition :

Proposition 1.3. *Soient E et F deux ensembles non vides ; une application f de E dans F est surjective si et seulement si elle est inversible à droite.*

Preuve. Si f est surjective, on définit une application g de F dans E en posant, pour tout $y \in F$, $g(y) = x(y)$, où $x(y)$ est un élément de $f^{-1}(\{y\})$ (qui est non vide par hypothèses) ; c'est l'axiome du choix (voir la section 1.3) auquel on recours ici pour justifier que l'on puisse réaliser ce processus de sélection qui permet de "piquer" un élément $g(y)$ dans chaque sous-ensemble (non vide par hypothèses) $f^{-1}(\{y\})$.
S'il existe un inverse à droite g, le fait que tout y de F s'écrive $y = f(g(y))$ implique bien que f est surjective. □

Proposition 1.4. *Soient E et F deux ensembles non vides ; une application f de E dans F est bijective si et seulement si elle est inversible à droite et à gauche ; les inverses g_1 (à gauche) et g_2 (à droite) sont alors égaux, on note $g_1 = g_2 = f^{-1}$ et on dit alors que f est inversible.*

Preuve. On combine les deux propositions 1.2 et 1.3. Si f admet un inverse à gauche g_1 et un inverse à droite g_2, f est donc bijective, la réciproque étant vraie aussi. Les deux inverses à gauche g_1 et à droite g_2 sont alors égaux : en effet, on a $g_1 \circ f = \mathrm{Id}_E$, ce qui implique, en composant par g_2 à droite et en utilisant l'associativité, $g_1 \circ (f \circ g_2) = g_1 \circ \mathrm{Id}_F = g_1 = \mathrm{Id}_E \circ g_2 = g_2$, donc finalement $g_1 = g_2$ comme voulu. On voit, si $f^{-1} = g_1 = g_2$, que $f^{-1} \circ f$ est l'application identité sur E et que $f \circ f^{-1}$ est l'application identité sur F. □.

1.9.5 Applications d'un ensemble fini dans un autre

Soient E et F deux ensembles finis, E ayant exactement p éléments et F ayant exactement n éléments. On a la proposition intéressante suivante :

Proposition 1.5. *S'il existe une application injective de E dans F, alors $n \geq p$. Si tel est le cas, il y a exactement $n(n-1)(n-2)\cdots(n-p+1)$ applications injectives de E dans F et le nombre entier*

$$A_n^p := n \times (n-1) \times \cdots \times (n-p+1)$$

est appelé nombre d'arrangements de p éléments parmi n. C'est aussi le nombre de p-uplets (ordonnés) distincts que l'on peut fabriquer à partir des éléments d'un ensemble de cardinal n.

Preuve. La preuve repose sur un très important principe en mathématiques (que le théoricien des nombres allemand Carl L. Siegel, 1896-1981, a contribué à populariser) : si l'on dispose de p allumettes et que l'on doive les ranger toutes dans n tiroirs, on en mettra nécessairement deux dans un même tiroir dès que $p > n$! Ceci implique que si f est une injection de E dans F, alors nécessairement le nombre d'éléments de E ne saurait excéder celui de F. Pour compter le nombre d'injections de E dans F (si $n \geq p$), il faut décider de numéroter les éléments de E, x_1,...,x_n. On a n choix possibles pour $f(x_1)$; une fois $f(x_1)$ choisi, il reste $n-1$ choix possibles pour $f(x_2)$, *etc.* ; finalement, il reste $n-(p-1)$ choix possibles pour $f(x_p)$ et le compte y est en faisant le produit. □

Proposition 1.6. *S'il existe une application surjective de E dans F, alors $p \geq n$.*

Preuve. Pour chaque élément y de F, il doit au moins exister un élément de E tel que $y = f(x)$. Il faut donc nécessairement que E contienne plus d'éléments que F. □

Si E et F sont deux ensembles finis de même cardinal, dire que f est une injection de E dans F équivaut à dire que f est une surjection de E dans F, ou encore que f est une bijection entre E et F.

1.9.6 Éléments de combinatoire

Si E et F sont deux ensembles finis (de cardinaux respectifs p et n), l'ensemble des applications de E dans F a pour cardinal n^p ; en effet, pour chaque élément de E, il y a n choix possibles pour l'image.

Ceci justifie que l'on note F^E l'ensemble des applications de E dans F.

En particulier, si E est de cardinal p, l'ensemble des applications de E dans $\{0,1\}$ est de cardinal 2^p. Par conséquent, *l'ensemble des parties d'un ensemble à p éléments a pour cardinal* 2^p.

Dans un ensemble à n éléments, on peut compter le nombre de parties à p éléments, $p = 0, ..., n$. Il y a 1 partie à 0 élément (la partie vide), n parties à 1 élément, *etc.* Si l'on note

$$\binom{n}{p}$$

le nombre de parties à p éléments dans un ensemble à n éléments, on voit que l'on a la relation

$$\binom{n}{p} = \binom{n-1}{p-1} + \binom{n-1}{p} \qquad (*)$$

pour tout n strictement plus grand que 1 et tout p entre 1 et n ; il suffit pour voir cela de trier d'un côté les parties à p éléments assujetties à contenir un élément marqué de l'ensemble (il en a autant que de parties à $p-1$ éléments dans un ensemble à $n-1$ éléments), de l'autre les parties à p éléments qui ne contiennent pas cet élément marqué (il y en a autant que de parties à p éléments dans un ensemble à $n-1$ éléments).

En convenant que

$$\binom{n}{0} = 1$$

pour tout $n \geq 0$ (la seule partie à zéro élément d'un ensemble à n éléments donné est la partie vide) et que

$$\binom{0}{1} = 0$$

(ce qui est logique, il n'y a pas d'élément dans l'ensemble vide !), la relation $(*)$ reste valable pour tout entier $n \geq 1$ et tout entier p entre 0 et n et cette liste de relations s'écrit sous forme d'un tableau triangulaire descendant, le *triangle de Pascal* (de fait, ce triangle était connu des mathématiciens arabes bien avant Pascal) :

$$\begin{array}{ccccccc} 1 & & & & & & \\ 1 & 1 & & & & & \\ 1 & 2 & 1 & & & & \\ 1 & 3 & 3 & 1 & & & \\ 1 & 4 & 6 & 4 & 1 & & \\ \vdots & \vdots & \vdots & & & & \end{array}$$

(on a indiqué sur la ligne n $(n = 0, 1, ...)$ la liste dans l'ordre des nombres $\binom{n}{p}$ pour $p = 0, ..., n$).

Le nombre $\binom{n}{p}$, pour $n \geq 1$ et p entre 0 et n est appelé *nombre de combinaisons de p éléments pris parmi n* ; on le note aussi C_n^p (par analogie avec la notation A_n^p utilisée pour le nombre d'arrangements) ; cependant on préfèrera dans ce cours utiliser la notation $\binom{n}{p}$, plus classique dans la terminologie anglo-saxonne.

Il existe une relation entre le nombre d'arrangements de p éléments parmi n et le nombre de combinaisons de p éléments parmi n lorsque p est entre 1 et n.

Proposition 1.7. *Si n est un entier non nul et p un entier entre 1 et n, les nombres entiers A_n^p et $\binom{n}{p}$ sont liés par la relation*

$$\binom{n}{p} \times p! = A_n^p \,,$$

où $p! := 1 \times 2 \times \cdots \times (p-1) \times p$, d'où la formule

$$\binom{n}{p} = \frac{n!}{p!\,(n-p)!}$$

(ce nombre est une "fausse" fraction, c'est un nombre entier !).

Remarque. Si la formule proposée dans cet énoncé pour exprimer le nombre de combinaisons de p éléments parmi n est intéressante du point de vue théorique (c'est d'ailleurs un fait surprenant que cette écriture fractionnaire se simplifie pour donner un entier !), il s'agit en revanche d'une formule inexploitable du point de vue théorique ($n!$ est très vite un entier énorme ingérable informatiquement, pensez par exemple à un nombre comme $1000! = 1000 \times 999 \times 998 \times \cdots$!).

Preuve. À une partie constituée de p éléments correspond autant d'arrangements de p éléments parmi n qu'il y a de possibilités de permuter entre eux ces p éléments. Or il y a $A_p^p = p!$ applications bijectives (ou injectives, c'est pareil) d'un ensemble à p éléments dans lui-même. On a la relation voulue. □

En comptant toutes les parties d'un ensemble à n éléments et en les classant suivant leur cardinal, on a la relation

$$2^n = \sum_{p=0}^{n} \binom{n}{p} \,.$$

qui devrait vous rappeler le cas particulier $(1+1)^n = 2^n$ de la formule du binôme dans le calcul algébrique commutatif :

$$(x+y)^n = \sum_{p=0}^{n} \binom{n}{p} x^p y^{n-p} \,.$$

On justifiera pourquoi cette relation entre la formule du binôme et les nombres de combinaisons $\binom{n}{p}$ (que l'on appelle aussi pour cette raison *coefficients binomiaux*) en comptant combien de fois $x^p y^{n-p}$ apparaît lorsque l'on développe

$$(x+y) \times (x+y) \times \cdots \times (x+y)$$

($x+y$ multiplié n fois par lui-même) ; il faut choisir x dans p de ces facteurs, y dans les $n-p$ facteurs restants et le nombre de possibilités de faire cela est précisément le nombre de parties à p éléments dans un ensemble à n éléments. Pareille formule n'est valable que sous la clause de commutativité, à savoir que $xy = yx$; vous verrez en effet ultérieurement que dans le cadre du calcul matriciel, la formule du binôme est de fait beaucoup plus compliquée car on ne peut plus cette fois "regrouper" (comme nous venons de le faire) les mots contenant p fois le symbole x et $n-p$ fois le symbole y.

FIN DU CHAPITRE 1

Chapitre 2

Nombres entiers, rationnels, réels et complexes

C'est une raison algébrique (l'impossibilité de résoudre dans $\mathbb{N}$ l'équation $a + x = b$ si a et b sont des entiers tels que $b < a$) qui a motivé la construction de $\mathbb{Z}$. Les fractions de nombres entiers se sont introduites ensuite naturellement, leur écriture décimale faisant intervenir, comme

$$\frac{22}{7} = 3,142857\ 142857\ 142857\cdots$$

un mot répété indéfiniment. Alors que tout sous-ensemble majoré (*resp.* minoré) de $\mathbb{Z}$ admet un plus grand (*resp.* plus petit) élément, ceci cesse d'être vrai pour les sous-ensembles de l'ensemble $\mathbb{Q}$ des fractions (dits aussi *nombres rationnels*). En effet, des nombres comme $\sqrt{2}$ et le nombre d'or

$$1 + \frac{1}{1 + \frac{1}{1+\frac{1}{1+\cdots}}} = \frac{\sqrt{5}+1}{2}$$

ne peuvent être des fractions (on le verra plus loin) et l'ensemble

$$\{x \in \mathbb{Q}\,;\, x^2 \leq 2\}$$

qui est un sous-ensemble majoré de $\mathbb{Q}$ n'admet pas de plus petit majorant dans $\mathbb{Q}$ (s'il y en avait un, ce serait $\sqrt{2}$ qui ne peut être dans $\mathbb{Q}$!)

C'est pour retrouver pour $E = \mathbb{Q}$ cette propriété indispensable à la notion de mesure, à savoir que tout sous-ensemble majoré (*resp.* minoré) d'un certain ensemble ordonné $(E, \leq)$ admette un plus petit majorant (*resp.* un plus grand minorant) dans E qu'il a fallu "inventer" (ou découvrir, car le monde de l'analyse nous entourant est réel) les nombres réels. à travers eux, la perception de l'infiniment petit se fera jour.

2.1 L'anneau $\mathbb{Z}$ des entiers relatifs

2.1.1 Construction de l'anneau ordonné $(\mathbb{Z}, +, \times)$

L'ensemble des entiers positifs $\mathbb{N}$ est, on l'a vu, équipé d'un ordre total $\leq$ (on écrit aussi $a < b$ si $a \leq b$ et $a \neq b$) tel que toute partie non vide A de $\mathbb{N}$ admette un plus petit élément (le plus grand de

tous les minorants de A, dit aussi *borne inférieure* de A) et que toute partie non vide et majorée B de $\mathbb{N}$ admet un plus grand élément (le plus petit de tous les majorants de B, dit aussi *borne supérieure* de B).

En revanche, si $a, b \in \mathbb{N}$ et si $a > b$, on ne peut résoudre dans $\mathbb{N}$ l'équation $x + a = b$. C'est le cas si $a > 0$ et $b = 0$. Cela prive le monoïde $(\mathbb{N}, +)$ d'avoir la propriété suivante : tout élément a admet un inverse pour l'addition, *i.e* un élément a' de $\mathbb{N}$ tel que $a + a' = a' + a = 0$.

L'ensemble $\mathbb{Z}$ des entiers relatifs a lui cette propriété ; le monoïde $(\mathbb{Z}, +)$ est même le plus petit monoïde que l'on puisse construire qui contienne $\mathbb{N}$, l'addition prolongeant l'addition sur $\mathbb{N}$, avec de plus la propriété que tout élément ait cette fois un opposé pour l'addition. L'ensemble $\mathbb{Z}$ est réalisé comme l'ensemble des classes de couples d'entiers positifs $[(a, b)]$, où l'on décide de mettre dans la même classe (ou encore d'identifier) deux couples quelconques (a_1, b_1) et (a_2, b_2) tels que $b_1 + a_2 = a_1 + b_2$; notons que c'est naturel de mettre dans la même "classe" deux tels couples puisque, formellement (si on avait droit aux soustractions), l'égalité $a_1 + a_2 = b_1 + b_2$ se lirait aussi $b_1 - a_1 = b_2 - a_2$; il est donc licite de faire l'identification si l'on pense un entier relatif comme la "différence" de deux éléments de $\mathbb{N}$. On peut interpréter a comme une "perte", b comme un "gain", le couple (a, b) étant lui pensé comme un bilan ; dire que $a_1 + b_2 = a_2 + b_1$ revient à dire que (a_1, b_1) et (a_2, b_2) correspondent au même bilan comptable.

On peut aussi (pour plus de simplicité et si l'on est moins féru de constructions abstraites) voir $\mathbb{Z}$ comme l'union de $\mathbb{N}$ et des opposés $-1, -2, ...$ des entiers positifs non nuls. En effet, la classe $[(a, b)]$ correspond à l'entier naturel strictement positif $b - a$ (et on la note ainsi) si $b > a$; cette même classe $[(a, b)]$ est notée (c'est une convention) $-(a-b)$ lorsque $a > b$ ($a-b$ est alors un entier naturel strictement positif) ; elle est notée enfin 0 si $a = b$.

On a un ordre total sur $\mathbb{Z}$ prolongeant l'ordre sur $\mathbb{N}$ en décidant que

$$\begin{aligned} &\forall a, b \in \mathbb{N}\,, \qquad -a \leq b \\ &\forall a, b \in \mathbb{N}\,, \qquad (-a \leq -b) \iff (b \leq a) \end{aligned}$$

L'addition des deux classes $[(a_1, b_1)]$ et $[(a_2, b_2)]$ est par définition la classe $[(a_1 + a_2, b_1 + b_2)]$. Pour la multiplication, c'est un peu plus compliqué, il faut remarquer formellement que

$$(b_1 - a_1)(b_2 - a_2) = a_1a_2 + b_1b_2 - a_1b_2 - a_2b_1$$

est définir donc le produit des classes $[(a_1, b_1)]$ et $[(a_2, b_2)]$ par

$$[(a_1, b_1)] \times [(a_2, b_2)] = [(a_1b_2 + a_2b_1, a_1a_2 + b_1b_2)]\,.$$

Si l'on veut être plus concret, on peut simplement dire que la multiplication sur $\mathbb{Z}$ prolonge la multiplication sur $\mathbb{N}$ comme suit :

$$\begin{aligned} &\forall a, b \in \mathbb{N}\,, \qquad (-a) \times b := -(ab) \\ &\forall a, b \in \mathbb{N}\,, \qquad (a) \times (-b) = ab\,. \end{aligned}$$

L'ordre sur $\mathbb{Z}$ est compatible avec ces deux opérations au sens suivant

$$\begin{aligned} ((a \leq b) \wedge (c \leq d)) &\implies a + c \leq b + d \\ ((a \leq b) \wedge (c \geq 0)) &\implies ac \leq bc \\ ((a \leq b) \wedge (c \leq 0)) &\implies bc \leq ac \end{aligned}$$

L'addition sur $\mathbb{Z}$ est commutative ($x+y = y+x$), associative ($(x+y)+z = x+(y+z)$), admet un élément neutre (ici 0 car $0+x = x+0 = x$ pour tout x) et tout élément x admet un opposé x' pour l'addition (ici

$x' = -x$ car $x + (-x) = (-x) + x = 0$). On dit que $(\mathbb{Z}, +)$ est un *groupe abélien*, qualificatif emprunté au mathématicien norvégien Niels Henryk Abel (1802-1829), contemporain du mathématicien français Evariste Galois (1811-1832) qui, comme lui, inventa la théorie des groupes au service du problème de la résolubilité (ou non) à la règle et au compas des équations algébriques. La théorie des groupes, invention des mathématiciens pour des questions au départ de nature exclusivement mathématique, est devenue à l'orée du XX-ème siècle un formidable outil au service d'autres disciplines scientifiques, telles en particulier la chimie (liaisons atomiques, classification périodique). On retire l'épithète *abélien* lorsque l'opération d'addition cesse d'être commutative.

La multiplication sur $\mathbb{Z}$ est commutative ($xy = yx$) , associative ($(x.y).z = x.(y.z)$), 1 est élément neutre (on dit plutôt *élément unité* pour éviter de faire la confusion avec 0 qui lui est élément neutre pour l'addition), ce qui signifie $1.x = x.1 = x$, mais aucun entier relatif différent de ± 1 n'admet d'inverse pour la multiplication ; donc $(\mathbb{Z}, \times)$ n'est pas un groupe, pas plus que $(\mathbb{Z} \setminus \{0\}, \times)$.

En revanche, la multiplication dans $\mathbb{Z}$ est distributive par rapport à l'addition et l'on dit que $(\mathbb{Z}, +, \times)$ a ainsi *une structure d'anneau commutatif unitaire* : c'est un groupe abélien pour l'addition, la multiplication est commutative, associative, distributive par rapport à l'addition et admet un élément neutre (dit *élément unité*). Les épithètes *commutatif* et *unitaire* sont respectivement retirées lorsque la multiplication cesse d'être commutative ($xy \neq yx$ pour certains x, y) ou qu'il n'y a plus d'élément unité pour cette même opération de multiplication (comme ici 1 qui vérifie $1.x = x.1 = x$ pour tout x).

Toute partie A non vide et minorée de $\mathbb{Z}$ admet un plus petit élément (le plus grand de tous les minorants de A, dit encore *borne inférieure* de A). Toute partie B non vide et majorée de $\mathbb{Z}$ admet un plus grand élément (le plus petit de tous les majorants de B, dit encore *borne supérieure* de B). On notera que bornes inférieure et supérieure d'un sous-ensemble A de $\mathbb{Z}$, pourvu qu'elles existent, sont tous les deux des éléments de A. Ceci ne sera plus le cas avec les sous-ensembles de $\mathbb{R}$, on le verra plus loin !

2.1.2 Un exemple de calcul algébrique dans $\mathbb{Z}$: l'identité de Bézout

Si a est un entier relatif, on pose $|a| = a$ si $a \in \mathbb{N}$, $|a| = a'$ si $a = -a'$ avec $a' \in \mathbb{N}$.

Si a et b sont deux entiers relatifs non tous les deux nuls, on appelle *plus grand diviseur commun* de a et b (ou encore PGCD(a, b)) le PGCD des entiers positifs $|a|$ et $|b|$. Les nombres a et b sont premiers entre eux si ce PGCD vaut 1. On attribue au mathématicien français Etienne Bézout (1730-1783) le résultat suivant, dont l'intérêt pratique tant en mathématiques pures qu'appliquées est aujourd'hui devenu capital (il conditionne ce que l'on appelle le *lemme des restes chinois* que vous verrez plus tard).

Théorème 2.1. *Soient a et b deux nombres entiers relatifs non tous les deux nuls et d leur PGCD ; il existe au moins un couple $(u_0, v_0) \in \mathbb{Z}^2$ tel que*

$$au_0 + bv_0 = d$$

(une telle relation est appelée identité de Bézout lorsque $d = 1$). De plus, si $ab \neq 0$ et si $a = da'$ et $b = db'$, a' et b' sont premiers entre eux et toutes les solutions de l'équation $au + bv = d$ (cas particulier d'une équation à coefficients entiers et dont on cherche les solutions entières, prototype de ce que l'on appelle, en hommage à Diophante d'Alexandrie, une équation diophantienne) sont tous les couples (u, v) de la forme que

$$u = u_0 + b'k\,,\ v = v_0 - a'k\,, \qquad k \in \mathbb{Z}\,.$$

Preuve. La construction de u_0 et v_0 est algorithmique et se fait en remontant depuis l'avant-dernière

ligne les calculs faits dans l'algorithme de division euclidienne de $|a|$ par $|b|$ (on suppose ici $b \neq 0$).

$$\begin{aligned} |a| &= |b|q_0 + r_0 \\ |b| &= r_0 q_1 + r_1 \\ r_0 &= r_1 q_2 + r_2 \\ \vdots &= \vdots \\ r_{N-3} &= q_{N-1} r_{N-2} + r_{N-1} \\ r_{N-2} &= q_N r_{N-1} + d \\ r_{N-1} &= d q_{N+1} + 0 \end{aligned}$$

On écrit

$$\begin{aligned} d &= r_{N-2} - q_N r_{N-1} \\ &= r_{N-2} - q_N(r_{N-3} - q_{N-1} r_{N-2}) \\ &= -q_N r_{N-3} + (1 + q_N q_{N-1}) r_{N-2} \\ \vdots &= \vdots \\ &= u_0 a + v_0 b \end{aligned}$$

Pour prouver l'autre volet de l'assertion lorsque $ab \neq 0$, on remarque que toute solution (u, v) de l'équation diophantienne $au + bv = d$ s'écrit

$$u = u_0 + u' , \qquad v = v_0 + v' ,$$

où (u', v') est solution de l'équation diophantienne *sans second membre* (on dit encore *homogène*, on retrouvera cette terminologie avec les systèmes linéaires ou les équations différentielles) $au' + bv' = 0$. Si l'on pose $a = da'$ et $b = db'$, a' et b' sont premiers entre eux (car on a "évacué" en mettant d en facteur tous les diviseurs communs). Le couple (u', v') est solution de $a'u' = -b'v'$; mais comme le PGCD de a' et b' est égal à 1, il existe (x, y) dans $\mathbb{Z}^2$ avec $xa' + yb' = 1$, soit donc $(1 - yb')u' = -b'v'$, soit $u' = b'(yu' - v')$ ce qui prouve que b' divise u', soit $u' = b'k$ pour $k \in \mathbb{Z}$; de même a' divise v' et l'on a $v' = k'a'$; comme $a'b'$ est non nul, on a $a'u' + v'b' = 0$ implique $k' = -k$ et l'on a donc $u = u_0 + kb'$, $v = v_0 - ka'$ comme voulu. □

Plus que l'énoncé du théorème 2.1 lui même, ce qui est très important (parce que très utile) est qu'il s'agisse d'une assertion dont la démonstration est constructive, c'est-à-dire s'articule sur un algorithme. Voici, en quelques lignes de code, la fonction qui calcule, étant donnés deux entiers relatifs a et b avec $b \neq 0$ à la fois le PGCD d de a et b, mais aussi une paire d'entiers relatifs (u, v) telle que $d = au + bv$ (il s'agit, on le remarquera, d'un algorithme inductif qui s'auto-appelle) :

```
fonction [PGCD,u,v]=bezout(a,b);

x=a ;
y= abs(b) ;
[q,r]=div(x,y);
si r==0
    PGCD = y;
    u=0 ;
    v=1;
sinon
```

```
 [d,u1,v1]=bezout(y,r);
 PGCD=d;
 u=v1;
 v=sign(b)*(u1- q*v1);
fin
```

On pourra s'entraîner à programmer cet algorithme et à le tester sur une calculatrice programmable, comme nous le ferons dans le cours sous un environnement scientifique (MATLAB 7).

Corollaire 2.1.1. (lemme de Gauss, Carl-Friedrich Gauss, 1777-1855) *Soient a et b deux éléments non nuls de $\mathbb{Z}$ avec* $\mathrm{PGCD}\,(a,b)=1$. *Si c est un nombre entier relatif tel que b divise ac, alors b divise c.*

Preuve. On prend u et v tels que $au+bv=1$ et on multiplie cette égalité par c, ce qui donne $c=acu+bcv$; comme b divise acu (puisque b divise ac) et bien sûr bcv, b divise c. Nous avons d'ailleurs utilisé ce raisonnement dans la preuve du théorème 2.1. □

Une application du lemme de Gauss : il n'existe pas de couple d'entiers strictement positifs (p,q) tel que $\sqrt{2}=p/q$. Prouvons ceci par l'absurde : si $x=p/q$ existe, on aurait

$$p^2=2q^2\,;$$

si l'on suppose p et q premiers entre eux (la fraction étant écrite sous forme irréductible, ce qu'il est licite de supposer après simplification), on déduirait du lemme de Gauss (appliqué d'abord avec $a=q$ et $b=p$, premiers entre eux et $c=2q$) que p divise $2q$, puis (toujours avec Gauss et cette fois $a=q$, $b=p$, $c=2$) que p divise 2, donc que $p=1$ ou $p=2$; le cas $p=1$ donnerait $1=2q^2$, ce qui est impossible car 2 ne peut diviser 1 dans $\mathbb{Z}$! le cas $p=2$ donnerait, après simplification, $2=q^2$, ce qui est impossible aussi (2 ne saurait être le carré d'un entier).

Avant de clôre ce paragraphe, remarquons que l'on peut naturellement prolonger l'algorithme de division euclidienne (par un entier strictement positif b) à $\mathbb{Z}$ tout entier en remarquant que our tout $a\in\mathbb{Z}$, il existe un unique couple (q,r), avec $q\in\mathbb{Z}$ et $r\in\{0,...,b-1\}$ tel que $a=bq+r$. Ceci est connu lorsque $a>0$ et peut se démontrer par récurrence sur $a=-1,-2,-3,...$ exactement comme on l'a fait dans la section 1.7.4 pour prouver le théorème d'Euclide. L'unicité du couple (q,r) s'obtient de la même manière. On peut donc énoncer cette nouvelle version du théorème d'Euclide :

Théorème d'Euclide dans $\mathbb{Z}$. Soit a un entier relatif et b un nombre entier strictement positif ; il existe un unique couple d'entiers (q,r) avec $q\in\mathbb{Z}$ et $r\in\{0,...,b-1\}$ tel que $a=bq+r$; le nombre q est dit *quotient* de a par b dans la division euclidienne, le nombre r est lui dit *reste* après division euclidienne de a par b.

2.2 Nombres rationnels et nombres réels

2.2.1 Fractions et développements décimaux périodiques : deux approches des rationnels

Un nombre *rationnel* est par définition le quotient d'un nombre entier relatif a par un entier strictement positif b. On a deux manières de définir un tel quotient.

1. *Le point de vue abstrait*

Il est calqué sur ce que l'on a fait pour construire $\mathbb{Z}$. On décide qu'un tel quotient a/b est une classe $[[(a,b)]]$ de couples $(a,b)\in\mathbb{Z}\times(\mathbb{N}\setminus\{0\})$ (a étant appelé *numérateur*, b *dénominateur*), deux couples (a_1,b_1) et (a_2,b_2) étant dans la même classe si et seulement si $a_1b_2=a_2b_1$ (c'est naturel d'identifier deux tels couples car $a_1b_2=a_2b_1$ s'écrit aussi formellement $a_1/b_1=a_2/b_2$).

L'ensemble des fractions ainsi obtenu est le plus petit ensemble (on le note $\mathbb{Q}$) contenant $\mathbb{Z}$ auquel on puisse prolonger l'addition et la multiplication de $\mathbb{Z}$ selon les règles

$$[[(a_1, b_1)]] + [[(a_2, b_2)]] = [[(a_1 b_2 + a_2 b_1, b_1 b_2)]] \,,$$

ce qui est juste la *réduction au même dénominateur* si l'on pense cette formule comme

$$\frac{a_1}{b_1} + \frac{a_2}{b_2} = \frac{a_1 b_2 + a_2 b_1}{b_1 b_2} \,,$$

et

$$[[(a_1, b_1)]] \times [[(a_2, b_2)]] = [[(a_1 a_2, b_1 b_2)]] \,,$$

de manière à ce que $(\mathbb{Q}, +, \times)$ soit un anneau commutatif unitaire et que de plus tout élément non nul (a, b) de $\mathbb{Q}$ admette un inverse (en l'occurrence (b, a) puisque (ab, ba) correspond au couple $(1, 1)$, lui-même identifié au nombre 1).

Un anneau commutatif unitaire où tout élément x distinct de l'élément neutre 0 pour l'addition admet un inverse pour la multiplication (comme $(\mathbb{Q}, +, \times)$) est appelé *corps commutatif*.

Ce qui manque à l'ensemble $\mathbb{Q}$ ainsi construit n'est pas une propriété algébrique ($(\mathbb{Q}, +, \times)$ est bien un corps commutatif, on l'a vu), mais une propriété relevant plus de l'analyse : $\mathbb{Q}$ ne vérifie pas la propriété de la borne supérieure !

En effet, comme nous l'avons déjà dit, le sous-ensemble

$$\{x \in \mathbb{Q} \,;\, x^2 \leq 2\}$$

est un sous-ensemble majoré de $\mathbb{Q}$ (il est majoré par 3 car $9 \geq 2$), mais qui ne possède pas de borne supérieure dans $\mathbb{Q}$. En effet, la borne supérieure "potentielle" serait le nombre $\sqrt{2}$ dont on a vu (avec le lemme de Gauss, section 2.1.2) qu'il ne pouvait être rationnel ; on aurait pu aussi envisager le recours aux développements en fraction continue : si x vérifie $x^2 = 2$, on a $x^2 - 1 = 1$, soit aussi

$$x = 1 + \frac{1}{1 + x} \,; \qquad (\dagger)$$

L'unicité de l'écriture d'une fraction en fraction continue finie exclut qu'un nombre rationnel x puisse vérifier la formule (†) ci-dessus (le fait d'identifier par exemple le développement en fraction continue de x et celui obtenu en substituant ce développement au second membre de (†) conduit à une absurdité).

2. *Le point de vue concret hérité de l'école primaire*

Lorsque l'on doit calculer une fraction (comme 22/7) de nombres entiers positifs, on pose la division et l'on écrit les décimales après la virgule. On trouve par exemple ici $\dfrac{22}{7} = 3 + 0,142857\,142857\,142857\cdots$, ce que l'on note encore $\dfrac{22}{7} = 3 + 0,\overline{142857}$ pour indiquer que le "mot" 142857 est répété indéfiniment. Si l'on songe que le nombre de restes possibles dans la division euclidienne par b est fini (il y a b reste possibles), on voit que forcément le développement décimal d'une fraction a/b avec $a \in \mathbb{N}$ et $b \in \mathbb{N} \setminus \{0\}$ est de la forme

$$\frac{a}{b} = m + 0, d_1 \cdots d_N \overline{d_{N+1} \cdots d_N} \,,$$

avec $m \in \mathbb{Z}$ (quotient de a par b dans la division euclidienne), $d_1, ..., d_M$ étant des chiffres entre 0 et 9, la barre surmontant le "mot" $d_{N+1} \cdots d_M$ signifiant que ce mot se reproduit de manière périodique indéfiniment dans l'écriture décimale.

On peut supposer aussi $a \in \mathbb{Z}$ et associer à la fraction a/b la donnée de l'entier m (quotient de a par b dans la division euclidienne (élargie à $\mathbb{Z}$ comme on l'a vu en fin de section 2.1.2), ce peut être un entier positif ou négatif, que l'on appelle *partie entière* de a/b) et de la suite $d_1, d_2, ...$ où, à partir d'un certain cran, un certain mot $d_{N+1} \cdots d_M$ finit par se répéter indéfiniment. Les nombres d_j, $j = 1, 2, ...$, sont dites *décimales* de a/b.

On note

$$\frac{a}{b} = m + 0, d_1 d_2 \cdots d_N \overline{d_{N+1} \cdots d_M} \, . \qquad (\dagger\dagger)$$

Définition 2.1. Le développement (††) est dit *développement décimal* de la fraction a/b.

On peut vérifier qu'un développement décimal ayant ainsi un motif répété correspond au développement d'une fraction.

En revanche, le développement décimal d'un éventuel nombre x tel que $x^2 = 2$ (que l'on peut chercher par approximations successives en tâtonnant) ne peut présenter de "mot" se répétant par périodicité.

On traitera un exemple pour se convaincre, comme

$$x = 12, 431\overline{572} \, .$$

On a

$$1000x - 12431 = 0, \overline{572} \, ,$$

soit

$$1000(1000x - 12431) = 572, \overline{572} = 572 + 0, \overline{572} \, ,$$

soit encore

$$1000(1000x - 12431) - 572 = 1000x - 12431 \, ,$$

d'où l'on déduit bien que x est une fraction.

Parmi les fractions, figure une catégorie de fractions intéressantes, celles des *fractions décimales*, de la forme

$$x = m + 0, d_1 \cdots d_N = m + \frac{d_1}{10} + \frac{d_2}{100} + \cdots + \frac{d_N}{10^N} \, ,$$

Il y a un petit problème cependant ! On remarque que le nombre 1 s'écrit aussi

$$1 = 0 + 0, \overline{9} \, ;$$

en effet, si on ajoute à $9/10 = 0, 9$ le décimal $9/100 = 0, 09$, puis $9/1000 = 0, 009$, et que l'on répète N fois cette opération, on trouve

$$9(1/10 + 1/100 + 1/1000 + \cdots + 1/10^N) = \frac{9}{10} \times \Big(\frac{1 - \frac{1}{10^N}}{1 - \frac{1}{10}} \Big) = 1 - \frac{1}{10^N} \, ,$$

quantité qui devient arbitrairement petite plus N est grand ; on se rapproche de 1 sans jamais l'atteindre, comme l'archer avançant chaque fois en jetant sa flèche aux 1/10 de la distance au but qu'il espère atteindre ; il atteindra certes ce but (à 1/9 de sa position initiale), mais au bout d'une "infinité" de jets ! La notion de *série convergente*, que vous retrouverez l'année prochaine en mathématiques, est déjà cachée derrière ce paradoxe. Ainsi le nombre décimal

$$m + 0, d_1 \cdots d_N$$

avec $d_N \in \{1, ..., 9\}$ a t-il comme autre développement décimal (infini cette fois) le développement

$$m + 0, d_1 \cdots (d_N - 1)\ \overline{9}$$

On fait ici surgir une notion d'infini dans l'écriture même des nombres aussi simples que les fractions.

Si l'on exclut les développements décimaux qui se terminent par une répétition infinie de 9, on peut considérer que l'ensemble des fractions est en correspondance bijective avec l'ensemble des développements décimaux du type (††), les développements se terminant par une infinité de 9 étant interdits.

L'addition et la multiplication des nombres décimaux se fait en posant les additions et multiplications comme habituellement (attention aux retenues, y compris pour la partie entière !). En revanche, addition et multiplication des nombres rationnels introduits de cette manière sont des opérations plus délicates, à cause du problème des retenues (essayez d'écrire l'algorithme permettant de calculer la somme de deux développements décimaux, vous verrez la difficulté, liée au fait que pour poser une addition on commence par la droite et que les développements décimaux sont en général illimités à droite !). Nous y reviendrons.

Remarquons pour en finir avec le corps des nombres rationnels que l'écriture d'une fraction a/b en fraction continue ne fait pas intervenir, elle, de développement infini, mais seulement un développement fini (on développera par exemple $22/7$ en fraction continue et l'on comparera avec le développement décimal).

Terminons ici cette section sur l'ensemble $\mathbb{Q}$ des nombres rationnels en expliquant pourquoi $\mathbb{Q}$ (de par sa construction abstraite proposée dans le point **1** ci-dessus) est un ensemble dénombrable, c'est-à-dire en bijection avec $\mathbb{N}$. Comme $\mathbb{Z}$ est dénombrable et $\mathbb{N} \setminus \{0\}$ aussi, on peut numéroter les éléments de $\mathbb{Z} \times (\mathbb{N} \setminus \{0\})$; en effet, on peut numéroter (par exemple comme suit) les éléments de $\mathbb{N} \times (\mathbb{N} \setminus \{0\})$:

$$\begin{array}{rcl} x_0 & = & (0,1) \\ x_1 & = & (1,1)\,,\ x_2 = (0,2)\,, \\ x_3 & = & (2,1)\,,\ x_4 = (1,2)\,,\ x_5 = (0,3)\,; \\ & & \textit{etc.} \end{array}$$

on peut aussi numéroter (sur le même principe) les éléments de $\{-n\,;\, n \in \mathbb{N}^*\} \times \mathbb{N}^*$ et par conséquent, en intercalant par exemple les deux numérotations, numéroter les éléments de

$$\mathbb{Z} \times \mathbb{N}^* = (\mathbb{N} \times \mathbb{N}^*) \cup \Big(\{-n\,;\, n \in \mathbb{N}^*\} \times \mathbb{N}^*\Big)\,.$$

Ceci nous autorise à numéroter les éléments de $\mathbb{Q}$ (on "saute" bien sûr les répétitions !). L'ensemble des nombres rationnels $\mathbb{Q}$ est donc dénombrable.

2.2.2 Une approche de l'ensemble des nombres réels

On introduit la notion de nombre réel ainsi :

Définition 2.2. Un nombre réel est un développement décimal illimité

$$x = m + 0, d_1 d_2 d_3 \cdots\,,$$

où m est un élément de $\mathbb{Z}$ et les d_j, $j = 1, 2, ...$, des chiffres pris dans la liste $\{0, ..., 9\}$.

L'ensemble des nombres réels contient $\mathbb{Q}$ si l'on identifie un nombre rationnel à son unique développement ne se terminant pas par une succession infinie de 9.

Un nombre réel correspond donc à la donnée d'un entier m et d'une suite de chiffres entre 0 et 9, c'est-à-dire d'une application $d : \mathbb{N} \to \{0, ..., 9\}$.

On a un ordre total sur l'ensemble des nombres réels, prolongeant l'ordre sur $\mathbb{Q}$. On note cet ordre $\leq$. écrire $x < y$ signifie $x \leq y$ et $x \neq y$. Voici comment est défini cet ordre :

Définition 2.3. Soient $x = m + 0, d_1 d_2 \cdots$ et $y = m' + 0, d'_1 d'_2 \cdots$ deux nombres réels ; on dit que $x \leq y$ si et seulement si $m \leq m'$ et si la suite de chiffres $d_1, d_2, ..., d_N, ...$ précède la suite $d'_1, d'_2, ..., d'_N, ...$ pour l'ordre lexicographique sur $\{0, ..., 9\}$.

La *partie entière* du nombre réel $x = m + 0, d_1 d_2 \cdots$ est par définition le plus grand entier $E(x)$ inférieur ou égal à x ; cet entier vaut donc m si l'un des d_j au moins est différent de 9 ; sinon $x = m + 1$ et $E(x) = x = m + 1$.

2.2.3 Suites de nombres réels

Définition 2.4. Une suite de nombres réels est par définition une application u de $\mathbb{N}$ dans $\mathbb{R}$.

La suite peut aussi démarrer au cran $n_0 \in \mathbb{N} \setminus \{0\}$; c'est alors une application de $\mathbb{N} \setminus \{0, ..., n_0 - 1\}$ dans $\mathbb{R}$ comme par exemple la suite définie par

$$u_n = \frac{1}{n}$$

pour tout entier $n \geq 1$.

Convention de notation. Si $u : n \mapsto u(n)$ est une telle suite de nombres réels, on note souvent u_n plutôt que $u(n)$ (on appelle ce nombre réel le *terme général* de la suite) et la suite u est aussi notée $(u_n)_n$ ou $(u_n)_{n \geq n_0}$ si l'on veut préciser le cran où elle "démarre".

ATTENTION ! Il vaut veiller à ne pas confondre la suite $(u_n)_n$ (qui est une fonction de $\mathbb{N}$ dans $\mathbb{R}$) avec l'ensemble

$$\{u_n \, ; \, n \in \mathbb{N}\}$$

qui lui est un sous-ensemble de $\mathbb{R}$! Dans la notion de suite, on exploite le fait que $\mathbb{N}$ est ordonné et que l'on puisse "lister" les entiers en passant d'un entier à son successeur. L'ordre est important pour la notion de suite alors que

$$\{u_n \, ; \, n \in \mathbb{N}\}$$

est un grand sac dans lequel on a mis "en vrac" toutes les valeurs réelles prises par la suite.

Définition 2.5. On dit que la suite $(x_n)_n$, où $x_n = m_n + 0, d_{n,1} d_{n,2} \cdots$ converge vers le nombre réel $x = m + 0, d_1 d_2 \cdots$ (on ajoute parfois "lorsque n tend vers l'infini", sinon ceci est sous-entendu), si la suite $(m_n)_n$ stationne, pour n assez grand, à la valeur m et si, pour chaque $p \in \mathbb{N} \setminus \{0\}$, la suite $(d_{n,p})_p$, où $d_{n,p}$ est la p-ème décimale de x_n, stationne, pour n assez grand, à la valeur d_p (correspondant à la p-ème décimale de x). On dit aussi que x est *limite* de la suite $(x_n)_n$.

Ainsi, tout nombre réel est limite d'une suite de nombres rationnels (même de nombres rationnels décimaux) ; si

$$x = m + 0, d_1 d_2 \cdots d_N \cdots ,$$

on pose, pour $n = 0$, $x_0 = m$ et, pour $n \geq 1$,

$$x_n = m + 0, d_1 \cdots d_{n-1} d_n \overline{0} \, .$$

La suite $(x_n)_n$ de nombres décimaux ainsi construite converge vers x.

La définition de la convergence exclut qu'une même suite de nombres réels puisse avoir deux limites distinctes.

Remarque. Si $(x_n)_{n\geq n_0}$ est une suite convergente, il existe a et b dans $\mathbb{R}$ tels que

$$\forall n \geq n_0\,, \qquad a \leq x_n \leq b\,.$$

En effet, si m est la valeur où stationne la suite $(m_n)_n$, il vient un moment où tous les x_n sont tels que $m \leq x_n \leq m+1$.

L'ordre sur $\mathbb{R}$ est compatible avec le passage à la limite : si $(x_n)_n$ converge vers x, $(y_n)_n$ vers y et que $x_n \leq y_n$ pour tout $n \in \mathbb{N}$, alors $x \leq y$.

L'ensemble des nombres réels possède la propriété essentielle suivante :

Proposition 2.1. *Toute suite croissante de nombres réels majorée par un nombre réel M est convergente vers un nombre réel x tel que $x \leq M$. Toute suite décroissante de nombres réels minorée par un nombre réel m est convergente vers un nombre réel x tel que $m \leq x$.*

Preuve. On fait la preuve pour les suites croissantes majorées. Soit $(x_n)_n$ une telle suite. Posons

$$x_n = m_n + 0, d_{n,1}d_{n,2}\cdots d_{n,p}\cdots$$

Les parties entières des x_n forment une suite croissante et majorée $(m_n)_n$ de nombres entiers ; cette suite stationne donc à la borne supérieure m de l'ensemble $A_0 := \{m_n\,;\, n \in \mathbb{N}\}$ des valeurs de la suite. La suite $(d_{n,1})_n$ est une suite croissante de nombres entre 0 et 9 qui stationne donc, quand n est grand, à la borne supérieure d_1 du sous-ensemble majoré de $\mathbb{N}$ défini par $A_1 := \{d_{n,1}\,;\, n \in \mathbb{N}\}$. Á partir du cran N_1 où cette suite $(d_{n,1})_n$ stationne, la suite $(d_{n,2})_n$ devient une suite croissante qui elle aussi finit par stationner sur un chiffre d_2 entre 0 et 9 ; et ainsi de suite, on finit par montrer que, pour chaque indice $p \in \mathbb{N}^*$, la suite $(d_{n,p})_n$ finit par être croissante au delà d'un certain cran N_p et donc par stationner sur un chiffre d_p entre 0 et 9. Le nombre réel

$$x = m + 0, d_1 d_2 \cdots d_p \cdots$$

est ainsi la limite de la suite $(x_n)_n$. □

2.2.4 Les opérations sur $\mathbb{R}$

Armés de la proposition 2.1, nous pouvons définir la somme et le produit de deux nombres réels $x = m + 0, d_1 d_2 \cdots$ et $y = m' + 0, d'_1 d'_2 \cdots$.

Nous savons additionner et multiplier depuis l'école primaire deux nombres décimaux (on pose les opérations, ça marche bien car les développements sont finis à droite et que l'on travaille à partir de la droite pour ajouter et multiplier en faisant attention aux retenues).

Pour le cas général (où x et y sont deux réels quelconques, on se ramène en fait à additionner les décimaux.

Chacun des nombres x et y est en effet limite croissante d'une suite de nombres décimaux ; le nombre x est limite de la suite $(x_n)_n$, le nombre y est limite de la suite $(y_n)_n$ (x_n est le nombre décimal obtenu en mettant à zéro toutes les décimales de x d'ordre strictement plus grand que n, y_n est le nombre décimal obtenu en mettant à zéro toutes les décimales de y d'ordre strictement plus grand que n).

La suite $(x_n + y_n)_n$ est une suite croissante majorée par $m + m' + 2$, elle est donc convergente vers un nombre réel que l'on convient d'appeler $x + y$.

Pour définir le produit de x et y, on définit d'abord le produit de x et de z pour un nombre décimal z fixé. La suite $x_n z$ est une suite, soit croissante majorée, soit décroissante minorée (cela dépend du signe de z) de nombres décimaux. Cette suite converge vers un nombre réel noté xz. Ensuite, on remarque que la suite $(xy_n)_n$ est une suite de nombres réels soit croissante majorée, soit décroissante minorée, qui converge donc vers un nombre que l'on convient d'appeler xy.

On définit donc ainsi une addition et une multiplication sur l'ensemble des nombres réels qui prolongent l'addition et la multiplication des nombres rationnels. Ces deux opérations sont compatibles avec l'ordre au sens suivant :

$$\begin{aligned} ((a \leq b) \wedge (c \leq d)) &\implies a + c \leq b + d \\ ((a \leq b) \wedge (c \geq 0)) &\implies ac \leq bc \\ ((a \leq b) \wedge (c \leq 0)) &\implies bc \leq ac \end{aligned}$$

et $(\mathbb{R}, +, \times)$ est encore un corps commutatif pour ces deux opérations.

Pour cet ordre total et l'addition sur $\mathbb{R}$, l'ensemble des nombres réels obéit à la *propriété d'Archimède* (on dit aussi que le corps ordonné $\mathbb{R}$ est *archimédien*).

Proposition 2.2. *Soient x et y deux nombres réels avec $x > 0$. Il existe un entier N tel que $Nx > y$.*

Preuve. Il existe un nombre décimal $b/10^n$ avec $b \in \mathbb{N} \setminus \{0\}$ et $0 < b/10^n < x$. D'autre part, soit a la partie entière de y. La division euclidienne de $a + 1$ par b donne $a + 1 = qb + r$ avec $r \in \{0, ..., b - 1\}$. Si $N = 10^n(q + 1)$, on a $bN/10^n = b(q + 1) \geq a + 1$ et l'on a donc aussi $Nx \geq a + 1 > y$. □

2.2.5 Le lemme des "gendarmes"

Définition 2.6. Deux suites $(x_n)_n$ et $(y_n)_n$ de nombres réels sont dites *adjacentes* si l'on a, pour tout $n \in \mathbb{N}$,

$$x_n \leq x_{n+1} \leq y_{n+1} \leq y_n$$

et si de plus $(y_n - x_n)_n$ tend vers le nombre réel 0 lorsque n tend vers l'infini.

Exemple. Par exemple, les deux suites $(x_n)_{n\geq 1}$ et $(y_n)_{n\geq 1}$ de nombres rationnels définies par

$$x_n = 1 - \frac{1}{2} + \cdots + \frac{1}{2n-1} - \frac{1}{2n} \qquad y_n = x_n + \frac{1}{2n+1}$$

(pour $n \geq 1$) sont adjacentes (à vérifier en exercice). Même chose pour les deux suites $(u_n)_{n\geq 1}$ et $(v_n)_{n\geq 1}$ de nombres rationnels définies par

$$u_n = 4\Big(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots + \frac{1}{4n-3} - \frac{1}{4n-1}\Big) \qquad v_n = u_n + \frac{4}{4n+1}\,.$$

L'ensemble des nombres réels obéit au *critère des suites adjacentes* ou encore *lemme des "gendarmes"* :

Proposition 2.3. *Soient $(x_n)_n$ et $(y_n)_n$ deux suites adjacentes de nombres réels. Les deux suites convergent vers une limite commune dans $\mathbb{R}$.*

Preuve. La suite $(x_n)_n$ est croissante et majorée par y_0, donc convergente vers une limite x. La suite $(y_n)_n$ est décroissante et minorée par x_0, donc converge vers une limite y. Comme $x_n \leq y_n$, on a $x \leq y$. La suite $(y_n - x_n)_n$ est une suite de nombres réels positifs tendant vers 0. Si

$$\begin{aligned} x_n &= m_n + 0, d_{n,1} d_{n,2}, ..., d_{n,p} \cdots \\ y_n &= m'_n + 0, d'_{n,1} d'_{n,2}, ..., d'_{n,p} \cdots \end{aligned}$$

la suite $(m_n - m'_n)_n$ stationne pour n grand à la valeur 0, ainsi que chaque suite $(d_{n,p} - d'_{n,p})_n$ pour $p = 1, 2, ...$ Les nombres réels x et y limites de ces deux suites sont donc égaux. □

Exemple 1. Les suites $(x_n)_n$ et $(y_n)_n$ de l'exemple précédent convergent vers une limite commune. Il en est de même pour les suites $(u_n)_n$ et $(v_n)_n$. Il se trouve que la limite commune des suites $(u_n)_n$ et $(v_n)_n$ est un nombre irrationnel jouant un rôle que l'on verra capital dans la section suivante (consacrée à l'introduction de $\mathbb{C}$), le nombre π, correspondant à la moité du périmètre du cercle de rayon 1.

Exemple 2. Des tests numériques (sous Mathematica) montrent aisément que la convergence des suites $(u_n)_n$ et $(v_n)_n$ de l'exemple 1 vers π est une convergence très lente ; ce sont des valeurs très grandes de n qui permettent de voir surgir les premières décimales de π (voir le test effectué en cours sous Mathematica en prenant $n = 100$, puis $n = 1000$, enfin $n = 10000$). Notons d'ailleurs que l'erreur entre la limite pressentie (en l'occurrence π) et le nombre rationnel u_n est majorée par $v_n - u_n = 4/(4n+1)$ (on cherchera en exercice à le justifier). Il existe des procédés autrement plus rapides pour calculer les décimales de π, tel celui que proposa le mathématicien anglais John Machin (1680-1752) qui remarqua (en exploitant des formules de trigonométrie que l'on évoquera au chapitre suivant) que les deux suites (encore adjacentes, on le justifiera en exercice) de termes généraux respectifs

$$\begin{aligned} \tilde{u}_n &= 4\sum_{k=0}^{2n-1}(-1)^k \frac{4(1/5)^{2k+1} - (1/239)^{2k+1}}{2k+1} \\ \tilde{v}_n &= 4\sum_{k=0}^{2n}(-1)^k \frac{4(1/5)^{2k+1} - (1/239)^{2k+1}}{2k+1} \end{aligned}$$

convergent elles aussi toutes les deux vers π ; ceci lui permit d'être le premier à calculer les cent premières décimales du nombre π. On remarquera cette fois que l'erreur entre π et son approximation "inférieure" $\tilde{u}_n$ est majorée par

$$4\frac{4(1/5)^{4n+1} - (1/239)^{4n+1}}{4n+1},$$

ce qui est une quantité négligeable comparée à la quantité $4/(4n+1)$ contrôlant supérieurement l'erreur entre π et son approximation inférieure u_n dans l'exemple 1 précédent. On fera le test de la convergence des deux suites adjacentes $(\tilde{u}_n)_n$ et $(\tilde{v}_n)_n$ en cours (sous Mathematica). Le calcul (fait sous machine) de $\tilde{u}_{100}$ et de $\tilde{v}_{100}$ fait apparaître deux fractions ayant même développement décimal au moins jusqu'à l'ordre 100 ; les 100 décimales ainsi obtenues sont donc nécessairement (de par cet effet de "pincement" lié au critère de convergence des suites adjacentes) les cent premières décimales du nombre irrationnel π.

2.2.6 Borne supérieure, borne inférieure d'un sous-ensemble de $\mathbb{R}$

Si A est un sous-ensemble majoré de $\mathbb{Q}$, A n'admet pas nécessairement de plus petit majorant (pour l'ordre de $\mathbb{Q}$), comme on l'a vu pour

$$A = \{x \in \mathbb{Q}\,;\, x^2 \leq 2\}\,.$$

En revanche, $\mathbb{R}$ (avec son ordre prolongeant l'ordre sur $\mathbb{Q}$) est le plus petit ensemble contenant $\mathbb{Q}$ et ayant la propriété de la borne supérieure (ou de la borne inférieure).

Proposition 2.4. *Soit A un sous-ensemble majoré non vide de $\mathbb{R}$. Alors l'ensemble des majorants de A dans $\mathbb{R}$ admet un plus petit élément dans $\mathbb{R}$, dit borne supérieure de A. Soit B un sous-ensemble minoré non vide de $\mathbb{R}$. Alors l'ensemble des minorants de B admet un plus grand élément, dit borne inférieure de B.*

Remarque importante. On peut caractériser la borne supérieure b d'un sous-ensemble non vide majoré A de $\mathbb{R}$ par les deux clauses :
- b est un majorant de A ($\forall x \in A\,, x \leq b$) ;
- si $y < b$, il existe $x \in A$ avec $y < x \leq b$ (tout réel y strictement inférieur à b cesse d'être un majorant de A).

De même, on peut caractériser la borne inférieure a d'un sous-ensemble non vide majoré B de $\mathbb{R}$ par les deux clauses :
- a est un minorant de B ($\forall x \in B\,, a \leq x$) ;
- si $a < y$, il existe $x \in B$ avec $a \leq x < y$ (tout réel y strictement supérieur à a cesse d'être un minorant de B).

Preuve de la proposition 2.4. On va montrer que tout sous-ensemble non vide et majoré A de $\mathbb{R}$ admet un plus petit majorant en utilisant le critère des suites adjacentes (proposition 2.3).

On fixe $n \in \mathbb{N}$ et on considère le sous-ensemble A_n de $\mathbb{Z}$ constitué des entiers relatifs k tels que $\dfrac{k}{10^n}$ majore A (comme A est majoré, le sous-ensemble A_n est non vide du fait que $\mathbb{R}$ est archimédien). Le sous-ensemble A_n est minoré : en effet, A est non vide et contient un élément x_0 ; si $k \in A_n$, on a $x_0 \times 10^n \leq k$; A_n est donc minoré (dans $\mathbb{Z}$) par la partie entière de $x_0 \times 10^n$. L'ensemble A_n admet donc un plus petit élément a_n et l'on pose

$$\begin{aligned} x_n : &= \frac{a_n - 1}{10^n} \\ y_n : &= \frac{a_n}{10^n}\,. \end{aligned}$$

Il est facile de voir que les suites $(x_n)_n$ et $(y_n)_n$ ainsi construites sont adjacentes, donc convergent vers une limite commune b. Comme y_n majore A, b majore A (puisque le passage à la limite est compatible avec l'ordre). Si b' était un majorant de A plus petit que b et distinct de b, on pourrait trouver un des nombres x_n (pour n assez grand) avec $b' < x_n \leq b$. Mais il y a toujours un élément de A entre x_n et y_n, donc au delà de x_n. Donc b' ne majore pas A et l'on a une contradiction. Nous avons ainsi prouvé par l'absurde que b était bien le plus petit majorant de A.

La preuve est identique pour montrer que tout sous-ensemble de $\mathbb{R}$ minoré admet un plus grand minorant (une borne inférieure). □

On introduit un instrument fondamental de mesure, la *valeur absolue*.

Définition 2.7. La valeur absolue $|x|$ d'un nombre réel x est par définition le plus grand des deux nombres x et $-x$, ou encore la borne supérieure de l'ensemble $\{x, -x\}$.

La valeur absolue obéit aux deux règles suivantes :

$$\begin{aligned} &\forall x, y \in \mathbb{R}\,, && |xy| = |x| \times |y| \\ &\forall x, y \in \mathbb{R}\,, && |x + y| \leq |x| + |y|\,. \end{aligned}$$

La seconde de ces règles, dite *inégalité triangulaire*, a pour conséquence la troisième règle suivante :

$$\forall x, y \in \mathbb{R}\,, \ \Big||x| - |y|\Big| \leq |x - y|\,.$$

Cette notion de valeur absolue nous permet de formuler de manière différente l'assertion :

«*La suite de nombres réels* $(x_n)_n$ *converge vers le nombre réel* x»

On a en effet la proposition suivante :

Proposition 2.5. *Soit $(x_n)_n$ une suite de nombres réels. La suite $(x_n)_n$ converge vers le nombre réel x si et seulement si :*

$$\forall \epsilon > 0\,, \qquad \exists N(\epsilon) \in \mathbb{N}\,, \qquad (n \geq N(\epsilon)) \Longrightarrow (|x_n - x| \leq \epsilon)\,. \tag{$*$}$$

Preuve. Pour chaque $\epsilon > 0$, il existe (puisque $\mathbb{R}$ est archimédien) un entier $N \in \mathbb{N}$ tel que $10^{-N} \leq \epsilon$. Si la suite $(x_n)_n$ converge vers x, alors pour n assez grand (dépendant de N, donc de ϵ), x_n et x ont les mêmes décimales jusqu'à l'ordre N. La valeur absolue $|x_n - x|$ est donc majorée par 10^{-N}, donc par ϵ. On vient donc de vérifier que si la suite (x_n) converge vers x, la clause $(*)$ est remplie.

Réciproquement, si $|x_n - x| \leq 10^{-N-1}$, x_n et x ont les mêmes décimales au moins jusqu'à l'ordre N ; si la clause $(*)$ est remplie, la suite $(x_n)_n$ converge donc vers x. □

Cette nouvelle caractérisation de la convergence permet de prouver (ce dont on pouvait se douter) que la prise de limite, déjà compatible avec l'ordre, est aussi compatible avec les opérations de prise de valeur absolue, d'addition, de multiplication, et de prise d'inverse sur $\mathbb{R}$.

Proposition 2.6. *Soient $(x_n)_n$ et $(y_n)_n$ deux suites de nombres réels respectivement convergentes vers les nombres réels x et y ; alors la suite $(|x_n|)_n$ converge vers $|x|$, la suite $(x_n + y_n)_n$ converge vers $x + y$ et la suite $(x_n y_n)_n$ converge vers xy. De plus, si $x \neq 0$, alors $x_n \neq 0$ pour $n \geq n_0$ et la suite $(1/x_n)_{n \geq n_0}$ converge vers $1/x$.*

Preuve. Pour le premier point, on utilise l'inégalité

$$||x_n| - |x|| \leq |x_n - x|$$

et la proposition 2.5.

Pour ce qui est de la somme, on remarque que

$$|x + y - x_n - y_n| \leq |x_n - x| + |y_n - y|$$

(inégalité triangulaire). Si $\epsilon > 0$, il existe $N(\epsilon)$ tel que

$$n \geq N(\epsilon) \Longrightarrow \Big((|x_n - x| \leq \epsilon/2) \wedge (|y_n - y| \leq \epsilon/2)\Big)\,.$$

En additionnant, on trouve que pour $n \geq N(\epsilon)$, $|x_n + y_n - x - y| \leq 2\epsilon/2 = \epsilon$.

Pour ce qui est du produit, on remarque que si les suites $(x_n)_n$ et $(y_n)_n$ convergent, il existe M tel que, pour tout $n \in \mathbb{N}$, $|x_n| \leq M$ et $|y_n| \leq M$ et, par passage à la limite $|x| \leq M$, $|y| \leq M$. On a aussi, par inégalité triangulaire

$$|x_n y_n - xy| \leq |x_n y_n - x_n y| + |x_n y - xy| \leq M(|y_n - y| + |x_n - x|)\,. \tag{$\dagger$}$$

Or, si $\epsilon > 0$, il existe $N(\epsilon)$ tel que

$$n \geq N(\epsilon) \Longrightarrow \Big((|x_n - x| \leq \epsilon/2M) \wedge (|y_n - y| \leq \epsilon/2M)\Big)\,.$$

En reportant dans $(\dagger)$, on trouve bien que pour $n \geq N(\epsilon)$, $|x_n y_n - xy| \leq \epsilon$, ce que l'on voulait pour assurer la convergence de $(x_n y_n)_n$ vers xy d'après la proposition 2.5.

Pour ce qui est de l'inverse enfin, on remarque que pour n assez grand $|x_n| \geq |x|/2$ car $(|x_n|)_n$ converge vers $|x|$. Ensuite, on remarque que

$$\Big|\frac{1}{x_n} - \frac{1}{x}\Big| = \frac{|x_n - x|}{|x_n|\,|x|} \leq \frac{2}{|x|^2}\,|x_n - x|\,,$$

quantité que l'on peut rendre inférieure ou égale à ϵ dès que $|x_n - x| \leq \epsilon |x|^2/2$. □

2.2.7 Intervalles de $\mathbb{R}$; la propriété des segments emboîtés ; non dénombrabilité de $\mathbb{R}$

Soient a et b deux nombres réels avec $a < b$. On définit quatre sous-ensembles de $\mathbb{R}$, dits *intervalles*, impliquant ces deux nombres a et b.

– l'*intervalle ouvert*
$$]a,b[:=\{x\in\mathbb{R}\,;\,a<x<b\}\,;$$
– l'*intervalle fermé* (ou *segment*)
$$[a,b]:=\{x\in\mathbb{R}\,;\,a\leq x\leq b\}\,;$$
– les deux intervalles *semi-ouverts* (respectivement à gauche et à droite)
$$]a,b]:=\{x\in\mathbb{R}\,;\,a<x\leq b\}$$
et
$$[a,b[:=\{x\in\mathbb{R}\,;\,a\leq x<b\}\,.$$

On peut étendre ces définitions au cas $a = b$. Dans ce cas, on convient que le segment $[a, a]$ est le singleton $\{a\}$ tandis que les trois autres intervalles sont vides. Ces intervalles sont dits *bornés* car ils admettent tous (lorsqu'ils sont non vides) une borne supérieure (b) et une borne inférieure (a). On définit par extension les intervalles non bornés

$$\begin{array}{rcl}]-\infty,b[& := & \{x\in\mathbb{R}\,;\,x<b\}\\ ]-\infty,b] & := & \{x\in\mathbb{R}\,;\,x\leq b\}\\ ]a,+\infty[& := & \{x\in\mathbb{R}\,;\,x>a\}\\ [a,+\infty[& := & \{x\in\mathbb{R}\,;\,x\geq a\}\,.\end{array}$$

Le premier et le troisième sont dits *ouverts*, les deux autres sont dits *fermés*.

Tous ces sous-ensembles de $\mathbb{R}$ sont *convexes*, au sens suivant : un sous-ensemble E de $\mathbb{R}$ est convexe et et seulement si, pour tout couple (x, y) d'éléments de E, le segment $[x, y]$ est inclus dans E.

Si I est un intervalle de $\mathbb{R}$ (de l'un des types suivants, borné ou non-borné), on appelle *intérieur* de I (et on note I°) le nouvel intervalle obtenu en retirant à I les bornes (supérieure ou inférieure) qu'il contient éventuellement. On appelle *adhérence* de I (et on note $\overline{I}$) le nouvel intervalle obtenu en ajoutant à I les bornes (supérieure ou inférieure) qu'il ne contenait éventuellement pas. On a donc $I^\circ \subset I \subset \overline{I}$ pour tout intervalle de $\mathbb{R}$. L'intérieur d'un intervalle non vide (par exemple $[a, a]$) peut fort bien être vide !

Une propriété importante de $\mathbb{R}$ est celle dite *segments emboîtés* :

Proposition 2.7. *Soit $([a_n, b_n])_n$ une suite de segments bornés de $\mathbb{R}$ non vides et emboîtés les uns dans les autres (c'est-à-dire $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$ pour tout $n \in \mathbb{N}$). Il existe au moins un point x appartenant à tous les intervalles $[a_n, b_n]$ pour tout $n \in \mathbb{N}$, ou encore, en termes ensemblistes,*

$$\bigcap_{n\in\mathbb{N}}[a_n,b_n]\neq\emptyset.$$

Preuve. La suite $(a_n)_n$ est une suite de nombres réels croissante majorée (par b_0), donc convergente d'après la proposition 2.1 vers un nombre réel x. La suite (b_n) est une suite de nombres réels décroissante minorée (par a_0), donc convergente d'après la proposition 2.1 vers un nombre réel y. Comme $a_n \leq b_n$

pour tout $n \in \mathbb{N}$, le fait que la prise de limite soit compatible avec l'ordre implique $x \leq y$. Le segment $[x, y]$ est inclus dans tous les segments $[a_n, b_n]$. □

Remarque. L'intersection des segments emboîtés $[a_n, b_n]$ (pour $n \in \mathbb{N}$) se réduit à un seul point si et seulement si les suites $(a_n)_n$ et $(b_n)_n$ sont adjacentes, ce qui signifie que la suite $(b_n - a_n)_n$ des "longueurs" (on dit aussi *diamètres*) des segments emboîtés converge vers 0.

Une application : $\mathbb{R}$ n'est pas dénombrable. Prouvons ici par l'absurde, en utilisant le fait que $\mathbb{R}$ vérifie la propriété des segments emboîtés, que $\mathbb{R}$ (au contraire de $\mathbb{Q}$) n'est pas dénombrable, c'est-à-dire qu'il n'est pas possible de "lister" les nombres réels en une liste (sans répétitions)

$$x_0, x_1, x_2, ..., x_n, ...$$

Supposons que ceci soit possible. Soit $[a_0, b_0]$ un segment de $\mathbb{R}$, avec $b_0 - a_0 > 0$, ne contenant pas x_0 (il suffit de prendre $b_0 < x_0$ ou $a_0 > x_0$). On peut construire un segment $[a_1, b_1]$ tel que $b_1 - a_1 > 0$, inclus dans $[a_0, b_0]$, et ne contenant pas le point x_1 (discuter suivant que r_1 est intérieur à $[a_0, b_0]$, est une borne de ce segment, ou est un point de $\mathbb{R} \setminus [a_0, b_0]$) ; en continuant ainsi de suite, on trouve une suite de segments emboîtés $[a_n, b_n]$ telle que l'intersection de tous ces segments ne contienne aucun des nombres x_k, $k \in \mathbb{N}$, donc aucun nombre réel, ce qui est en contradiction avec le fait que $\mathbb{R}$ vérifie la propriété des segments emboîtés (et donc que l'intersection des $[a_n, b_n]$ pour $n \in \mathbb{N}$ soit non vide) !

Définition 2.8. On dit qu'un sous-ensemble E de $\mathbb{R}$ est un voisinage d'un point x si et seulement s'il existe un intervalle ouvert non vide $]x - \epsilon_1, x + \epsilon_2[$ tel que

$$x \in]x - \epsilon_1, x + \epsilon_2[\subset E ;$$

un sous-ensemble de $\mathbb{R}$ vide ou voisinage de chacun de ses points est dit *ouvert*. Un sous-ensemble de $\mathbb{R}$ est dit *fermé* si et seulement si son complémentaire est ouvert.

Un sous-ensemble de $\mathbb{R}$ n'a aucune raison d'être ouvert ou fermé ! Par exemple, $\mathbb{Q}$ n'est ni ouvert ni fermé car tout intervalle ouvert non vide de $\mathbb{R}$ contient toujours au moins un nombre irrationnel et un nombre rationnel, même en fait un nombre décimal.

Cependant, à tout sous ensemble E de $\mathbb{R}$, on peut associer un sous-ensemble ouvert E° inclus dans E (E° est d'ailleurs le plus grand ouvert – au sens de l'inclusion –inclus dans E) et un sous-ensemble fermé $\overline{E}$ contenant E (qui est le plus petit fermé – au sens de l'inclusion – contenant E). L'ensemble E° est dit *intérieur* de E, l'ensemble $\overline{E}$ est dit *adhérence* de E. Ces deux ensembles sont définis respectivement ainsi :

- un nombre x est dans E° si et seulement si E est un voisinage de x ;
- un nombre x est dans $\overline{E}$ si et seulement si x est limite d'une suite $(x_n)_n$ de points de E.

Il est très possible que E° soit vide (même si E ne l'est pas) ; en revanche, on a toujours $E \subset \overline{E}$. L'ensemble $\overline{E} \setminus E^\circ$ est appelé *frontière* de E ; la frontière de E est l'ensemble des points qui sont à la fois limite d'une suite $(x_n)_n$ de points de E et d'une suite $(y_n)_n$ de points de $\mathbb{R} \setminus E$: par exemple, la frontière de $\mathbb{R} \setminus \mathbb{Z}$ est exactement l'ensemble $\mathbb{Z}$, mais la frontière de $\mathbb{R} \setminus \mathbb{Q}$ est $\mathbb{R}$ tout entier car tout nombre réel s'approche à la fois par une suite de nombres décimaux et par une suite de nombres irrationnels (on expliquera pourquoi). De même, $\mathbb{Z}$ a pour intérieur l'ensemble vide, pour adhérence $\mathbb{Z}$ et donc pour frontière $\mathbb{Z}$, tandis que $\mathbb{Q}$ a toujours pour intérieur l'ensemble vide, pour adhérence $\mathbb{R}$ (tout nombre réel s'approche par une suite de décimaux), et donc pour frontière $\mathbb{R}$ tout entier ! Il faut donc se méfier parfois des intuitions en ce qui concerne la frontière qui peut (comme dans le cas $E = \mathbb{R} \setminus \mathbb{Q}$ ou $E = \mathbb{Q}$ dont la frontière est $\mathbb{R}$ tout entier) être beaucoup plus grande que l'ensemble et même le contenir !

2.2.8 La droite numérique achevée

Un sous-ensemble de $\mathbb{R}$ non majoré n'admet pas de borne supérieure (puisque l'ensemble des majorants est vide !) De même un sous-ensemble de $\mathbb{R}$ non minoré n'admet pas de borne inférieure. Pour pallier à ce fait, on complète $\mathbb{R}$ par deux éléments notés $-\infty$ et $+\infty$ afin de réaliser la *droite numérique achevée* $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$.

Remarque. Il existe une autre manière d'ajouter non pas deux, mais juste un point à l'infini à $\mathbb{R}$. On considère l'ensemble

$$S := \{(x, y) \in \mathbb{R}^2 \,;\, x^2 + y^2 = 1\} \setminus \{(0, 1)\}$$

qui est le cercle unité privé du pôle nord $(1, 0)$ dans le plan réel. L'application

$$f \;: (x, y) \longmapsto \frac{x}{1 - y}$$

réalise une bijection entre cet ensemble et $\mathbb{R}$ (on pourra s'exercer à faire un dessin comme celui fait en cours). On peut donc compléter $\mathbb{R}$ en ajoutant un point "à l'infini", en l'occurrence celui qui correspond au pôle nord $(0, 1)$ du cercle unité, seul point à ne pas avoir d'image par l'application f. Cependant, dans ce que nous ferons ici, nous ajouterons à $\mathbb{R}$ deux points à l'infini pour rendre compte de ce qui se passe pour les valeurs du passé infini et du futur infini, l'ensemble des nombres réels étant pensé, comme c'est souvent le cas en physique, comme l'axe des temps.

Définition 2.9. Une suite $(x_n)_n$ de nombres réels converge vers $+\infty$ si et seulement si

$$\forall A > 0\,,\ \exists N = N(A) \in \mathbb{N}\,, \quad (n \geq N(A)) \Longrightarrow (x_n \geq A)\,.$$

Une suite $(x_n)_n$ de nombres réels converge vers $-\infty$ si et seulement si

$$\forall A > 0\,,\ \exists N = N(A) \in \mathbb{N}\,, \quad (n \geq N(A)) \Longrightarrow (x_n \leq -A)\,.$$

Un sous-ensemble E de $\overline{\mathbb{R}}$ est un *voisinage de* $+\infty$ si et seulement s'il contient un sous-ensemble de la forme $]b, +\infty[\cup \{+\infty\}$. Un sous-ensemble E de $\overline{\mathbb{R}}$ est un *voisinage de* $-\infty$ si et seulement s'il contient un sous-ensemble de la forme $]-\infty, a[\cup \{-\infty\}$.

Voici quelques règles fondamentales (autres que celles énoncées dans la proposition 2.5) concernant la compatibilité des prises de limite (y compris vers $+\infty$ et $-\infty$) et les opérations d'addition et de multiplication :

$$\begin{aligned}
(x_n \longrightarrow 0) \wedge (\exists a, b \in \mathbb{R},\ \forall n\,, a \leq y_n \leq b) &\Longrightarrow (x_n y_n \longrightarrow 0)\\
(x_n \longrightarrow +\infty) &\Longrightarrow (\frac{1}{x_n} \longrightarrow 0_+)\\
(x_n \longrightarrow -\infty) &\Longrightarrow (\frac{1}{x_n} \longrightarrow 0_-)\\
(x_n \longrightarrow +\infty) \wedge (\exists a \in \mathbb{R},\ \forall n\,, y_n \geq a) &\Longrightarrow x_n + y_n \longrightarrow +\infty\\
(x_n \longrightarrow -\infty) \wedge (\exists b \in \mathbb{R},\ \forall n\,, y_n \leq b) &\Longrightarrow x_n + y_n \longrightarrow -\infty\\
(x_n \longrightarrow +\infty) \wedge (\exists c > 0,\ \forall n >> 0\,, y_n \geq c) &\Longrightarrow x_n y_n \longrightarrow +\infty\\
(x_n \longrightarrow -\infty) \wedge (\exists c > 0,\ \forall n >> 0\,, y_n \geq c) &\Longrightarrow x_n y_n \longrightarrow -\infty\\
(x_n \longrightarrow +\infty) \wedge (\exists c > 0,\ \forall n >> 0\,, y_n \leq -c) &\Longrightarrow x_n y_n \longrightarrow -\infty\\
(x_n \longrightarrow -\infty) \wedge (\exists c > 0,\ \forall n >> 0\,, y_n \leq -c) &\Longrightarrow x_n y_n \longrightarrow +\infty
\end{aligned}$$

(converger vers 0_+ signifie converger vers 0 en restant positif pour n assez grand, converger vers 0_- signifie converger vers 0 en restant négatif pour n assez grand ; la notation $\forall n >> 0$ se lit "*pour tout n assez grand*").

En revanche si $(x_n)_n$ est une suite tendant vers $+\infty$ et $(y_n)_n$ une suite tendant vers $-\infty$, on ne peut *a priori* rien dire de la suite $(x_n + y_n)_n$ (on dit que l'on est en face d'une *indétermination*). On dit aussi que la forme $+\infty - \infty$ est *une forme indéterminée*. Il faut travailler plus pour en dire quelque chose et lever cette indétermination : par exemple, c'est seulement en écrivant

$$n^2 - n = n^2(1 - \frac{1}{n})$$

que l'on peut conclure que $(n^2 - n)_n$ tend vers $+\infty$ lorsque n tend vers $+\infty$. Suivant ce même principe, une suite de terme général

$$a_0 n^p + a_1 n^{p-1} + \cdots + a_p \,,$$

où $p \in \mathbb{N} \setminus \{0\}$ et $a_0, ..., a_p$ sont des nombres réels avec $a_0 \neq 0$, tend vers $+\infty$ si $a_0 > 0$ et vers $-\infty$ si $a_0 < 0$ (lorsque n tend vers l'infini). Ceci se voir en remarquant

$$a_0 n^p + a_1 n^{p-1} + \cdots + a_p = a_0 n^p \Big(1 + \frac{a_1}{a_0}\,\frac{1}{n} + \cdots \frac{a_p}{a_0}\,\frac{1}{n^p}\Big) = a_0 n^p (1 + v_n) \,,$$

où la suite $(v_n)_n$ tend vers 0.

On retrouve le même problème d'indétermination à lever lorsque la suite $(x_n)_n$ converge vers $+\infty$ ou $-\infty$, que la suite $(y_n)_n$ converge vers 0, et que l'on veuille étudier la suite $x_n y_n$ (on dit que les formes $(+\infty) \times 0$ ou $(-\infty) \times 0$ sont des *formes indéterminées*). Par exemple, on ne peut pas sans travailler davantage conclure immédiatement au comportement lorsque n tend vers l'infini de la suite de terme général

$$x_n = \frac{1}{n} \times \log n$$

dont on montrera plus tard qu'elle converge vers 0.

2.3 Le plan $\mathbb{R}^2$ et les nombres complexes

2.3.1 Le plan $\mathbb{R}^2$

L'ensemble des couples (x, y) de nombres réels hérite d'une structure de groupe abélien avec l'opération d'addition définie par

$$(x_1, y_1) + (x_2, y_2) := (x_1 + x_2, y_1 + y_2) \,.$$

La "multiplication" "naïve"

$$(x_1, y_1) \times (x_2, y_2) := (x_1 \times x_2, y_1 \times y_2)$$

n'est pas une opération intéressante (pas de distributivité par rapport à l'addition par exemple) et on l'oubliera.

On dispose par contre aussi d'une action externe de $\mathbb{R}$ sur $\mathbb{R}^2$ de la manière suivante : à $a \in \mathbb{R}$ et (x, y) dans $\mathbb{R}^2$, on associe

$$a \cdot (x, y) := (a \times x, a \times y) = (ax, ay) \,.$$

Addition et multiplication externe vérifient les règles de compatibilité

$$\begin{aligned} a \cdot ((b \cdot (x, y))) &= (a \times b) \cdot (x, y) \\ (a + b) \cdot (x, y) &= a \cdot (x, y) + b \cdot (x, y) \\ a \cdot ((x_1, y_1) + (x_2, y_2)) &= a \cdot (x_1, y_1) + a \cdot (y_1, y_2) \\ 1 \cdot (x, y) &= (x, y) \,. \end{aligned}$$

On dit que $\mathbb{R}^2$ équipé de l'addition et de cette action externe du corps $\mathbb{R}$ dessus est un $\mathbb{R}$-*espace vectoriel* (on note cette structure $(\mathbb{R}^2, +, \cdot)$).

On note que $(\mathbb{R}, +, \times)$ est un aussi $\mathbb{R}$-espace vectoriel sur lui-même avec l'action externe $a \cdot x = a \times x$.

Définition 2.10. Une fonction l de $\mathbb{R}$ dans $\mathbb{R}$ est dite $\mathbb{R}$-*linéaire* si et seulement si L vérifie

$$\forall x, y \in \mathbb{R}\,,\ \forall a, b \in \mathbb{R}\,,\ l(a \times x + b \times y) = a \times l(x) + b \times l(y)\,.$$

Par analogie, une fonction L de $\mathbb{R}^2$ dans $\mathbb{R}^2$ est dite $\mathbb{R}$-*linéaire* si et seulement si

$$\forall (x_1, y_1), (x_2, y_2) \in \mathbb{R}^2\,,\ \forall a, b \in \mathbb{R}\,,\ L(a \cdot (x_1, y_1) + b \cdot (x_2, y_2)) = a \cdot L(x_1, y_1) + b \cdot L(x_2, y_2)\,.$$

Pour se donner une fonction $\mathbb{R}$-linéaire l de $\mathbb{R}$ dans lui-même, il suffit de se donner $l(1)$ car alors

$$l(x) = l(x \times 1) = x \times l(1)\,.$$

Pour se donner une fonction $\mathbb{R}$-linéaire L de $\mathbb{R}^2$ dans lui-même, il suffit de se donner $L(1,0) = (a, c)$ et $L(0,1) = (b, d)$; en effet, on a alors, pour tout (x, y) dans $\mathbb{R}^2$,

$$L(x, y) = x \cdot L(1,0) + y \cdot L(0,1) = \Big(ax + by, cx + dy\Big)\,.$$

On peut donc représenter l'application L par un tableau à deux lignes et deux colonnes, le tableau suivant

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\,,$$

que l'on appelle *matrice* de L relativement au système (l'ordre des éléments est important) $\mathcal{B} = [(1,0), (0,1)]$.

Proposition 2.6. *La composée $l = l_2 \circ l_1$ de deux applications $\mathbb{R}$-linéaires l_1 et l_2 de $\mathbb{R}$ dans $\mathbb{R}$ est une application linéaire l de $\mathbb{R}$ dans $\mathbb{R}$ avec $l(1) = l_1(1) \times l_2(1)$.*
La composée de deux applications $\mathbb{R}$-linéaires L_1 et L_2 de $\mathbb{R}^2$ dans $\mathbb{R}^2$ est une application linéaire $L = L_2 \circ L_1$ de $\mathbb{R}^2$ dans $\mathbb{R}^2$. La matrice de l'application L dans le système $\mathcal{B}$ s'écrit

$$\begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} \bullet \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} = \begin{pmatrix} a_2a_1 + b_2c_1 & a_2b_1 + b_2d_1 \\ c_2a_1 + d_2c_1 & c_2b_1 + d_2d_1 \end{pmatrix}\,.$$

RÈGLE DE CALCUL IMPORTANTE *Pour calculer le coefficient du tableau de $L = L_2 \circ L_1$ qui se trouve au carrefour de la ligne d'indice i ($i = 1, 2$) et de la colonne d'indice j ($j = 1, 2$), on multiplie "terme à terme" la i-ème ligne de la matrice de L_2 relativement au système $\mathcal{B}$ par la j-ème colonne de la matrice de L_1 relativement au système $\mathcal{B}$.*

On vérifiera que le produit $\bullet$ entre matrices 2×2 n'est pas commutatif (car la composition d'applications à laquelle ce produit correspond ne l'est pas puisqu'en général $L_2 \circ L_1 \neq L_1 \circ L_2$) ; en revanche la composition entre applications $\mathbb{R}$ linéaires de $\mathbb{R}$ dans lui-même est une opération commutative (car la multiplication est une opération commutative sur $\mathbb{R}$). Le produit $\bullet$ entre matrices 2×2 est en revanche une opération associative (comme la composition des applications).

On peut faire agir $\mathbb{R}$ de manière externe sur l'ensemble $\mathcal{M}_{2,2}(\mathbb{R})$ des matrices 2×2 à coefficients réels en posant, pour tout x dans $\mathbb{R}$ et toute matrice $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ de $\mathcal{M}_2(\mathbb{R})$,

$$x \cdot \begin{pmatrix} a & b \\ c & d \end{pmatrix} := \begin{pmatrix} x \times a & x \times b \\ x \times x & x \times d \end{pmatrix}\,.$$

On peut aussi définir une addition (commutative et associative) sur $\mathcal{M}_{2,2}(\mathbb{R})$ en posant

$$\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} + \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} = \begin{pmatrix} a_1 + a_2 & b_1 + b_2 \\ c_1 + c_2 & d_1 + d_2 \end{pmatrix} .$$

Additionner les matrices de L_1 et L_2 relativement au système $\mathcal{B}$ revient à écrire la matrice relativement au système $\mathcal{B}$ de l'application

$$(x, y) \longmapsto L_1(x, y) + L_2(x, y) .$$

Proposition 2.7. *L'ensemble $\mathcal{M}_{2,2}(\mathbb{R})$ équipé de l'addition et de la multiplication externe hérite d'une structure de $\mathbb{R}$-espace vectoriel.*

Preuve. On laisse la vérification en exercice. □

Exercice. Pour s'entraîner au calcul du produit de deux matrices, on vérifiera qu'une matrice

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

vérifie toujours la formule

$$A \bullet A - (a + d) \cdot A + (ad - bc) \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

On en déduira, si $ad - bc \neq 0$, qu'il existe une matrice B, à savoir la matrice

$$B := \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

telle que

$$A \bullet B = B \bullet A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} .$$

Les applications linéaires de $\mathbb{R}^2$ dans $\mathbb{R}^2$ jouant un rôle important en physique (suivant le principe de moindre action) sont celles qui préservent les angles (orientés) des figures (et par voie de conséquence préservent les formes des objets), si l'on convient d'y ajouter l'application nulle qui envoie tout point (x, y) sur le point $(0, 0)$. Ces applications sont les applications linéaires de $\mathbb{R}^2$ dans $\mathbb{R}^2$ obtenues en composant une homothétie de rapport $r \geq 0$ et de centre l'origine avec une rotation autour de $(0, 0)$ d'angle θ. La matrice d'une telle application dans le système $\mathcal{B}$ est

$$r \cdot \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} = a \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b \cdot \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

avec $a = r\cos\theta$ et $b = r\sin\theta$.

On remarque que deux applications linéaires du type précédent, de matrices respectivement

$$r_1 \cdot \begin{pmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{pmatrix} = \begin{pmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{pmatrix}$$

et

$$r_2 \cdot \begin{pmatrix} \cos\theta_2 & -\sin\theta_2 \\ \sin\theta_2 & \cos\theta_2 \end{pmatrix} = \begin{pmatrix} a_2 & -b_2 \\ b_2 & a_2 \end{pmatrix}$$

commutent pour la composition et que la matrice de leur composée est

$$\begin{pmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{pmatrix} \bullet \begin{pmatrix} a_2 & -b_2 \\ b_2 & a_2 \end{pmatrix} = \begin{pmatrix} a_1a_2 - b_1b_2 & -(a_1b_2 + b_1a_2) \\ a_1b_2 + b_1a_2 & a_1a_2 - b_1b_2 \end{pmatrix} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

avec $a = a_1a_2 - b_1b_2$ et $b = a_1b_2 + b_1a_2$.

2.3.2 Le corps $(\mathbb{C}, +, \times)$

Si x est un nombre réel, le nombre x^2 est toujours positif et il n'existe par conséquent aucun nombre réel tel que $x^2 + 1 = 0$. On peut exprimer ceci différemment en termes d'applications linéaires de $\mathbb{R}$ dans $\mathbb{R}$ en affirmant *qu'il n'existe pas d'application linéaire l de $\mathbb{R}$ dans $\mathbb{R}$ qui, composée avec elle-même, donne l'application linéaire $x \to -x$.*

Dans $\mathbb{R}^2$, nous disposons non plus d'un, mais de deux degrés de liberté et la situation est donc différente. Chercher à factoriser sous la forme $L \bullet L$ l'application linéaire $(x, y) \mapsto (-x, -y)$ se fait en cherchant une matrice 2×2 A à coefficients réels telle que

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

Si l'on se limite à trouver L de manière à ce que sa matrice relativement au système $\mathcal{B}$ soit de la forme

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}$$

on trouve deux applications linéaires L possibles (les rotations d'angle un demi-droit dans le sens des aiguilles d'une montre ou le sens inverse, dit aussi sens trigonométrique), celles dont les matrices relativement à $\mathcal{B}$ sont

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

Ces deux matrices A vérifient $A \bullet A = -\mathrm{Id}_{\mathbb{R}^2}$, équation qui formellement correspond à $X^2 + 1 = 0$.

Définition 2.11. On appelle nombre complexe toute matrice 2×2 de nombres réels de la forme

$$a \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + b \cdot i\,,$$

où $i = \sqrt{-1}$ désigne la matrice

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

On notera pour simplifier cette matrice $a + ib$ et l'on dira que a est la *partie réelle* du nombre complexe $a + ib$, b sa *partie imaginaire*; de plus le nombre complexe $a + ib$ est dit *affixe* du couple (a, b) de $\mathbb{R}^2$.

Proposition 2.8. *l'ensemble des nombres complexes $\mathbb{C}$ peut être équipé d'une addition (l'addition des matrices) et d'une multiplication (le produit des matrices) selon les règles*

$$\begin{aligned} (a_1 + ib_1) + (a_2 + ib_2) &= (a_1 + a_2) + i(b_1 + b_2) \\ (a_1 + ib_1) \times (a_2 + ib_2) &= (a_1a_2 - b_1b_2) + i(a_1b_2 + b_1a_2)\,. \end{aligned}$$

Muni de ces deux opérations $(\mathbb{C}, +, \times)$ hérite d'une structure de corps commutatif et l'identification entre a et $a + i0$ permet de considérer $\mathbb{R}$ avec ses deux opérations comme un sous-corps de $(\mathbb{C}, +, \times)$.

Preuve. On se contentera de montrer que tout nombre complexe $a + ib$ avec a et b non tous les deux nuls admet un inverse pour la multiplication. Pour cela, on remarque que

$$(a + ib) \times (a - ib) = a^2 + b^2\,,$$

ce qui permet de définir l'inverse de $a + ib$ par

$$\frac{1}{a + ib} = \frac{a - ib}{a^2 + b^2}\,.$$

Par exemple

$$\frac{1}{2+3i} = \frac{2-3i}{13} = \frac{2}{13} - \frac{3}{13}i\,.$$

Le reste des vérifications relève de la routine. □

2.3.3 Module et argument

Le nombre complexe $a+ib$ (si a et b ne sont pas tous les deux nuls) correspond, on l'a vu, à la composée d'une rotation autour de l'origine d'angle θ à déterminer et d'une homothétie de rapport $r \geq 0$ (aussi à déterminer). Pour trouver r et θ, on pose

$$a = r\cos\theta\,,\ b = r\sin\theta\,,$$

et l'on voit donc que

$$a^2 + b^2 = r^2(\cos^2\theta + \sin^2\theta) = r^2\,,$$

ce qui donne

$$r = \sqrt{a^2+b^2}\,.$$

Ce nombre $r = \sqrt{a^2+b^2}$ est dit *module* (les physiciens disent *amplitude*) du nombre complexe $a+ib$. On le note $|a+ib|$. On pose $|0| = 0$.

L'angle θ déterminé *via* ses lignes trigonométriques $\cos\theta$ et $\sin\theta$ (ou encore par un point sur le cercle du plan $\mathbb{R}^2$ de centre $(0,0)$ et de rayon 1) est dit *argument* du nombre complexe non nul $a+ib$ (les physiciens disent aussi *phase*). Si 2π désigne le périmètre du cercle unité, l'argument θ du nombre complexe non nul $z = a+ib$ est seulement déterminé à 2π-près. On dit qu'un nombre réel de la forme $\theta + 2k\pi$, avec $k \in \mathbb{Z}$, est une *détermination* de l'argument du nombre complexe non nul $z = r\cos\theta + ir\sin\theta$. La famille de toutes ces déterminations est notée $\arg z$. La détermination de l'argument de z comprise dans l'intervalle $[0, 2\pi[$ (qui est unique) est appelée *détermination principale de l'argument* du nombre complexe non nul z. On la note $\operatorname{Arg} z$.

L'écriture $z = r(\cos\theta + i\sin\theta)$ du nombre complexe z est dite *écriture trigonométrique de* z, tandis que l'écriture $z = a+ib$ est dite *écriture cartésienne* de z.

Si $z = a+ib = r\cos\theta + ir\sin\theta$ est un nombre complexe, le nombre $a-ib$ est dit *conjugué* de $a+ib$ et noté $\overline{z}$. On a les règles suivantes :

$$\begin{aligned}
|z| &= |\overline{z}| \\
\arg\overline{z} &= -\arg z\,, \forall z \in \mathbb{C}\setminus\{0\} \\
|z_1 z_2| &= |z_1| \times |z_2| \\
\overline{z_1 + z_2} &= \overline{z_1} + \overline{z_2} \\
\overline{z_1 z_2} &= \overline{z_1} \times \overline{z_2} \\
|z|^2 &= z \times \overline{z} \\
\frac{1}{z} &= \frac{\overline{z}}{|z|^2}\,,\ \forall z \in \mathbb{C}\setminus\{0\} \\
||z_1| - |z_2|| &\leq |z_1 - z_2| \\
|z_1 + z_2| &\leq |z_1| + |z_2|\,,
\end{aligned}$$

l'égalité dans l'une des deux inégalités ci-dessus n'ayant lieu que si l'un des z_i est nul ou si z_1 et z_2 sont non nuls et ont même argument.

Quant à la formule

$$\arg(z_1 z_2) = \arg z_1 + \arg z_2\,, \forall z_1, z_2 \in \mathbb{C} \setminus \{0\}\,,$$

il faut la manier avec soin : elle signifie que si θ_i, $i = 1, 2$, est une détermination de l'argument de z_i, $\theta_1 + \theta_2$ est une détermination de l'argument de $z_1 z_2$. C'est la même signification qu'il faut accorder à la formule

$$\arg \overline{z} = -\arg z\,, \forall z \in \mathbb{C} \setminus \{0\}\,.$$

2.3.4 La fonction exponentielle complexe et les formules de Moivre et d'Euler

Une fonction importante de $\mathbb{R}$ dans $]0, +\infty[$ est la fonction exponentielle ; cette fonction a du point de vue outil de calcul le mérite d'échanger les opérations d'addition et de multiplication, selon la règle :

$$\forall x_1\,, x_2 \in \mathbb{R}\,,\ \exp(x_1 + x_2) = \exp x_1 \times \exp x_2\,.$$

On note aussi, pour tout x réel, $\exp x = e^x$; le nombre $e = \exp 1$ est un nombre réel (dont on verra plus tard qu'il est irrationnel) appelé à jouer un rôle important.

Outre cette propriété d'échanger les deux opérations, l'application exponentielle est une application fondamentale autant en mathématiques qu'en physique car correspondant à des phénomènes d'évolution régis par une équation différentielle très simple dont nous parlerons au prochain chapitre.

De fait, la fonction exponentielle se prolonge en une fonction de $\mathbb{C}$ dans $\mathbb{C} \setminus \{0\}$ (notée aussi exp) et suivant la même règle d'échange des opérations d'addition et de multiplication, c'est-à-dire

$$\forall z_1\,, z_2 \in \mathbb{C}\,,\ \exp(z_1 + z_2) = \exp z_1 \times \exp z_2\,.$$

Définition 2.12. Si $z = x + iy$ est un nombre complexe écrit sous forme cartésienne, on pose

$$\exp(x + iy) := \exp x \times (\cos y + i \sin y)\,.$$

Le nombre complexe $\exp(x + iy)$ correspond donc à l'application linéaire de $\mathbb{R}^2$ dans $\mathbb{R}^2$ obtenue en composant l'homothétie de centre $(0, 0)$ et de rapport $\exp x$ avec la rotation autour de $(0, 0)$ d'angle orienté y. Ceci explique bien pourquoi l'on a la formule

$$\forall z_1\,, z_2 \in \mathbb{C}\,,\ \exp(z_1 + z_2) = \exp z_1 \times \exp z_2 \qquad (\dagger)$$

et donc en particulier

$$\forall z \in \mathbb{C}\,,\ \exp z \times \exp(-z) = 1\,,$$

d'où $\exp z \neq 0$.

Outre encore la propriété qu'elle a d'échanger les deux opérations, l'application exponentielle considérée comme une application de $\mathbb{C}$ dans $\mathbb{C} \setminus \{0\}$ est encore une application fondamentale autant en mathématiques qu'en physique car attachée aux phénomènes d'évolution dans le plan gouvernés par un système différentiel (vous verrez cela plus tard) ainsi qu'aux phénomènes temporels de nature ondulatoire (vous l'avez rencontrée en physique).

On note aussi $\exp z = e^z$.

L'image par l'application exponentielle de la droite du plan paramétrée par

$$t \longmapsto x_0 + iy_0 + t(u_0 + iv_0)$$

(c'est-à-dire passant par le point (x_0, y_0) et dirigée par le vecteur (u_0, v_0)) est soit un cercle de centre $(0, 0)$ et de rayon e^{x_0} (si $u_0 = 0$), soit (lorsque $u_0 \neq 0$) une spirale tourbillonnant depuis l'origine (sur laquelle elle s'écrase) jusqu'à l'infini (on expliquera pourquoi en exercice) ; ceci justifie pourquoi la fonction exponentielle s'avère un outil majeur pour modéliser les phénomènes tourbillonnaires (cyclones, tourbillons dans un écoulement turbulent ...) ; le cas $u_0 = 0$ correspond à une situation stable (on a affaire à un cercle et non une spirale), tandis que le cas $u_0 \neq 0$ correspond à la situation instable. On a représenté sur la figure ci-dessous les deux modèles de situation.

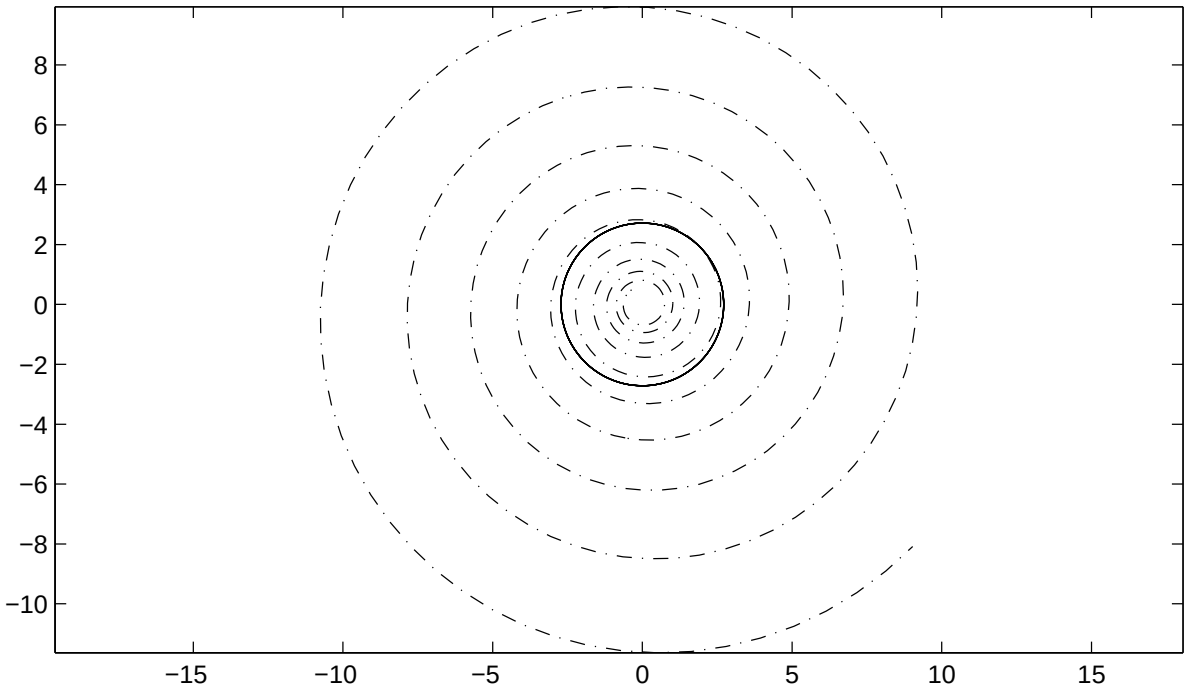

FIGURE 2.1 – L'image des droites par l'application exponentielle complexe : modèle stable $[x_0 = 1, y_0 = 0, u_0 = 0]$(en trait plein) et modèle instable $[x_0 = 1, y_0 = 0, u_0 = 1/2]$(en pointillés)

Sur la figure suivante, on a représenté l'image d'un cercle (en l'occurrence de rayon $r = 7$) par l'application exponentielle complexe ; on constatera que l'image du cercle de rayon 1 est une courbe sans point double, puis que la situation se complique au fur et à mesure que r augmente (des points doubles apparaîssent, on expliquera pourquoi, dès que le diamètre du cercle est au moins égal à 2π).

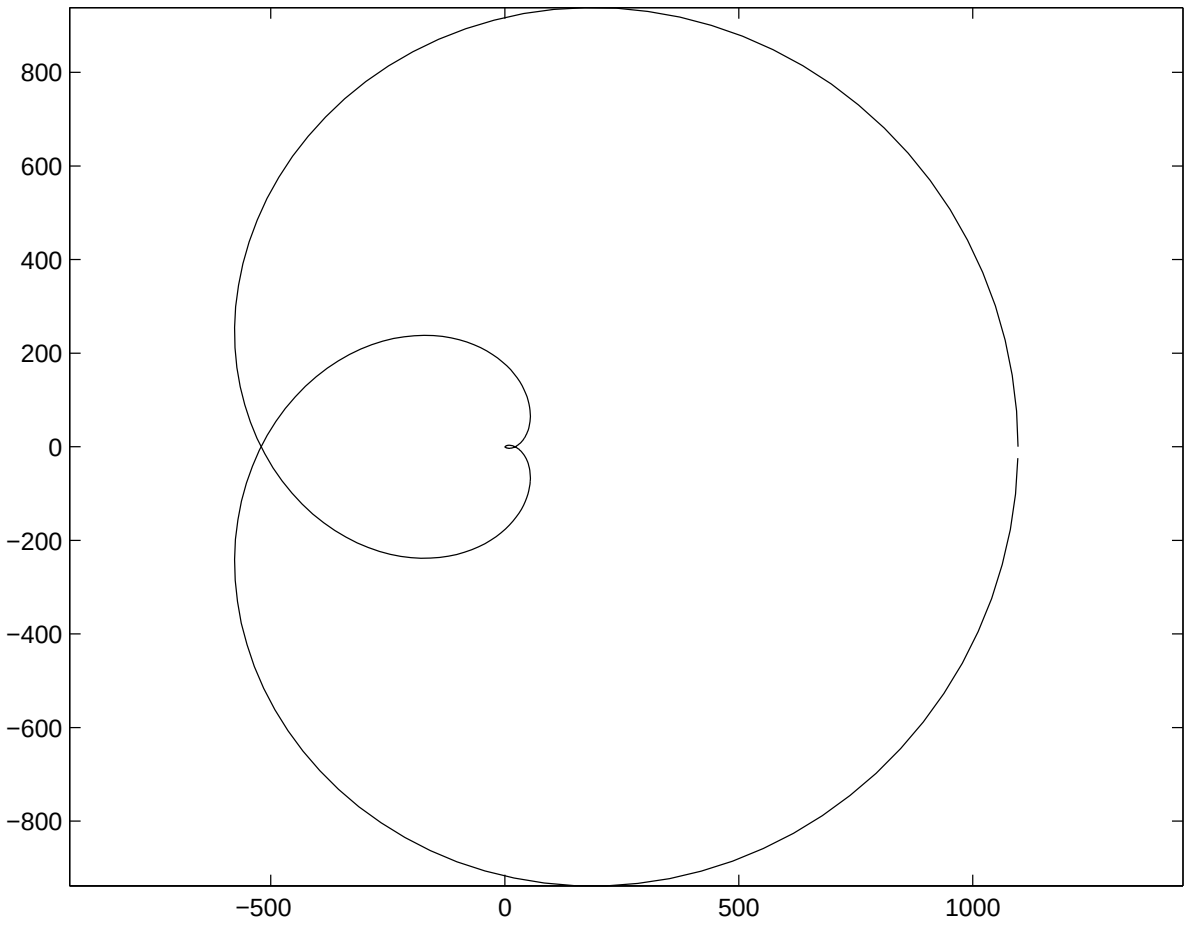

FIGURE 2.2 – L'image du cercle de centre $z = 0$ et de rayon $r = 7$ par l'application exponentielle complexe

Les fonctions cos et sin déduites de la fonction exponentielle *via* les formules

$$\begin{aligned} \cos z &:= \frac{\exp(iz) + \exp(-iz)}{2} \\ \sin z &:= \frac{\exp(iz) - \exp(-iz)}{2i} \end{aligned}$$

sont aussi des fonctions très importantes (du point de vue de la physique par exemple) dans le champ complexe. Sur la figure ci-dessous, on a représenté en trait plein (resp. en pointillés) l'image du cercle de centre $z = 0$ et de rayon 7 par l'application cos (resp. sin). Plus le rayon du cercle augmente, plus les points doubles se multiplient pour l'image par sin !

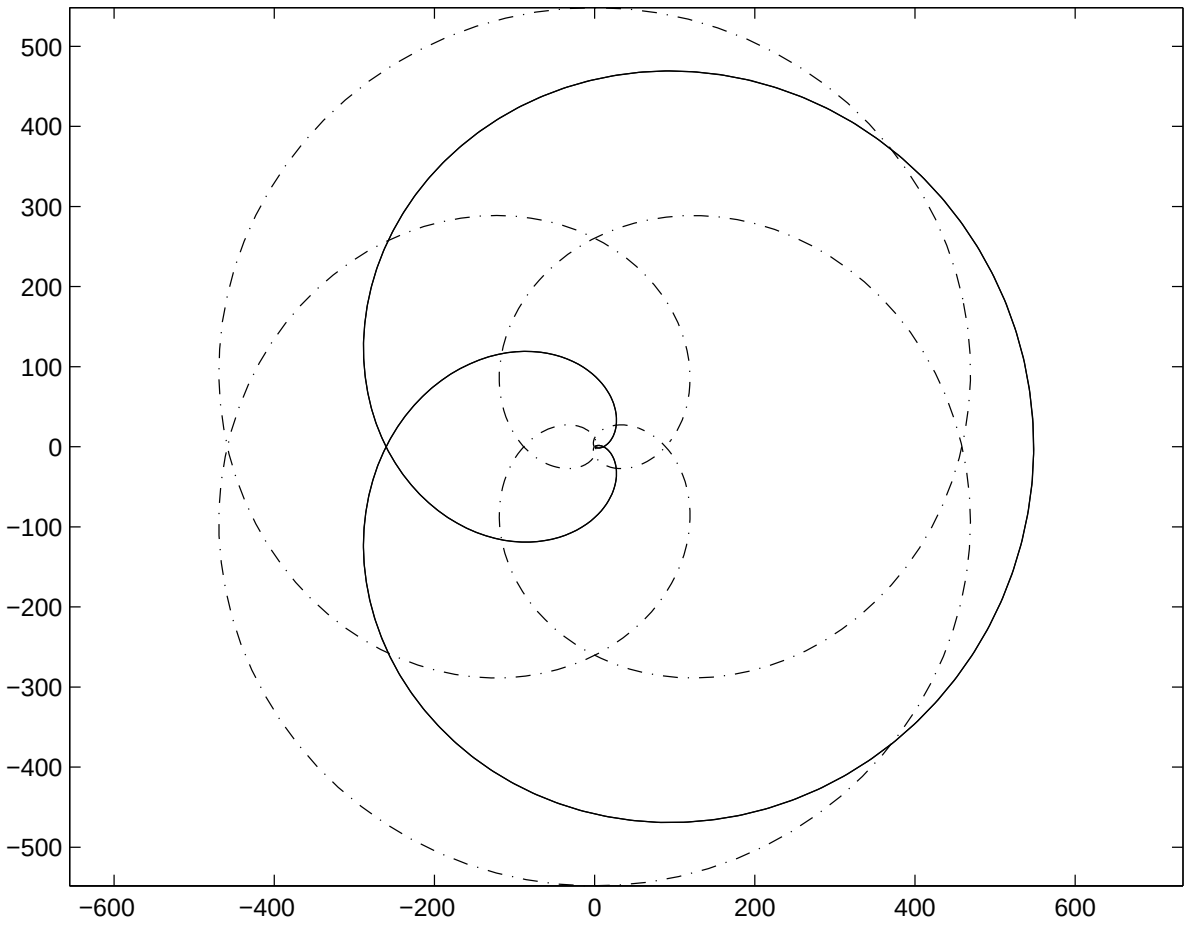

FIGURE 2.3 – L'image du cercle de centre $z = 0$ et de rayon $r = 7$ par les applications cos et sin

Utilisant l'application exponentielle, on remarque que l'écriture d'un nombre complexe sous forme trigonométrique est

$$z = r\exp(i\theta) = re^{i\theta}\,,$$

où θ est une détermination arbitraire de $\arg z$. De plus, les trois nombres e, i, π sont liés par la relation capitale

$$e^{2i\pi} = 1$$

dont on reparlera (cette relation n'est pas un jeu d'écriture, elle implique vraiment quatre êtres mathématiques : une fonction, la fonction exponentielle, et trois nombres, à savoir la base des logarithmes néperiens e, le demi-périmètre du cercle unité π et la racine carrée $i = \sqrt{-1}$.

Comme application de la très importante relation (†), nous énoncerons deux listes de formules, celles attribuées au mathématicien français Abraham de Moivre (1667-1754), à l'origine des tous premiers développements de l'analyse complexe (1707) et celles (en quelque sorte "inverses") attribuées au mathématicien suisse Leonhard Euler (1707-1783).

FORMULES DE MOIVRE. *Pour tout $\theta \in \mathbb{R}$, pour tout $n \in \mathbb{N}$, on a*

$$\begin{aligned}\cos(n\theta) &= \sum_{\{p\in\mathbb{N}\,;\,2p\le n\}} \binom{n}{2p} (-1)^p (\sin\theta)^{2p} (\cos\theta)^{n-2p} \\ \sin(n\theta) &= \sum_{\{p\in\mathbb{N}\,;\,2p+1\le n\}} \binom{n}{2p+1} (-1)^p (\sin\theta)^{2p+1} (\cos\theta)^{n-(2p+1)}\,.\end{aligned}$$

Preuve. On utilise la formule

$$(\cos\theta + i\sin\theta)^n = (e^{i\theta})^n = \cos(n\theta) + i\sin(n\theta)\,.$$

En appliquant la formule du binôme dans le corps commutatif $(\mathbb{C}, +, \times)$, il vient

$$(\cos\theta + i\sin\theta)^n = \sum_{k=0}^{n} \binom{n}{k} i^k (\sin\theta)^k (\cos\theta)^{n-k}\,.$$

On remarque ensuite que $i^{2p} = (-1)^p$ et que $i^{2p+1} = (-1)^p i$ (car $i^2 = -1$). En identifiant les parties réelles et imaginaires, on obtient les formules voulues. □

Remarque. On remarque qu'il existe un polynôme T_n de degré exactement n tel que

$$\cos(n\theta) = T_n(\cos\theta)\,;$$

on a $T_0(X) = 1$, $T_1(X) = X$, $T_2(X) = 2X^2 - 1$, à vous de trouver $T_3(X)$, $T_4(X)$, *etc.* On pourra vérifier ceci soit directement avec les formules de Moivre, soit en utilisant un raisonnement par récurrence. Le polynôme T_n, dit n-ème polynôme de Tchebychev, joue un rôle très important dans les questions relatives à l'approximation des fonctions par des fonctions polynomiales.

FORMULES d'EULER *Pour tout $\theta \in \mathbb{R}$, pour tout $n \in \mathbb{N}$,*

$$\begin{aligned}(\cos\theta)^n &= \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k} e^{i(n-2k)\theta} = \frac{1}{2^n}\sum_{k=0}^{n}\binom{n}{k}\cos((n-2k)\theta) \\ (\sin\theta)^n &= \frac{(-i)^n}{2^n}\sum_{k=0}^{n}\binom{n}{k}(-1)^k e^{i(n-2k)\theta} \\ &= \frac{(-1)^p}{2^{2p}}\sum_{k=0}^{2p}\binom{2p}{k}(-1)^k\cos((2(p-k)\theta)\ si\ n=2p \\ &= \frac{(-1)^p}{2^{2p+1}}\sum_{k=0}^{2p+1}\binom{2p+1}{k}(-1)^k\sin((2(p-k)+1)\theta)\ si\ n=2p+1\,.\end{aligned}$$

Preuve. On part des deux formules

$$\begin{aligned}\cos\theta &= \frac{1}{2}(e^{i\theta} + e^{-i\theta}) \\ \sin\theta &= \frac{1}{2i}(e^{i\theta} - e^{-i\theta}) = \frac{(-i)}{2}(e^{i\theta} - e^{-i\theta})\end{aligned}$$

(connues au lycée comme les formules d'Euler) que l'on élève à la puissance n, puis l'on applique la formule du binôme. On remarque ensuite que les membres de gauche des formules sont réels, ce qui

fait que dans le membre de droite, on peut se contenter de prendre la partie réelle (la partie imaginaire étant automatiquement nulle du fait de l'égalité). □

Remarque. Les formules d'Euler permettent d'exprimer sous forme d'une combinaison linéaire finie

$$\sum_{k\in\mathbb{Z}} a_k e^{ik\theta}$$

toute expression de la forme $P(\cos\theta, \sin\theta)$ où

$$P(X,Y) = \sum_{k_1}\sum_{k_2} a_{k_1,k_2} X^{k_1} Y^{k_2}\,,$$

la somme étant finie et les coefficients a_{k_1,k_2} étant des nombres complexes. Ce procédé est dit *linéarisation des expressions trigonométriques* et jouera un rôle majeur dans la simplification des calculs de primitives ou d'intégrales de telles expressions, on le verra au chapitre suivant.

2.3.5 Résolution dans $\mathbb{C}$ de l'équation algébrique $z^n = A$

Comme on l'a vu, l'équation $z^2 + 1 = 0$ admet deux solutions dans $\mathbb{C}$, les deux nombres $z = i$ et $z = -i$. L'équation $z^2 = 1$ a, elle aussi, deux racines complexes (ici réelles), $z = \pm 1$.

Plus généralement, considérons, si A est un nombre complexe non nul et n un entier strictement positif, l'équation $z^n = A$. écrivons A sous forme trigonométrique

$$A = Re^{i\theta}$$

et cherchons les solutions éventuelles z de l'équation $z^n = A$ sous la forme $z = re^{i\varphi}$.

On doit donc avoir

$$z^n = r^n e^{ni\varphi} = Re^{i\theta}\,;$$

l'égalité de ces deux nombres complexes équivaut à l'égalité de leurs modules et de leurs arguments, c'est-à-dire :

$$\begin{aligned} r^n &= R \\ \exists k\in\mathbb{Z}\,,\ n\varphi &= \theta + 2k\pi\,. \end{aligned}$$

Ceci équivaut à

$$\begin{aligned} r &= R^{1/n} \\ \exists k\in\mathbb{Z}\,,\ \varphi &= \frac{\theta}{n} + \frac{2k\pi}{n}\,. \end{aligned}$$

Ceci équivaut encore, si l'on utilise la division euclidienne de k par n, $k = nq + m$:

$$\begin{aligned} r &= R^{1/n} \\ \exists m\in\{0,...,n-1\}\,,\ \exists q\in\mathbb{Z},\ \varphi &= \frac{\theta}{n} + \frac{2m\pi}{n} + 2\pi q\,, \end{aligned}$$

ou encore

$$\begin{aligned} |z| &= R^{1/n} \\ \exists m\in\{0,...,n-1\}\,,\ \arg z &= \arg\left(\exp\left(i\frac{\theta+2\pi m}{n}\right)\right). \end{aligned}$$

On est donc en mesure de prouver l'important résultat suivant :

Proposition 2.9 *Si $n \in \mathbb{N} \setminus \{0\}$ et $A = Re^{i\theta} \in \mathbb{C} \setminus \{0\}$, il existe exactement n nombres complexes solutions de l'équation algébrique $z^n = A$. Ces n nombres sont les n nombres distincts*

$$z = R^{1/n} \exp\left(i\frac{\theta + 2\pi m}{n}\right),\ m = 0, ..., n-1\,.$$

Preuve. On vient de voir avec ce qui précède que ces nombres sont tous solutions de $z^n = A$. Il reste à montrer que ces nombres sont bien distincts. Si l'on avait

$$R^{1/n} \exp\left(i\frac{\theta + 2\pi m_2}{n}\right) = R^{1/n} \exp\left(i\frac{\theta + 2\pi m_1}{n}\right)$$

avec $0 \leq m_1 < m_2 \leq n-1$, on aurait

$$\exp\left(i\frac{2\pi(m_2 - m_1)}{n}\right) = 1\,,$$

ce qui est impossible car $0 < \dfrac{m_2 - m_1}{n} 2\pi < 2\pi$, ce qui prouve que l'argument du nombre

$$\exp\left(i\frac{2\pi(m_2 - m_1)}{n}\right)$$

est différent de celui de 1 (dont les déterminations sont les $2k\pi$, $k \in \mathbb{Z}$). □

Exemple. Si $A = 1$, l'équation $z^n = 1$ admet n racines distinctes, les n nombres

$$z_m = \exp(\frac{2mi\pi}{n})\,,\ m = 0, ..., n-1\,.$$

Les affixes de ces points sont les n sommets (dont le point $(1, 0)$) du polygone régulier à n côtés inscrit dans le cercle de centre $(0, 0)$ et de rayon 1. Ce polygone est unique dès qu'il a pour sommet le point $(1, 0)$ et on obtient ses autres sommets en découpant à partir du point $(1, 0)$ le cercle de rayon 1 en n arcs de longueur égales $(\dfrac{2\pi}{n})$; voir la figure ci-dessous.

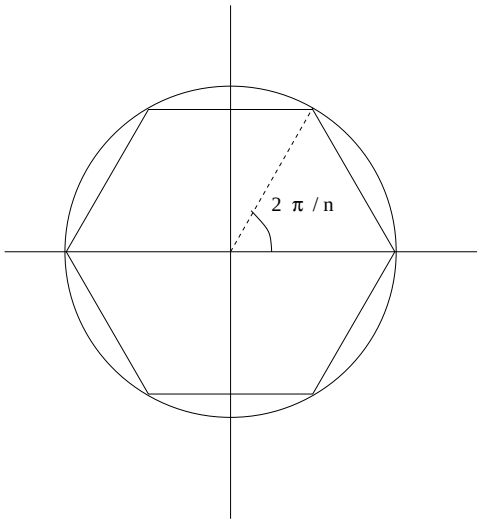

FIGURE 2.4 – Racines n-èmes de l'unité

Les nombres complexes $\exp(\frac{2mi\pi}{n})$ avec $\mathrm{PGCD}\,(m,n) = 1$ sont dits *racines primitives n-èmes de l'unité.*

2.3.6 Résolution des équations du second degré

Plus généralement, si a, b, c sont trois nombres complexes avec $a \neq 0$, on peut écrire, pour tout $z \in \mathbb{C}$,

$$\begin{aligned} az^2+bz+c &= a(z^2+\frac{b}{a}z)+c \\ &= a\Big((z+\frac{b}{2a})^2-\frac{b^2-4ac}{4a^2}\Big)\,. \end{aligned}$$

Dire que $az^2+bz+c=0$ revient à écrire

$$(z+\frac{b}{2a})^2=\frac{b^2-4ac}{4a^2}\,.$$

Soit u une racine complexe de l'équation

$$X^2=b^2-4ac\,; \qquad (\dagger)$$

deux cas se présentent :

- soit $b^2-4ac=0$, auquel cas l'équation (†) a une seule racine $z=0$;
- soit $b^2-4ac\neq 0$, auquel cas l'équation (†) a deux racines distinctes u et $-u$, affixes de points symétriques par rapport à l'origine.

Il en résulte deux configurations concernant la résolution de l'équation $az^2+bz+c=0$ dans $\mathbb{C}$:

- si $b^2-4ac=0$, l'équation $az^2+bz+c=0$ devient

  $$(z+\frac{b}{2a})^2=0$$

  et admet comme seule solution $z=-\dfrac{b}{2a}$;
- si $b^2-4ac\neq 0$, l'équation $az^2+bz+c=0$ admet deux racines distinctes

  $$z=\frac{-b\pm u}{2a}\,,$$

  où $\pm u$ sont les deux nombres complexes solutions de l'équation $X^2=b^2-4ac$.

On a représenté sur la figure ci-dessous les nombres complexes $-b/2a$, le nombre $(b^2-4ac)/(4a^2)=\delta/(4a^2)$ et ses deux racines carrées $\pm\sqrt{\delta}/2a=\pm u/2a$, puis les deux racines z_1 et z_2 de l'équation $az^2+bz+c=0$ (on a posé $\delta/(4a^2)=Re^{i\theta}$).

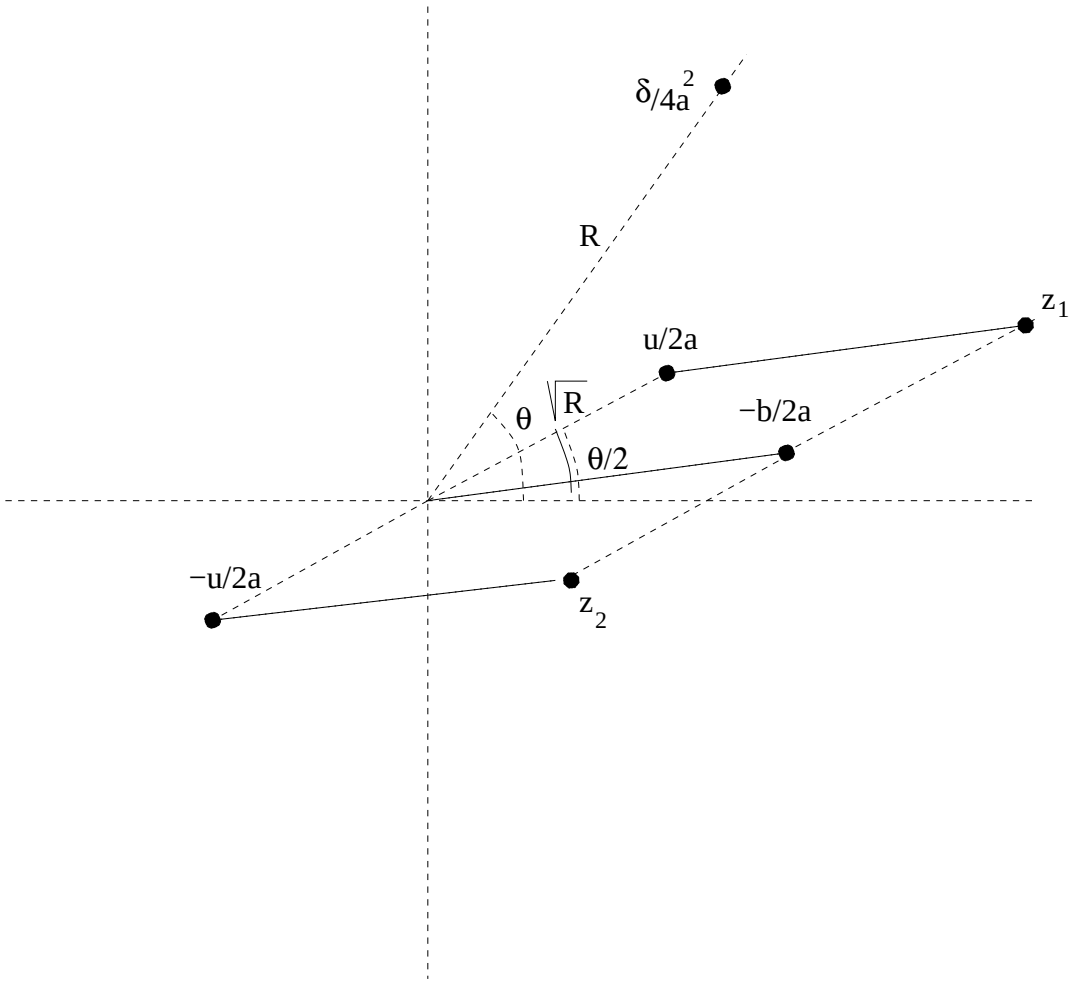

FIGURE 2.5 – Résolution graphique de $az^2 + bz + c = 0$

Si a, b, c sont des nombres réels avec $a \neq 0$, trois cas sont à distinguer :

- $b^2 - 4ac = 0$, auquel cas l'équation $az^2 + bz + c = 0$ a une racine unique dans $\mathbb{C}$ qui d'ailleurs se trouve être réelle $(z = -\dfrac{b}{2a})$;
- $b^2 - 4ac > 0$, auquel cas l'équation $az^2 + bz + c = 0$ a ses deux racines (distinctes) dans $\mathbb{R}$;
- $b^2 - 4ac < 0$, auquel cas l'équation $az^2 + bz + c = 0$ a ses deux racines (distinctes et conjuguées) dans $\mathbb{C} \setminus \mathbb{R}$.

Dans le premier cas $-b/2a$ est réel, les deux nombres $u/2a$ et $-u/2a$ aussi et tous les points de la figure précédente sont sur l'axe réel, en particulier les deux racines z_1 et z_2. Dans le second cas, on a $\theta = \pi$ et les points représentant $u/2a$ et $-u/2a$ sont sur l'axe $y'Oy$, symétriques par rapport à l'origine ; on a $u = i\sqrt{-\delta} \in i\mathbb{R}$ et les deux racines z_1 et z_2 correspondent à des points symétriques par rapport à l'axe $x'Ox$ (voir la figure ci-dessous).

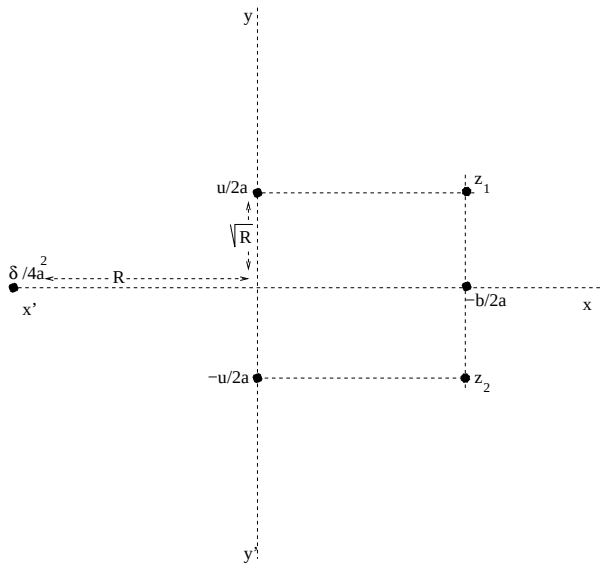

FIGURE 2.6 – Résolution graphique de $az^2 + bz + c = 0$, a, b, c réels et $\delta < 0$

Remarque. Le calcul de la racine carrée $z = x + iy$ d'un nombre $A = u + iv \neq 0$ donné sous forme cartésienne est grandement facilité par les deux relations :

$$\begin{aligned} u = \operatorname{Re} A &= x^2 - y^2 \\ |u + iv| = \sqrt{u^2 + v^2} = |z|^2 &= x^2 + y^2 \,, \end{aligned}$$

relations d'où l'on déduit

$$\begin{aligned} x &= \pm\sqrt{\frac{\sqrt{u^2+v^2}+u}{2}} \\ y &= \pm\sqrt{\frac{\sqrt{u^2+v^2}-u}{2}} \,. \end{aligned}$$

Cependant, on obtient ainsi quatre nombres complexes au lieu des deux prévus (l'équation du second degré $z^2 = A$ ayant deux racines), à savoir les quatre nombres

$$\pm\sqrt{\frac{\sqrt{u^2+v^2}+u}{2}} \pm i\sqrt{\frac{\sqrt{u^2+v^2}-u}{2}} \,;$$

il faut utiliser la relation supplémentaire

$$2xy = v = \operatorname{Im} A \,,$$

relation dont on retient juste

$$\operatorname{signe}(xy) = \operatorname{signe} v$$

pour conclure et isoler les deux couples (x_0, y_0) et $(-x_0, -y_0)$ solutions parmi les quatre trouvés. Par exemple, les deux racines de $z^2 = -3 + 7i$ sont à prendre parmi les quatre nombres

$$\pm\sqrt{\frac{\sqrt{58}-3}{2}} \pm i\sqrt{\frac{\sqrt{58}+3}{2}} \,;$$

comme $xy = 7 > 0$, il ne reste que deux choix possibles et l'on trouve ainsi que les deux racines de l'équation $z^2 = -3+7i$ sont

$$z = \pm\left(\sqrt{\frac{\sqrt{58}-3}{2}} + i\sqrt{\frac{\sqrt{58}+3}{2}}\right).$$

Cette étude de la résolution des équations du second degré augure d'un résultat plus général attribué à l'encyclopédiste Jean Le Rond d'Alembert (1717-1783) : *toute équation algébrique*

$$a_0 z^n + a_1 z^{n-1} + \cdots + a_n \,,$$

les a_j étant des nombres complexes avec $a_0 \neq 0$, admet au moins une racine dans $\mathbb{C}$. Plusieurs démonstrations vous en seront présentées plus tard dans votre parcours mathématique.

FIN DU CHAPITRE 2

Chapitre 3

Fonctions numériques et modélisation

L'objectif dans ce chapitre est l'étude des fonctions numériques d'une variable réelle, c'est-à-dire des fonctions d'un sous-ensemble D de $\mathbb{R}$ (dit *domaine de définition* de la fonction) et à valeurs dans $\mathbb{R}$. L'ensemble $\mathbb{R}$ pouvant être pensé comme l'axe des temps, ces fonctions modélisent souvent des grandeurs physiques (mesurées dans un système d'unités adéquat) évoluant pendant un certain laps de temps. On mettra en équation l'étude de cette évolution de manière à la fois à la prédire et à la contrôler, mais il nous faudra auparavant introduire les concepts de *continuité* et de *dérivabilité*. Cela nous donnera l'opportunité d'associer à un phénomène d'évolution un *modèle mathématique*, puis ensuite d'étudier ce modèle.

3.1 Limite d'une fonction en un point de $\mathbb{R}$ ou de $\mathbb{R} \cup \{-\infty, +\infty\}$

Soit D un sous-ensemble de $\mathbb{R}$ et f une fonction de D dans $\mathbb{R}$.

Définition 3.1. Si a est un point de $\overline{D}$ (c'est-à-dire la limite d'une suite de points de D, on dit que f *admet une limite finie* $l \in \mathbb{R}$ *en* a si et seulement si pour toute suite $(x_n)_{n\geq 0}$ de points de D convergeant vers a (il en existe puisque $a \in \overline{D}$!), on a

$$\lim_{n\to+\infty} f(x_n) = l\,.$$

Dans le cas particulier où a est un point appartenant au domaine de définition de f, alors, dire que f admet une limite l en a implique automatiquement que cette limite l soit égale à $f(a)$. On dit dans ce cas (c'est une définition sur laquelle nous reviendrons en détails dans la section 3.2 plus loin) que f est *continue* au point a. Lorsque a est un point du domaine de définition d'une fonction f, dire que cette fonction est continue en ce point a, c'est exactement dire que la limite de f en a existe (cette limite valant d'ailleurs automatiquemenr $f(a)$), ce qui signifie que pour toute suite $(x_n)_{n\in\mathbb{N}}$ de points de D convergeant vers a, on a

$$\lim_{n\to+\infty} f(x_n) = f(a)\,.$$

Dire que f n'est pas continue en a revient donc à exhiber une suite $(x_n)_{n\in\mathbb{N}}$ de points de D qui tende vers a sans que la suite $(f(x_n))_{n\in\mathbb{N}}$ tende vers $f(a)$.

Lorsque a est un point de $\overline{D}\setminus D$, on peut aussi parler pour f de *convergence de* f *vers* $\pm\infty$ *au point* a :

– On dit que la fonction f *converge vers* $+\infty$ *au point* $a \in \overline{D}\setminus D$ (ou encore *admet* $+\infty$ *pour limite au point* a, ou bien encore *tend vers* $+\infty$ *en* a) si et seulement si pour toute suite $(x_n)_{n\geq 0}$ de

points de D convergeant vers a, on a

$$\lim_{n\to+\infty} f(x_n) = +\infty\,;$$

– on dit enfin que la fonction f *converge vers* $-\infty$ *au point* $a \in \overline{D} \setminus D$ (ou encore *admet* $-\infty$ *pour limite au point* a, ou bien encore *tend vers* $-\infty$ *en* a) si et seulement si pour toute suite $(x_n)_{n\geq 0}$ de points de D convergeant vers a,

$$\lim_{n\to+\infty} f(x_n) = -\infty\,.$$

Remarque. Si a est un point de D et si f (définie sur D) admet une limite en a, alors nécessairement cette limite est égale à $f(a)$.

Si D est tel qu'il existe une suite de points de D tendant vers $+\infty$ (c'est-à-dire si D n'est pas majoré), on peut aussi parler de *limite de la fonction* f *au point* $+\infty$:

– on dit que f admet une limite finie $l \in \mathbb{R}$ lorsque x tend vers $+\infty$ dans D si et seulement si pour toute suite $(x_n)_{n\geq 0}$ de points de D convergeant vers $+\infty$ dans D (il en existe par hypothèses sur D !), on a

$$\lim_{n\to+\infty} f(x_n) = l\,;$$

– on dit que f converge vers $+\infty$ lorsque x tend vers $+\infty$ dans D si et seulement si pour toute suite $(x_n)_{n\geq 0}$ de points de D convergeant vers $+\infty$ dans D, on a

$$\lim_{n\to+\infty} f(x_n) = +\infty\,;$$

– on dit que f tend vers $-\infty$ lorsque x tend vers $+\infty$ dans D si et seulement si pour toute suite $(x_n)_{n\geq 0}$ de points de D convergent vers $+\infty$ dans D, on a

$$\lim_{n\to+\infty} f(x_n) = -\infty\,.$$

S'il existe une suite de points de D tendant vers $-\infty$ (ce qui signifie que D n'est pas minoré), on peut parler pour une fonction f de D dans $\mathbb{R}$ de convergence vers une limite finie l ou vers $\pm\infty$ lorsque x tend vers $-\infty$ dans D.

On a la proposition suivante, permettant de formuler mathématiquement le fait qu'une fonction tende vers une limite finie $l \in \mathbb{R}$ au point $a \in \overline{D}$ ou vers $\pm\infty$ en un point $a \in \overline{D} \setminus D$:

Proposition 3.1. *Soit D un sous-ensemble de $\mathbb{R}$, f une fonction de D dans $\mathbb{R}$ et l un nombre réel. Alors f admet l pour limite au point $a \in \overline{D}$ si et seulement si :*

$$\forall \epsilon > 0\,,\ \exists \eta > 0\,,\ \forall x \in D\,,\ |x-a| < \eta \Longrightarrow |f(x) - l| < \epsilon\,. \qquad (\dagger)$$

De même, si $a \in \overline{D} \setminus D$, f tend vers $+\infty$ au point a si et seulement si

$$\forall A > 0\,,\ \exists \eta > 0\,,\ \forall x \in D\,,\ |x-a| < \eta \Longrightarrow f(x) > A\,.$$

Enfin, toujours si $a \in \overline{D} \setminus D$, f tend vers $-\infty$ au point a si et seulement si

$$\forall A > 0\,,\ \exists \eta > 0\,,\ \forall x \in D\,,\ |x-a| < \eta \Longrightarrow f(x) < -A\,.$$

Preuve. On se contentera de prouver la première assertion (la preuve des autres est identique).

Supposons tout d'abord (†) vraie et fixons $\epsilon > 0$ arbitraire ; il lui correspond donc d'après (†) un certain nombre $\eta > 0$. Si $(x_n)_n$ est une suite de points de D convergeant vers a, alors pour n assez grand, on a $|x_n - a| < \eta$; mais alors, puisque (†) est vraie, on a $|f(x_n) - l| \leq \epsilon$ pour n assez grand ; comme ϵ est arbitraire, ceci prouve que la suite $(f(x_n))_n$ converge vers l (si l'on se souvient de ce que signifie la convergence d'une suite $(u_n)_n$ de nombres réels vers une limite réelle, voir la proposition 2.5).

On prouve que l'assertion $\lim\limits_{x\to a} f(x) = l$ implique l'assertion (†) en utilisant un raisonnement par contraposition. La négation de (†) s'écrit :

$$\exists \epsilon > 0\,,\ \forall \eta > 0\,,\ \exists x \in D\,, (|x - a| < \eta) \wedge (|f(x) - l| \geq \epsilon)\,. \qquad \text{(non †)}$$

En particulier, pour chaque $n \in \mathbb{N} \setminus \{0\}$, il existe $x_n \in D$ avec $|x_n - a| < 1/n$ et $|f(x_n) - l| \geq \epsilon$; la suite $(x_n)_{n\geq 1}$ converge vers a tandis que la suite $(f(x_n))_{n\geq 1}$ ne converge pas vers l, ce qui prouve que l'assertion $\lim\limits_{x\to a} f(x) = l$ est fausse. On a donc bien prouvé ainsi par contraposition que

$$(\lim_{x\to a} f(x) = l) \Longrightarrow (\dagger)\,,$$

ce qui achève la preuve de l'équivalence de ces deux assertions. □

De même, si D est non majoré, on a

$$\begin{aligned}
\lim_{x\to+\infty} f(x) = l &\iff \Big(\forall \epsilon > 0\,,\ \exists B > 0\,,\ \forall x \in D\,,\ x > B \Longrightarrow |f(x) - l| < \epsilon\Big)\\
\lim_{x\to+\infty} f(x) = +\infty &\iff \Big(\forall A > 0\,,\ \exists B > 0\,,\ \forall x \in D\,,\ x > B \Longrightarrow f(x) > A\Big)\\
\lim_{x\to+\infty} f(x) = -\infty &\iff \Big(\forall A > 0\,,\ \exists B > 0\,,\ \forall x \in D\,,\ x > B \Longrightarrow f(x) < -A\Big)\,.
\end{aligned}$$

On a des équivalences analogues si D est non minoré pour les trois types de convergence vers $-\infty$.

On remarque que la limite d'une fonction f en un point a de $\overline{D}$ (a pouvant être éventuellement $-\infty$ ou $+\infty$), limite qui peut être un nombre réel l ou bien valoir $\pm\infty$ si $a \in \overline{D}\setminus D$, peut fort bien ne pas exister (par exemple la fonction $f\ : x \in \mathbb{R} \longmapsto \sin x$ ne tend ni vers une limite l, ni vers $+\infty$, ni vers $-\infty$ lorsque x tend vers $+\infty$). En revanche, si cette limite existe, elle est unique ; une fonction numérique ne saurait avoir deux limites distinctes (appartenant à la droite numérique achevée) en un point a adhérent à son domaine de définition D (ou en $+\infty$ si D n'est pas majoré, ou bien en $-\infty$ si D n'est pas minoré).

Toutes les règles de calcul concernant la compatibilité entre la prise de limite et les opérations sur $\mathbb{R}$ (addition, multiplication, prise d'inverse si cela est possible, prise de valeur absolue) énoncées pour les suites dans la proposition 2.6 ou dans la section 2.2.8 rejaillissent sur les limites de fonctions (puisque tester si une fonction a une limite en un point de la droite achevée revient à le tester sur les suites). On énoncera pas toutes ces règles mais il est utile en exercice que vous vous exerciez à les "retranscrire" toutes du contexte des suites vers celui des fonctions.

Nous donnerons quelques règles de compatibilité entre prise de limite et composition des applications :

Proposition 3.2. *Soit D et E deux sous-ensembles de $\mathbb{R}$, f une fonction de D dans $\mathbb{R}$, g une fonction de E dans $\mathbb{R}$ avec $f(D) \subset E$. On suppose que*

- *$a \in \overline{D}$ et $\lim\limits_{x\to a, x\in D} f(x) = l \in \mathbb{R}$ (notons que ceci implique que l est automatiquement dans $\overline{E}$) ;*
- *$\lim\limits_{y\to l, y\in E} g(y) = L \in \mathbb{R}$.*

Alors

$$\lim_{x\to a\,,x\in D}(g\circ f)(x)=L\,.$$

Preuve. Il s'agit de montrer :

$$\forall\epsilon>0\,,\ \exists\eta>0\,,\ \forall x\in D\,,\ |x-a|<\eta\Longrightarrow|g(f(x))-L|<\epsilon\,. \qquad (*)$$

Or, on sait que

$$\forall\epsilon>0\,,\ \exists\eta_1>0\,,\ \forall y\in E\,,\ |y-l|<\eta_1\Longrightarrow|g(y)-L|<\epsilon \qquad (1)$$

puisque $\lim_{y\to l,y\in E} g(y)=L$. D'autre part, puisque $\lim_{x\to a,x\in D} f(x)=l$, il existe $\eta>0$ tel que

$$\forall x\in D\,,\ |x-a|<\eta\Longrightarrow|f(x)-l|<\eta_1\,. \qquad (2)$$

En combinant (1) et (2), on voit que si $x\in D$ et $|x-a|<\eta$, on a $|f(x)-l|<\eta_1$ et donc par ricochet $y=f(x)\in E$ et par voie de conséquence $|g(y)-L|=|g(f(x))-L|<\epsilon$, ce qui était ce que l'on voulait montrer. □

On a aussi des propositions tout à fait identiques du type par exemple de la suivante :

Proposition 3.4. *Soit D et E deux sous-ensembles de $\mathbb{R}$, f une fonction de D dans $\mathbb{R}$, g une fonction de E dans $\mathbb{R}$ avec $f(D)\subset E$. On suppose que*

- *$a\in\overline{D}$ et $\lim_{x\to a,x\in D} f(x)=+\infty\in\mathbb{R}$ (notons que ceci implique que E est automatiquement non majoré) ;*
- *$\lim_{y\to+\infty,y\in E} g(y)=L\in\mathbb{R}\cup\{-\infty,+\infty\}$.*

Alors

$$\lim_{x\to a\,,x\in D}(g\circ f)(x)=L\,.$$

Ou bien encore :

Proposition 3.5. *Soit D et E deux sous-ensembles de $\mathbb{R}$, f une fonction de D dans $\mathbb{R}$, g une fonction de E dans $\mathbb{R}$ avec $f(D)\subset E$. On suppose que*

- *D non majoré et $\lim_{x\to+\infty,x\in D} f(x)=l\in\mathbb{R}$ (notons que ceci implique que $l\in\overline{E}$) ;*
- *$\lim_{y\to l,y\in E} g(y)=L\in\mathbb{R}\cup\{-\infty,+\infty\}$.*

Alors

$$\lim_{x\to+\infty\,,x\in D}(g\circ f)(x)=L\,.$$

Les preuves des propositions 3.4 et 3.5 sont identiques à celle de la proposition 3.3. □

Si a est un point de $\overline{D}$, on dira aussi que f a une limite à droite $l\in\mathbb{R}\cup\{-\infty,+\infty\}$ si et seulement a est adhérent à $D\cap]a,+\infty[$ et si la restriction de f à $D\cap]a,+\infty[$ (c'est-à-dire l'application qui à $x\in D\cap]a,+\infty[$ associe $f(x)$) a pour limite l lorsque x dans vers a dans $D\cap]a,+\infty[$. Idem pour la notion de limite à gauche en a. Lorsque a est adhérent à la fois à $D\cap]a,+\infty[$ et $D\cap]-\infty,a[$, on peut donc dire que f a une limite en a si f a des limites à droite et à gauche en a et si ces deux limites sont égales (comme éléments de la droite numérique achevée). Si $a\in A$, ces deux limites sont nécesairement alors de plus égales à $f(a)$.

Exemples.

– Soit f la fonction de $\mathbb{R}$ dans $[0,1]$ qui à un nombre réel x associe le nombre $x - E(x)$, où $E(x)$ désigne la partie entière de x (c'est-à-dire le plus grand nombre $m \in \mathbb{Z}$ inférieur ou égal à x). En tout point x de $\mathbb{R} \setminus \mathbb{Z}$, la fonction f admet une limite égale à $f(x)$ (on vérifiera ce fait en utilisant les développements décimaux). En revanche, en tout point de $\mathbb{Z}$, la fonction f a pour limite à gauche $l = 0$ et pour limite à droite $l = 1$.
– Soit f la fonction définie sur $D =]-\pi, \pi[\setminus\{0\}$ par

$$f(x) := \frac{\cos x}{\sin x}\,.$$

En tout point x de D, f a pour limite $f(x)$; en revanche, au point $x = 0$, f a pour limite à droite $+\infty$ (car $x \mapsto \cos x$ a pour limite 1 en 0 tandis que $x \mapsto \sin x$ a pour limite en zéro à droite 0 par valeurs supérieures) et à gauche $-\infty$ (car $x \mapsto \cos x$ a pour limite 1 en 0 tandis que $x \mapsto \sin x$ a pour limite en zéro à gauche 0 par valeurs inférieures). Au point π, on vérifiera que f a pour limite $-\infty$ tandis qu'en $-\pi$, f a pour limite $+\infty$.

Terminons par une proposition intéressante :

Proposition 3.6. *Soit f est une fonction monotone (c'est-à-dire croissante ou décroissante) sur un intervalle ouvert $]a, b[$ (borné ou non borné). Alors f a une limite à gauche et à droite en tout point de $]a, b[$.*

Preuve. Ceci résulte du fait que $\mathbb{R}$ vérifie la propriété de la borne supérieure. Supposons par exemple f croissante et notons, pour $x_0 \in]a, b[$, $M(x_0)$ la borne supérieure de $f(]a, x_0[)$ (sous-ensemble de $\mathbb{R}$ majoré par $f(x_0)$). Par définition de $M(x_0)$, il existe, pour tout $\epsilon > 0$, un nombre $x < x_0$ tel que

$$M(x_0) - \epsilon < f(x) \leq M(x_0)\,;$$

mais alors, puisque f est croissante, on a

$$\forall x' \in [x, x_0[\,,\ M(x_0) - \epsilon < f(x') \leq M(x_0)\,;$$

on a donc

$$\lim_{x \to x_0, x < x_0} f(x) = M(x_0) = \sup\{f(x)\,;\, x < x_0\}$$

tandis que (par le même raisonnement) la limite à droite de f au point x_0 est

$$\lim_{x \to x_0, x > x_0} f(x) = m(x_0) = \inf\{f(x)\,;\, x > x_0\}\,.$$

On écrira ce qui se passe si f est décroissante. □

3.2 Continuité d'une fonction en un point et sur un ensemble

Définition 3.2. Soit f une fonction numérique de $D \subset \mathbb{R}$ à valeurs dans $\mathbb{R}$ et x_0 un point de D. On dit que f est *continue en x_0* si et seulement si

$$\lim_{x \to x_0} f(x) = f(x_0)\,.$$

Si A est un sous-ensemble de D, la fonction f est dite *continue sur A* si et seulement si f est continue en tout point de A. Si f est continue sur D, on dit que f est continue partout (sous-entendu bien sûr seulement là où elle est définie !).

Remarque. On pourrait simplement se contenter de dire que f est continue en x_0 si et seulement si f admet une limite finie en x_0 ; en effet, on a vu que, si cette limite existe, c'est nécessairement $f(x_0)$.

Si x_0 est un point de D adhérent à $D\cap]x_0,+\infty[$, on dit que f *est continue à droite en* x_0 si

$$\lim_{x\to x_0, x>x_0} f(x) = f(x_0)\,.$$

Idem pour la notion de *continuité à gauche en un point* x_0 adhérent à $D\cap]-\infty,x_0[$. Si x_0 est un point de D adhérent à la fois à $D\cap]x_0,+\infty[$ et à $D\cap]-\infty,x_0[$, dire que f est continue en x_0 équivaut donc à dire que f est continue à gauche et à droite en x_0.

Proposition 3.7. *Une fonction continue sur un intervalle fermé* $[a,b]$ *est bornée sur cet intervalle et atteint ses bornes.*

Nota Lire soigneusement la preuve de cette proposition plus délicate que les autres (c'est normal, le matériel dont nous disposons s'accumule de jour en jour plus nous avançons dans le cours) est un exercice qui ne saurait être que profitable. Il ne s'agit pas de savoir la refaire "par cœur" (cela n'a pas d'intérêt et en plus, il y a plein d'autres preuves) mais d'en comprendre les idées clef. J'y reviendrai dans le prochain cours, mais relisez la avant !

Preuve. Soit f une telle fonction. Montrons par l'absurde que f est majorée ; si ce n'était pas le cas, tous les ensembles

$$A_n := \{x\in[a,b]\,;\, f(x)\geq n\}\,,\ n=0,1,2,...,$$

seraient non vides (et bien sûr tous majorés par b). Posons

$$x_n := \sup\{x\,;\, x\in A_n\}$$

(ce nombre existe et est un élément de $[a,b]$ car $\mathbb{R}$ vérifie la propriété de la borne supérieure ou inférieure). La suite $(x_n)_n$ est par construction même (on le vérifiera) une suite décroissante de nombres réels tous entre a et b (en effet $A_{n+1}\subset A_n$, l'ensemble des majorants de A_n est donc inclus dans celui des majorants de A_{n+1}). Cette suite converge donc vers un point $x\in[a,b]$. Comme f est continue sur $[a,b]$, donc en particulier en x, la suite $(f(x_n))_n$ devrait converger vers $f(x)$, ce qui est impossible car $f(x_n)\geq n$ pour tout n, ce qui implique $\lim_{n\to\infty} f(x_n)=+\infty$, d'où la contradiction. La fonction f est donc majorée sur $[a,b]$ et l'on peut poser $M=\sup(f([a,b])$.

D'après le fait que M est borne supérieure de $f([a,b])$, il existe, pour chaque $n\in\mathbb{N}\setminus\{0\}$ un élément x de $[a,b]$ tel que $M-1/n<f(x)\leq M$; on pose

$$u_n := \inf\{x\in[a,b]\,;\, M-1/n<f(x)\leq M\}\,.$$

La suite $(u_n)_{n\geq 1}$ est une suite croissante majorée, donc convergente vers un nombre $\xi\in[a,b]$; comme f est continue en ξ, on a

$$f(\xi)=\lim_{n\to+\infty} f(u_n)\geq \lim_{n\to+\infty}(M-\frac{1}{n})\geq M\,;$$

mais comme on a aussi $f(\xi)\leq\sup(f([a,b])=M$, on a $f(\xi)=M$.

On raisonne de manière identique pour montrer que $f([a,b])$ est minoré et qu'il existe un point η de $[a,b]$ tel que $f(\eta)=\inf(f([a,b])$. □

Une vision intuitive de la continuité d'une fonction f sur un intervalle I (ouvert, semi-ouvert ou fermé, borné ou non) est la suivante : *f est continue sur I si et seulement si le graphe de f, c'est-à-dire le sous-ensemble de $\mathbb{R}^2$ défini comme*

$$G(f) := \{(x,f(x))\,;\, x\in I\}$$

peut se tracer au crayon sans que l'on ait jamais à lever celui-ci. G. Peano a construit par exemple une courbe continue du plan (c'est-à-dire une paire de fonctions continues de $[0,1]$ dans $\mathbb{R}$) telle que l'image

par l'application (γ_1, γ_2) de $[0,1]$ soit $[0,1]^2$. Ainsi le carré $[0,1]^2$ peut-il être balayé d'un trait de crayon "continu" (on ne lève jamais le crayon pour faire ce traçé) ! Ceci repose sur le théorème suivant :

Théorème des valeurs intermédiaires. *Soit f une fonction continue sur un intervalle fermé $[a,b]$; l'image de f est l'intervalle fermé $[\inf(f([a,b])), \sup(f([a,b]))]$, ce qui signifie que f prend toute valeur intermédiaire entre les deux nombres $\inf(f([a,b]))$ et $\sup(f([a,b]))$, ces nombres étant eux-mêmes atteints d'après la proposition 3.7.*

Remarque. Ceci est faux si l'on remplace $[a,b]$ par une union $I_1 \cup I_2$ de deux intervalles fermés disjoints (même s'il est toujours vrai que toute fonction continue sur une telle union y est bornée et atteint ses bornes. La fonction qui à $x \in [0,1] \cup [2,3]$ associe x évite les valeurs $y \in]1,2[$; pourtant $\inf(f([0,1] \cup [2,3])) = 0$ et $\sup(f([0,1] \cup [2,3])) = 3$ et donc $]1,2[\subset [0,3]$!

Nota Je n'ai pas traité cette preuve en cours, mais je vous invite à la lire soigneusement et à tenter de comprendre comment le raisonnement (subtil, car par l'absurde) s'enchevêtre. Ceci montre qu'autant le résultat est intuitivement évident, autant en fait il s'agit d'un résultat profond et non tout à fait évident. La conséquence que nous en donnerons avec la preuve du théorème de d'Alembert pour les équations algébriques à coefficients réels et de degré impair montre que ce théorème des valeurs intermédiaires a des applications non triviales du tout !

Preuve. C'est une nouvelle fois une conséquence du fait que $\mathbb{R}$ vérifie la propriété de la borne supérieure. On fait une preuve par l'absurde en supposant qu'il existe un nombre y tel que

$$\Big(\inf(f([a,b])) < y < \sup(f([a,b]))\Big) \wedge \Big(\forall x \in [a,b]\,,\ f(x) \neq y\Big)$$

(c'est la négation de l'assertion du théorème puisque l'on sait déjà que les deux valeurs $m = \inf(f([a,b])$ et $M = \sup(f([a,b]))$ sont, elles, atteintes sur $[a,b]$ d'après la proposition 3.7). Le nombre $f(a)$ est donc soit strictement inférieur à y, soit strictement supérieur à y.

On suppose tout d'abord que $f(a) < y$ et l'on considère le sous-ensemble de $[a,b]$ défini par

$$E := \{x \in [a,b]\,;\ f([a,x]) \subset]-\infty, y[\}\,.$$

Ce sous-ensemble de $\mathbb{R}$ est non vide majoré et admet donc une borne supérieure $\beta \in [a,b]$.

Montrons que nécessairement $\beta = b$. Si ce n'était pas le cas, on aurait $\beta < b$ et $f(\beta) \neq y$. Deux cas sont alors à envisager :

- soit $f(\beta) - y > 0$ (ce qui impose $\beta > a$), mais alors, du fait de la continuité de f en β, il existerait $\eta > 0$ tel que
$$|x - \beta| < \eta \Longrightarrow f(x) - y > 0\,;$$
ceci est en contradiction avec le fait que dans l'intervalle $]\beta - \eta, \beta] \cap [a,b]$, il doive exister un nombre x tel que $f([a,x]) \subset]-\infty, y[$, donc en particulier tel que $f(x) - y < 0$ (définition de la borne supérieure) ;
- soit $f(\beta) - y < 0$, mais alors, du fait de la continuité de f en β, il existerait $\eta > 0$ tel que
$$|x - \beta| < \eta \Longrightarrow f(x) - y < 0\,;$$
mais alors il existerait $x \in]\beta, \beta + \eta[\cap [a,b]$ tel que $f([a,x]) \subset]-\infty, y[$, ce qui contredit la définition de y comme borne supérieure de E.

On a donc nécessairement $\beta = b$ et $f([a,b]) \subset]-\infty, \alpha[$, ce qui est impossible puisque la valeur $M > y$ doit être prise par f sur $[a,b]$.

Si $f(a) > y$, on considère le sous-ensemble de $[a,b]$ défini par

$$F := \{x \in [a,b]\,;\ f([a,x]) \subset]y, +\infty[\}\,.$$

Ce sous-ensemble de $\mathbb{R}$ est non vide minoré et admet donc une borne inférieure α. Comme précédemment on montre que $\alpha = a$, donc que $f([a,b]) \subset]y, +\infty[$, ce qui est contradictoire avec le fait que la valeur $m < y$ est atteinte par f sur $[a,b]$. □

Application. Toute fonction polynomiale de la forme $x \to a_0x^p + a_1x^{p-1} + \cdots + a_p$ avec $a_0 \neq 0$, les a_j réels et p impair, s'annule en au moins un point de $\mathbb{R}$ (puisque la limite en $-\infty$ est l'infini avec le signe de $-a_0$ et que la limite en $+\infty$ est l'infini avec cette fois le signe de a_0). On vient donc de montrer ici le théorème de d'Alembert dans le cas particulier des équations algébriques à coefficients réels et de degré impair. Il n'existe d'ailleurs pas de preuve "purement algébrique" du théorème fondamental de l'algèbre ; toute preuve implique un voyage du côté de l'analyse comme celui ci !

3.3 Opérations sur les fonctions continues.

Les règles concernant les limites se répercutent en les règles suivantes pour la continuité :

- si f et g sont deux fonctions définies sur D et toutes les deux continues en $x_0 \in D$, $f+g$, fg, f/g lorsque $g(x_0) \neq 0$, sont aussi continues en x_0 ;
- si f est une fonction définie sur D et continue en $x_0 \in D$, $x \to |f(x)|$ est aussi définie sur D et continue en x_0 ;
- si f est une fonction définie sur D, g une fonction définie sur $f(D)$, x_0 un point de D tel que f soit continue en x_0 et g continue en $f(x_0)$, alors $g \circ f$ est définie sur D et continue au point x_0 (on applique la proposition 3.2).

3.4 Fonctions strictement monotones sur un intervalle

Définition 3.3. Soit I un intervalle de $\mathbb{R}$ (de n'importe quel type, borné ou non borné) ; on dit qu'une fonction f de I dans $\mathbb{R}$ est *strictement croissante* sur I si et seulement si

$$\forall x_1, x_2 \in I\,,\ (x_1 < x_2) \Longrightarrow (f(x_1) < f(x_2))\,;$$

on dit qu'une fonction f de I dans $\mathbb{R}$ est *strictement décroissante* sur I si et seulement si

$$\forall x_1, x_2 \in I\,,\ (x_1 < x_2) \Longrightarrow (f(x_1) > f(x_2))\,;$$

une fonction strictement croissante ou strictement décroissante sur I est dite *strictement monotone* sur l'intervalle I.

Le résultat suivant est très important car il nous permet d'introduire la notion de fonction inverse :

Proposition 3.8. *Soit I un intervalle de $\mathbb{R}$ (de n'importe quel type, ouvert, semi-ouvert ou fermé, borné ou non borné) et f une fonction strictement monotone et continue de I dans $\mathbb{R}$; alors f est une bijection entre I et son image $f(I)$ (qui d'ailleurs est aussi un intervalle J du même type que I) et l'application inverse f^{-1} est aussi une application strictement monotone continue de J dans I (strictement croissante si f est strictement croissante, strictement décroissante si f est strictement décroissante).*

Preuve. On suppose f strictement croissante pour fixer les idées. Clairement, f est injective sur I puisque $x_1 \neq x_2$ implique soit $x_1 < x_2$, auquel cas $f(x_1) < f(x_2)$, soit $x_2 < x_1$, auquel cas $f(x_2) < f(x_1)$. L'application f est donc une application bijective entre I et son image $f(I)$ et l'application inverse f^{-1} existe bien (de $f(I)$ dans I). Si $y_1 = f(x_1) < y_2 = f(x_2)$, on a nécessairement $x_1 < x_2$ (puisque sinon $y_2 \geq y_1$). L'application f^{-1} est donc bien strictement croissante aussi. Remarquons que nous n'avons pas encore utilisé la continuité de f à ce point de notre raisonnement. On va l'exploiter maintenant.

Si $y_1 = f(a)$ et $y_2 = f(b)$ sont deux points de $f(I)$ avec $y_1 < y_2$, alors le théorème des valeurs intermédiaires implique que f prenne sur $[a, b]$ toute valeur entre y_1 (incluse) et y_2 (incluse), donc que $f(I)$ contienne l'intervalle $[y_1, y_2]$. Ainsi $f(I)$ ne peut présenter de "trou" et est donc un intervalle J. Cet intervalle J est du même type que I et admet comme borne inférieure $m = \inf\{f(x); x \in I\} \in \mathbb{R} \cup \{-\infty\}$ et comme borne supérieure $M = \sup\{f(x); x \in I\} \in \mathbb{R} \cup \{+\infty\}$ (ces bornes étant incluses si les bornes correspondantes de I le sont, par exemple $f([a, b]) = [m, M]$, $f(]a, b]) =]m, M]$, $f([a, b[) = [m, M[$, $f(]a, b[) =]m, M[$, $f(]-\infty, b]) =]m, M]$, *etc.* si $a, b \in \mathbb{R}$).

La fonction f^{-1} est strictement monotone sur l'intervalle $J = f(I)$ et admet donc, d'après la proposition 3.6, une limite à gauche $f^{-1}(y_-)$ et à droite $f^{-1}(y_+)$ en tout point y de J (avec $f^{-1}(y_-) \leq f^{-1}(y) \leq f^{-1}(y_+)$). Mais, comme $f^{-1}(J) = I$ et ne peut donc avoir de "trou", les limites à gauche et à droite de f^{-1} en y sont égales et valent donc d'après le lemme des gendarmes $f^{-1}(y)$. La fonction f^{-1} est donc continue sur $J = f(I)$. □

Remarque. Le résultat est faux si l'on omet l'hypothèse "f continue" et que l'on garde toutes les autres ; par exemple la fonction $[0, 2] \to [0, 3]$ définie par $f(x) = x$ si $x \in [0, 1[$ et $f(x) = x + 1$ pour $x \in [1, 2]$ a pour image $[0, 1[\cup[2, 3]$ qui n'est pas un intervalle !

Exemples.

• La fonction $x \longmapsto \sin x$ est strictement croissante sur $[-\pi/2, \pi/2]$, à valeurs dans $[-1, 1]$; elle admet donc une fonction réciproque Arcsin $: [-1, 1] \longrightarrow [-\pi/2, \pi/2]$, continue sur $[-1, 1]$ et strictement croissante sur cet intervalle (bijective de $[-1, 1]$ dans $[-\pi/2, \pi/2]$) ; on a

$$\Big((x \in [-\pi/2, \pi/2]) \wedge (y = \sin x)\Big) \Longleftrightarrow \Big((y \in [-1, 1]) \wedge (x = \operatorname{Arcsin} y)\Big) \ ;$$

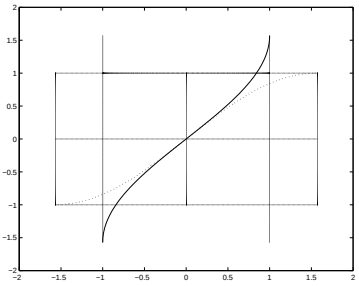

FIGURE 3.1 – Graphes de sin (en pointillés) et de Arcsin (en plein)

• La fonction $x \longmapsto \cos x$ est strictement décroissante sur $[0, \pi]$, à valeurs dans $[-1, 1]$; elle admet donc une fonction réciproque Arcos $: [-1, 1] \longrightarrow [0, \pi]$, continue sur $[-1, 1]$ et strictement décroissante (bijective de $[-1, 1]$ dans $[0, \pi]$) ; on a

$$\Big((x \in [0, \pi]) \wedge (y = \cos x)\Big) \Longleftrightarrow \Big((y \in [-1, 1]) \wedge (x = \operatorname{Arcos} y)\Big) \ ;$$

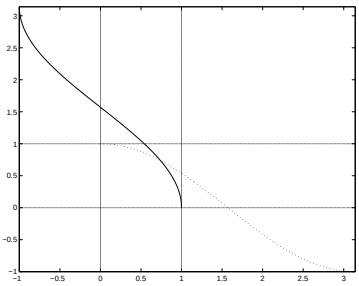

FIGURE 3.2 – Graphes de cos (en pointillés) et de Arcos (en plein)

• La fonction $x \longmapsto \exp x$ est strictement croissante de $\mathbb{R}$ dans $]0,+\infty[$; elle admet donc une fonction réciproque log $:]0,+\infty[\longrightarrow \mathbb{R}$, continue, strictement croissante (bijective de $]0,+\infty[$ dans $\mathbb{R}$) ; on a

$$\Big((x \in \mathbb{R}) \wedge (y = \exp(x))\Big) \Longleftrightarrow \Big((y \in]0,+\infty[) \wedge (x = \log y)\Big)\,.$$

Le graphe de la fonction exponentielle est "tourné' vers le haut, celui de la fonction logarithme regarde vers le bas. La fonction exponentielle est une fonction convexe, la fonction logarithme est une fonction concave. On verra plus tard pourquoi l'exponentielle impose toujours à l'infini sa limite aux fonctions puissances (comme $x \to x^2$) tandis que ce sont les fonctions puissances qui imposent leur limite au logarithme.

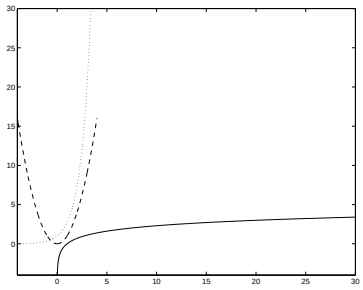

FIGURE 3.3 – Graphes de exp (en pointillés), de log (en plein) et de $x \to x^2$ (en tirets)

Soit f une application strictement monotone entre deux intervalles I et J de $\mathbb{R}$, comme sur la figure ci-dessous. Soit $G(f)$ le graphe de f et $G(f^{-1})$ le graphe de f^{-1}. On a l'assertion suivante :

$$\forall x \in \mathbb{R}\,,\ \forall y \in \mathbb{R}\,,\ \Big((x \in I) \wedge (y = f(x))\Big) \Longleftrightarrow \Big((y \in J) \wedge (x = f^{-1}(y))\Big)\,.$$

Comme

$$\begin{aligned} G(f) &= \{(x,y) \in I \times J\,;\, y = f(x)\} \\ G(f^{-1}) &= \{(y,x) \in J \times I\,;\, x = f^{-1}(y)\} \\ &= \{(y,x) \in J \times I\,;\, y = f(x)\}\,, \end{aligned}$$

on voit que le graphe de f^{-1} s'obtient à partir du graphe de f en prenant l'image de l'ensemble $G(f)$ par l'application linéaire de $\mathbb{R}^2$ dans $\mathbb{R}^2$ qui à (x,y) associe (y,x) ; cette application correspond à la

symétrie par rapport à la première bissectrice. Ceci reste vrai si l'on remplace I et J par deux ensembles D et E tels que $f\ : D \to E$ réalise une bijection entre D et E, d'inverse $f^{-1}\ : E \to D$.

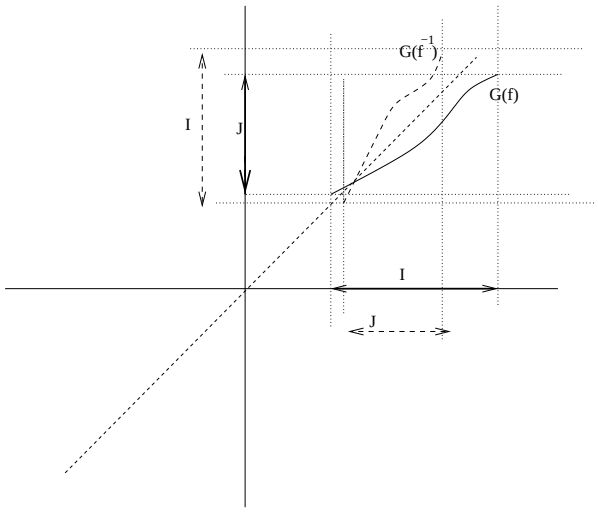

FIGURE 3.4 – Graphe de la fonction inverse

Ceci reste vrai dès que f est une bijection (non nécessairement monotone) entre deux sous-ensembles D et E de $\mathbb{R}$: le graphe de l'application inverse f^{-1} s'obtient toujours en prenant l'image du graphe de f par la symétrie par rapport à la première diagonale (la première diagonale est l'ensemble des points de $\mathbb{R}^2$ défini comme $\{(x,x)\,;\, x \in \mathbb{R}\}$). On verra des exemples dans la section 3.6.

3.5 Dérivabilité en un point et sur un intervalle

Les applications les plus simples (hormis les applications constantes) sont les applications dites *affines*, c'est-à-dire de la forme $f\ : x \in \mathbb{R} \longmapsto ax + b$; le graphe d'une telle application est une droite et l'on peut prédire immédiatement la valeur en un point x intermédiaire entre deux points distincts x_1 et x_2 dont on connaît les images par :

$$f(x) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}\,(x - x_1) + f(x_1)\,. \qquad (*)$$

Il est donc important de savoir "approcher" une fonction f quelconque par une fonction aussi simple, au moins près d'un point x_0 du domaine de définition de f. C'est ceci qui nous conduit à la notion de *dérivabilité* en un point x_0.

Définition 3.4. Soit f une fonction définie dans un voisinage V d'un point x_0 de $\mathbb{R}$ (c'est-à-dire au moins dans un intervalle ouvert $]x_0 - \eta_1, x_0 + \eta_2[$ contenant x_0). On dit que f est *dérivable au point* x_0 et admet comme nombre dérivé en x_0 le nombre $a(x_0)$ si et seulement si on peut écrire, pour $h \in]-\eta, \eta[$ avec $]x_0 - \eta, x_0 + \eta[\subset V$:

$$f(x_0 + h) = f(x_0) + a(x_0)h + h\epsilon(h)\,,$$

où la fonction ϵ définie dans $]-\eta, \eta[\backslash\{0\}$ par

$$\epsilon(h) := \frac{f(x_0 + h) - f(x_0) - a(x_0)h}{h}$$

est telle que

$$\lim_{h\to 0} \epsilon(h) = 0\,.$$

Ceci signifie qu'au voisinage de x_0, la fonction f ne diffère de la fonction affine

$$x \to a(x_0)(x - x_0) + f(x_0)$$

que par un "terme d'erreur" de la forme $(x-x_0)\epsilon(x-x_0)$ avec $\lim_{h\to 0} \epsilon(h) = 0$, c'est-à-dire un terme d'erreur négligeable par rapport à $|x - x_0|$ lorsque x se rapproche de x_0. On écrit un tel terme $o(|x - x_0|)$ pour signifier qu'il est négligeable devant $|x - x_0|$; ce sont les notations aujourd'hui communément utilisées (en mathématiques autant qu'en physique) introduites par le mathématicien allemand Landau (1877-1938), spécialiste de théorie analytique des nombres : une fonction φ définie au voisinage de 0 (sauf éventuellement en 0) et qui est un $o(1)$ est une fonction tendant vers 0 en 0 ; si $n \in \mathbb{N}$, c'est un $o(|h|^n)$ si $h \to \varphi(h)/|h|^n$ tend vers 0 lorsque h tend vers 0.

Si f est définie au voisinage de x_0 et que l'on examine le graphe de f au voisinage de x_0, dire que f est dérivable en x_0 signifie que la droite $D_{x_0,h}$ joignant les points $(x_0, f(x_0))$ et $(x_0 + h, f(x_0 + h))$ pour h non nul et proche de 0 tend vers une "droite limite", précisément la droite passant par (x_0, y_0) et de pente $a(x_0)$; en effet, le nombre

$$\frac{f(x_0 + h) - f(x_0)}{h}$$

qui représente la pente de la droite $D_{x_0,h}$ (voir figure ci-dessous) tend vers $a(x_0)$. Cette droite limite T_{x_0} d'équation $y = f(x_0) + a(x_0)(x - x_0)$ est dite *tangente géométrique* en $(x_0, f(x_0))$ au graphe de f tandis que l'application affine

$$x \to f(x_0) + a(x_0)(x - x_0)$$

est dite *application affine tangente* à f au point x_0. Pour calculer numériquement f près de x_0, on assimilera f à sa fonction affine tangente et l'on se ramènera à calculer les valeurs d'une fonction affine, ce qui se fait très rapidement dès que l'on connaît la valeur en deux points distincts (voir la formule $(*)$ ci-dessus). Attention ! Un graphe (comme celui de la fonction $x \longmapsto \sqrt{|x|}$ près de l'origine) peut présenter une tangente géométrique sans que la fonction soit dérivable : ceci se produit si la tangente géométrique s'avère être une droite verticale, comme sur l'exemple mentionné. Dire que la fonction f est dérivable en un point x_0 intérieur à son domaine de définition revient donc à dire que le graphe de F admet au point (x_0, y_0) une tangente géométrique non verticale ; la fonction $x \longmapsto |x|$ n'est par exemple pas dérivable en 0 car le graphe présente un point "anguleux", le point $(0, 0)$: il y a en effet une demi-tangente géométrique à droite $(y = x)$, une demi-tangente géométrique à gauche $(y = -x)$, mais pas de tangente car ces deux droites sont distinctes !

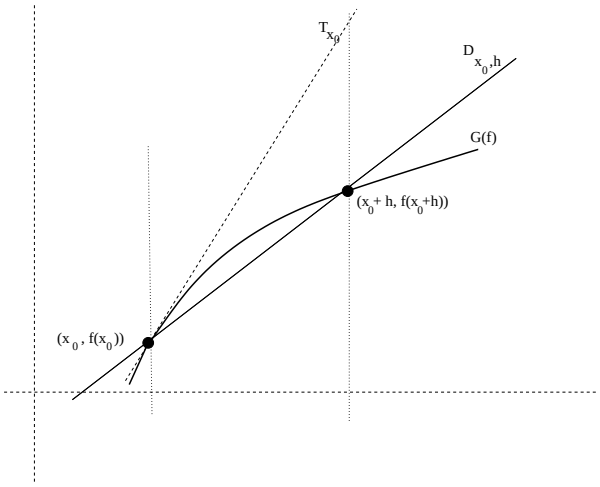

FIGURE 3.5 – Tangente géométrique

Lorsque f est définie au moins dans un intervalle semi-ouvert $[x_0, x_0+\eta[$, on dit de même que f *admet une dérivée à droite* en x_0 avec pour dérivée à droite en x_0 le nombre réel $a_+(x_0)$ si et seulement si pour $h \in]0, \eta[$

$$f(x_0+h) = f(x_0) + a_+(x_0)h + h\epsilon_+(h)\,,$$

où la fonction ϵ_+ définie dans $]0, \eta[$ par

$$\epsilon_+(h) := \frac{f(x_0+h) - f(x_0) - a_+(x_0)h}{h}$$

est telle que

$$\lim_{h\to 0_+} \epsilon_+(h) = 0\,.$$

Lorsque f est définie au moins dans un intervalle semi-ouvert $]x_0-\eta, x_0]$, on dit aussi que f *admet une dérivée à gauche* en x_0 avec pour dérivée à gauche en x_0 le nombre réel $a_-(x_0)$ si et seulement si pour $h \in]0, \eta[$

$$f(x_0-h) = f(x_0) - a_-(x_0)h - h\epsilon_-(h)\,,$$

où la fonction ϵ_- définie dans $]0, \eta[$ par

$$\epsilon_-(h) := \frac{f(x_0) - f(x_0-h) - a_-(x_0)h}{h}$$

est telle que

$$\lim_{h\to 0_+} \epsilon_-(h) = 0\,.$$

Si f est définie au moins dans $[x_0, x_0+\epsilon[$ et admet une dérivée à droite en x_0, la droite limite des droites $D_{x_0,h}$ avec $h > 0$ est dite *demi-tangente à droite* en $(x_0, f(x_0))$ au graphe de f ; c'est la droite d'équation $y = a_+(x_0)(x-x_0) + f(x_0)$. On a une définition analogue pour la demi-tangente à gauche si f est définie au moins dans $]x_0-\epsilon, x_0]$ et a une dérivée à gauche en x_0.

Proposition 3.9. *Une fonction définie au voisinage de x_0 et dérivable en x_0 est continue en x_0. Ceci vaut aussi si la fonction est au moins définie d'un côté de x_0 et admet une dérivée (du côté où elle est définie) en x_0.*

Preuve. On fait la preuve dans le premier cas (f définie dans un voisinage V de x_0 et dérivable en x_0). On peut écrire alors, pour x dans $]x_0 - \eta, x_0 + \eta[\backslash\{x_0\}$,

$$f(x) = f(x_0) + a(x_0)(x - x_0) + (x - x_0)\epsilon(x - x_0)$$

avec $\lim_{h\to 0} \epsilon(h) = 0$. Pour $|x - x_0| \leq \eta_1$, on peut affirmer que $|\epsilon(x - x_0)| < 1$, donc que

$$|f(x) - f(x_0)| \leq (|a(x_0)| + 1)\,|x - x_0|\,;$$

si l'on choisit $|x - x_0| < \inf(\eta_1, \alpha/(|a(x_0)| + 1))$ avec $\alpha > 0$ arbitraire, on voit que $|f(x) - f(x_0)| < \alpha$, ce qui prouve

$$\lim_{x\to\alpha} f(x) = f(x_0)\,,$$

donc que f est continue en x_0. □

La réciproque de cette proposition est fausse (par exemple $x \to |x|$ n'est pas dérivable en $x = 0$). Il existe même des fonctions continues sur un intervalle et dérivables en aucun point de cet intervalle ; ce sont par exemple les structures fractales (pensez au bord d'un flocon de neige !). On doit au mathématicien tchèque Bernard Bolzano (1781-1848), au mathématicien allemand Karl Weierstrass (1815-1897), au suédois Von Koch (1870-1924), au français H. Lebesgue (1875-1941), nombre de tels exemples. L'exemple de Bolzano est présenté dans le chapitre 2 du support de cours. Voici par exemple le début du cheminement conduisant à la construction du flocon de Von Koch : pour chaque segment, on retire le segment médian et on le remplace par la ligne brisée composée des deux côtés du triangle équilatéral s'appuyant sur le segment médian.

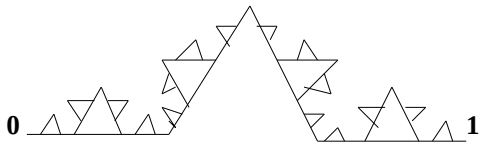

FIGURE 3.6 – Construction du flocon de von Koch (étape 2)

Avec la fonction définie par

$$t \in [0,1] \longmapsto \lim_{n\to+\infty} \sum_{k=1}^{n} \frac{\sin(\pi k^2 t)}{k^2}\,,$$

B. Riemann avait aussi proposé au milieu du XIX-ème siècle un autre exemple intéressant de fonction continue partout et dérivable nulle part. Le graphe de cette fonction, hérissé partout de "piquants" est reproduit ci-dessous :

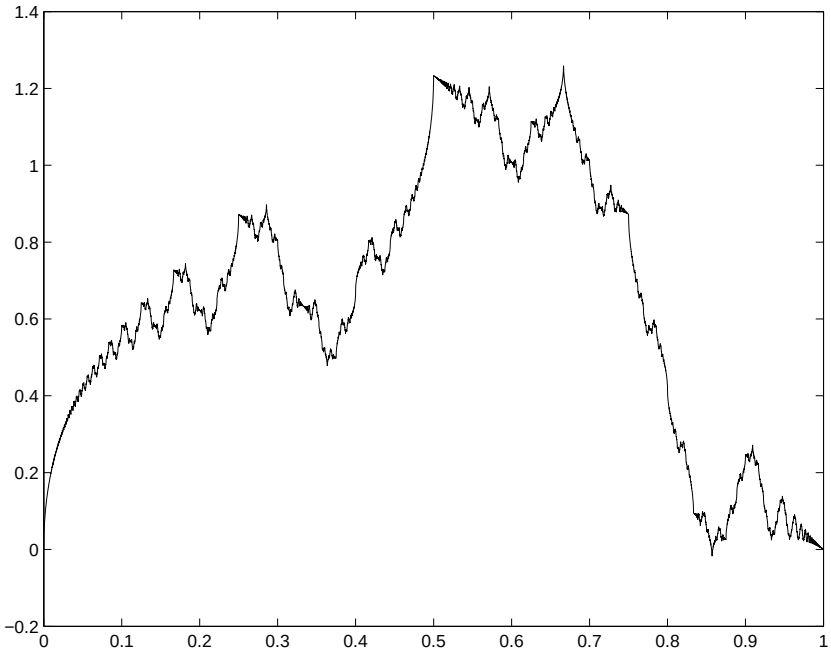

FIGURE 3.7 – La fonction de Riemann : un exemple de graphe fractal

Notons aussi que l'on peut parler de continuité en un point quelconque du domaine de définition d'une fonction alors que l'on ne parle de dérivabilité qu'en un point intérieur à ce domaine de définition ; ceci est plus restrictif !

Définition 3.10. *Si I est un intervalle ouvert de $\mathbb{R}$ (ou plus généralement un sous-ensemble ouvert de $\mathbb{R}$), une fonction f définie sur I est dérivable sur I si et seulement si elle est dérivable en tout point de I. En tout point x_0 de I où la dérivée existe, on la note $f'(x_0)$ et l'on définit ainsi une nouvelle fonction sur l'ensemble des points de I où f est dérivable ; cette nouvelle fonction est dite fonction dérivée de f sur I.*

On vérifiera les règles de calcul suivantes :

- si f et g sont définies au voisinage de x_0 et dérivables en x_0, $f+g$ l'est aussi avec $(f+g)'(x_0) = f'(x_0) + g'(x_0)$;
- si f et g sont définies au voisinage de x_0 et dérivables en x_0, fg l'est aussi car on peut écrire :

$$\begin{aligned} f(x_0+h)g(x_0+h) &= (f(x_0) + hf'(x_0) + o(|h|)) \times (g(x_0) + g'(x_0)h + o(|h|)) \\ &= f(x_0g(x_0) + h(f(x_0)g'(x_0) + g(x_0)f'(x_0)) + o(|h|) \end{aligned}$$

  et l'on voit que le nombre dérivé est $(fg)'(x_0) = f(x_0)g'(x_0) + g(x_0)f'(x_0)$.
- la fonction $x \to x^n$, avec $n \in \mathbb{Z}$ est dérivable sur $\mathbb{R} \setminus \{0\}$ (sur $\mathbb{R}$ si $n \in \mathbb{N}$) et sa fonction dérivée est :

$$x \longrightarrow nx^{n-1} ;$$

si $n \geq 0$, on a en effet, avec la formule du binôme

$$\begin{aligned}(x_0+h)^n &= x_0^n + nhx_0^{n-1} + \sum_{k=2}^{n} \binom{n}{k} h^k x_0^{n-k} \\ &= x_0^n + nx_0^{n-1}h + o(|h|)\,,\end{aligned}$$

d'où la dérivabilité de $f_n : x \to x^n$ avec $f_n'(x_0) = nx_0^{n-1}$. Si $n = -m < 0$, on écrit, si $x_0 \neq 0$, toujours en utilisant la formule du binôme,

$$\begin{aligned}\frac{1}{(x_0+h)^m} &= \frac{1}{x_0^m} - \frac{(x_0+h)^m - x_0^m}{(x_0+h)^m x_0^m} \\ &= \frac{1}{x_0^m} - h\frac{mx_0^{m-1} + o(1)}{x_0^{2m} + o(1)} \\ &= \frac{1}{x_0^m} - hmx_0^{-m-1} + o(|h|)\,,\end{aligned}$$

d'où le fait que $x \to x^n$ est dérivable (si $n < 0$) en tout point de $\mathbb{R} \setminus \{0\}$ avec comme dérivée la fonction $x \to nx^{n-1}$.

Concernant la composition des fonctions, on a la règle essentielle, dite de Leibniz (Gottfried Wilhelm von Leibniz, philosophe et mathématicien allemand, 1646-1716) que les anglo-saxons appellent aussi *chain rule* :

Théorème (règle de Leibniz) *Soit f une fonction définie au voisinage de x_0, dérivable en x_0 ; soit g une fonction définie au voisinage de $f(x_0)$, dérivable en $f(x_0)$; alors $g \circ f$ est dérivable en x_0, de dérivée $g'(f(x_0)) \times f'(x_0)$.*

Preuve. On a, pour H voisin de 0,

$$g(f(x_0)+H) = g(f(x_0)) + g'(f(x_0))H + o(|H|)$$

(car g est dérivable en $f(x_0)$). Mais on a aussi, pour h voisin de 0,

$$f(x_0+h) = f(x_0) + hf'(x_0) + o(|h|)$$

puisque f est dérivable en x_0. On a donc, en combinant les deux choses que, pour h voisin de 0,

$$\begin{aligned}(g\circ f)(x_0+h) &= g(f(x_0)+hf'(x_0)+o(|h|)) = g(f(x_0)+[hf'(x_0)+o(|h|)]) \\ &= g(f(x_0)) + g'(f(x_0)) \times (hf'(x_0)+o(|h|)) + o(|hf'(x_0)+o(|h|)|) \\ &= g(f(x_0)) + hg'(f(x_0))f'(x_0) + o(|h|)\,,\end{aligned}$$

d'où le résultat voulu. □

Applications :

- Si f est dérivable en x_0 de dérivée $f'(x_0)$ et que f ne s'annule pas en x_0, la fonction $1/f$ est dérivable en x_0, de dérivée
$$(1/f)'(x_0) = -\frac{f'(x_0)}{f^2(x_0)}$$
(on compose f avec $y \to 1/y$).
- Si f et g sont dérivables en x_0 et que g ne s'annule pas en x_0, f/g est aussi dérivable en x_0 et
$$(f/g)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g^2(x_0)}$$
(on combine le résultat pour $1/g$ et le résultat pour le produit de f et $1/g$).

– Si f est bijective d'un intervalle ouvert I contenant x_0 dans un intervalle J contenant $f(x_0)$ et que f et f^{-1} sont dérivables, l'une en x_0, l'autre en $f(x_0)$, alors on a la relation

$$(f^{-1})'(f(x_0)) \times f'(x_0) = 1\,,$$

en particulier $f'(x_0) \neq 0$ (on a $g \circ f = \mathrm{Id}$).

On a le résultat suivant :

Proposition 3.11. *Soit f une fonction strictement monotone sur un intervalle ouvert I, dérivable en tout point de I et telle que $f'(x_0) \neq 0$ pour tout $x_0 \in I$ (en fait $f'(x_0) > 0$ pour tout x_0 de I si f est strictement croissante, $f'(x_0) < 0$ pour tout x_0 de I si f est strictement décroissante) ; la fonction continue $f^{-1} : f(I) \to I$ est alors dérivable en tout point y_0 de $f(I)$ avec*

$$(f^{-1})'(y_0) = \frac{1}{f'(f^{-1}(y_0))}\,.$$

Preuve. On utilise la caractérisation géométrique de la dérivabilité (à savoir l'existence au point du graphe d'une tangente géométrique de pente le nombre dérivé). Le graphe de f a au point (x_0, y_0) une tangente géométrique, à savoir la droite d'équation $y = f(x_0) + (x - x_0)f'(x_0)$. Si l'on transforme la figure par symétrie par rapport à la première diagonale, on constate que le graphe de f^{-1} admet au point $(y_0, x_0) = (y_0, f^{-1}(y_0))$ pour tangente géométrique la droite d'équation $x = f(x_0)+(y-x_0)f'(x_0)$, soit encore

$$\begin{aligned} y &= x_0 + \frac{x - f(x_0)}{f'(x_0)} = f^{-1}(y_0) + \frac{x - f(x_0)}{f'(x_0)} \\ &= f^{-1}(y_0) + \frac{x - f(f^{-1}(y_0))}{f'(x_0)} \\ &= f^{-1}(y_0) + \frac{x - y_0}{f'(f^{-1}(y_0))}\,. \end{aligned}$$

Ceci montre que f^{-1} est dérivable au point y_0, de dérivée $1/f'(f^{-1}(y_0))$. □

Ce résultat nous permettra d'étudier le comportement des inverses de certaines fonctions classiques. À l'opposé de ce résultat, on a la proposition suivante, aussi très importante et que l'on admettra pour l'instant (cela résultera du théorème fondamental de l'analyse dont on parlera dans la section 3.7) :

Proposition 3.12. *Soit f une fonction dérivable sur un intervalle ouvert I, de dérivée identiquement nulle sur cet intervalle ; la fonction f est alors constante sur I.*

Remarque. Il faut cependant prendre garde à l'intuition ! Il existe en effet des fonctions continues, croissantes et surjectives de $[0, 1]$ dans $[0, 1]$, dérivables et de dérivée nulle en tous les points de $[0, 1]$ hormis un ensemble au plus dénombrable de points (où la fonction n'est pas dérivable) ! Ce sont les célèbres *escaliers du diable* (voir par exemple la figure ci-dessous où nous avons introduit les trois premières fonctions d'une suite approchant un tel escalier.

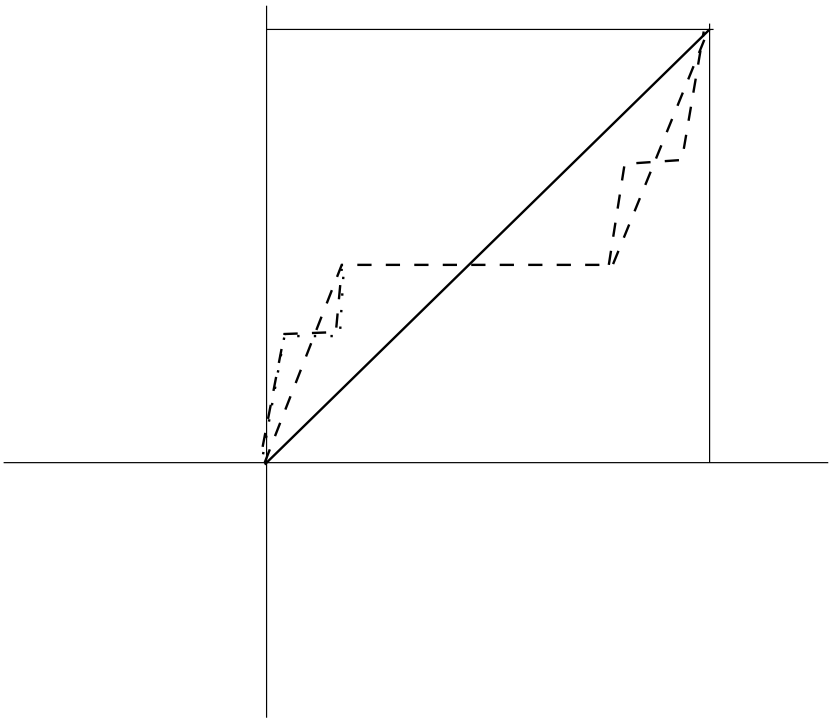

FIGURE 3.8 – Une approche d'*escalier du diable*

3.6 Quelques fonctions classiques et leurs inverses

3.6.1 La fonction exponentielle et le logarithme népérien

À la question naturelle : peut-on construire sur $[0, +\infty[$ une fonction qui domine toutes les fonctions $x \to x^n$, $n \in \mathbb{N}$?, on aimerait répondre (de manière simpliste) en posant

$$f(x) := \lim_{n \to +\infty} \sum_{k=0}^{n} x^k \, ;$$

mais on sait malheureusement que

$$\sum_{k=0}^{n} x^k = \begin{cases} \dfrac{x^{n+1} - 1}{x - 1} \text{ si } x \neq 1 \\ n + 1 \text{ si } x = 1 \end{cases}$$

et que cette quantité tend vers $+\infty$ lorsque n tend vers $+\infty$ (lorsque $x > 1$). Tout n'est cependant pas désespéré car l'on pourrait, pour corriger le tir, introduire, si x un nombre réel positif ou nul, la suite $(u_n(x))_{n \geq 0}$ de terme général

$$u_n(x) := \frac{x^n}{n!} \, .$$

On a

$$u_{n+1}(x) = \frac{x}{n+1} u_n(x) \, ,$$

ce qui implique, dès que n est supérieur où égal à la partie entière $E[2x]$ de $2x$, que

$$u_{n+1}(x) \leq \frac{u_n(x)}{2}\,.$$

La suite de terme général

$$U_n(x) := \sum_{k=0}^{n} u_k(x) = \sum_{k=0}^{n} \frac{x^k}{k!}$$

est une suite croissante de nombres réels majorée car, pour $n > E[2x]$, on a

$$\begin{aligned} U_n(x) &= \sum_{k=0}^{E[2x]} \frac{x^k}{k!} + \sum_{k=E[2x]+1}^{n} \frac{x^k}{k!} \\ &\leq \sum_{k=0}^{E[2x]} \frac{x^k}{k!} + \frac{x^{E[2x]+1}}{(E[2x]+1)!} \sum_{k=0}^{n-E[2x]-1} \frac{1}{2^k} \\ &\leq \sum_{k=0}^{E[2x]+1} \frac{x^k}{k!} + 2\frac{x^{E[2x]+1}}{(E[2x]+1)!}\,. \end{aligned}$$

La suite $(U_n(x))_{n\geq 0}$ est donc une suite de nombres réels croissante majorée, donc convergente vers une limite que l'on convient d'appeler $\exp x$ ou e^x.

Si maintenant $x < 0$ et si $(U_n(x))_{n\geq 0}$ désigne toujours la suite de terme général

$$U_n(x) := \sum_{k=0}^{n} u_k(x) = \sum_{k=0}^{n} \frac{x^n}{n!}\,,$$

on vérifie que la suite $(U_{2p}(x))_{p\geq 0}$ est une suite croissante, que la suite $(U_{2p+1}(x))_{p\geq 0}$ est une suite décroissante et que la suite de terme général $U_{2p+1}(x) - U_{2p}(x)$ tend vers 0 lorsque p tend vers l'infini. Les deux suites $(U_{2p}(x))_{p\geq 0}$ et $(U_{2p+1}(x))_{p\geq 0}$ sont donc des suites adjacentes ayant une limite commune et la suite $(U_n(x))_{n\geq 0}$ converge donc vers une limite (la limite commune des deux suites ci-dessus), limite que l'on appelle encore $\exp x$.

Définition 3.11. *La fonction exponentielle $x \longmapsto \exp x$ est la fonction définie sur $\mathbb{R}$ par*

$$\exp x = e^x := \lim_{n\to+\infty} \sum_{k=0}^{n} \frac{x^k}{k!}\,.$$

Grâce à la formule du binôme, on a, si x_1 et x_2 sont deux nombres réels et n un entier strictement positif

$$\Big(\sum_{k=0}^{n} \frac{x_1^k}{k!}\Big) \times \Big(\sum_{l=0}^{n} \frac{x_2^l}{l!}\Big) = \sum_{m=0}^{n} \frac{(x_1+x_2)^m}{m!} + \sum_{\substack{k_1,\,k_2\,\in\,\mathbb{N} \\ n\,<\,k_1+k_2\,\leq\,2n}} \frac{x_1^{k_1} x_2^{k_2}}{k_1!\,k_2!}\,;$$

on vérifiera en exercice que la suite de terme général

$$V_n(x_1,x_2) := \sum_{\substack{k_1,\,k_2\,\in\,\mathbb{N} \\ n\,<\,k_1+k_2\,\leq\,2n}} \frac{x_1^{k_1} x_2^{k_2}}{k_1!\,k_2!}$$

tend vers 0 lorsque n tend vers l'infini car

$$\begin{aligned} |V_n(x_1,x_2)| &\leq \Big(\sum_{k_1=E[n/2]+1}^{2n} \frac{|x_1|^{k_1}}{k_1!}\Big)\times e^{|x_2|} + \Big(\sum_{k_2=E[n/2]+1}^{2n} \frac{|x_2|^{k_2}}{k_2!}\Big)\times e^{|x_1|} \\ &\leq (e^{|x_1|} - U_{E[n/2]}(|x_1|))e^{|x_2|} + (e^{|y|} - U_{E[n/2]}(|x_2|))e^{|x_1|}\,, \end{aligned}$$

d'où l'on déduit la formule majeure :

$$\forall x_1 \in \mathbb{R}\,,\ \forall x_2 \in \mathbb{R}\,,\ \exp(x_1+x_2) = \exp(x_1)\times\exp(x_2)\,. \qquad (\dagger)$$

Cette formule implique que la fonction exponentielle ne s'annule jamais et reste strictement positive puisque $e^x \geq 1$ si $x \geq 0$ et que $e^x = 1/e^{-x} \in]0,1[$ si $x < 0$. De plus, pour $x \geq 0$, la fonction exponentielle est telle que

$$\frac{x^{n+1}}{(n+1)!} \leq e^x$$

pour tout $n \in \mathbb{N}$, ce qui implique que pour tout entier $n \in \mathbb{N}$,

$$\begin{aligned} \lim_{x\to+\infty} \frac{e^x}{x^n} &= +\infty \\ \lim_{x\to-\infty} e^x x^n &= 0\,, \end{aligned}$$

ce que l'on résume en général en disant que *l'exponentielle impose sa limite aux fonctions puissances.*

On peut écrire, pour tout $h \in \mathbb{R}$,

$$\begin{aligned} \exp(h) &= 1 + h + \Big(h^2 \times \lim_{n\to+\infty} \sum_{k=2}^{n} \frac{h^{k-2}}{k!}\Big) \\ &= 1 + h + \Big(h^2 \times \lim_{n\to+\infty} \sum_{k=0}^{n-2} \frac{h^{k}}{(k+2)!}\Big)\,; \end{aligned}$$

comme

$$\Big|\sum_{k=0}^{n-2} \frac{h^k}{(k+2)!}\Big| \leq \sum_{k=0}^{n-2} \frac{|h^k|}{k!} \leq e^{|h|} \leq e$$

si $|h| \leq 1$, la fonction exponentielle est dérivable en $x = 0$, de dérivée 1 en ce point. Comme

$$\frac{e^{x+h} - e^x}{h} = e^x \times \Big(\frac{e^h - 1}{h}\Big)$$

pour tout $h \in \mathbb{R}\setminus\{0\}$, la fonction exponentielle est dérivable (donc continue) en tout point x de $\mathbb{R}$ avec

$$\frac{d}{dx}[e^x] = e^x\,.$$

La fonction exponentielle s'auto-dérive en elle même (on verra que la seule fonction dérivable sur un intervalle de $\mathbb{R}$ et se dérivant en elle-même est la fonction exponentielle), d'où son importance majeure dans les problèmes d'évolution en physique ou en biologie (par exemple les problèmes liés aux processus de désintégration atomique ou aux processus d'évolution de population du type "proie-prédateur"). On y reviendra.

Une autre approche de la fonction exponentielle consiste à remarquer (on le justifiera plus loin) qu'on l'obtient comme la limite suivante :

$$\forall x \in \mathbb{R}\,,\ \exp x = \lim_{n\to+\infty} \left(1+\frac{x}{n}\right)^n . \qquad (*)$$

On justifiera cette formule $(*)$ plus loin. On peut cependant expliquer comment on peut la deviner en raisonnant comme suit. On aurait pu deviner la formule $(*)$ pour $x > 0$ en essayant de modéliser de manière numérique la recherche d'une fonction dérivable au voisinage de $[0, x]$ et s'auto-dérivant en elle-même. Si la fonction (inconnue, mais supposée valant 1 en $x = 0$) est échantillonnée aux points $0, x/N, ..., (N-1)x/N, x, ...$ avec N entier très grand et que l'on pose $u_k = f(kx/N)$, $k = 0, ..., N$, on voit que la suite $(u_k)_k$ doit se plier à la règle récurrente :

$$u_{k+1} - u_k = f((k+1)x/N) - f(kx/N) \simeq \frac{x}{N} f(kx/N) = \frac{x}{N}\, u_k\,;$$

partant de $u_0 = 1$, on aboutit à

$$u_k = \left(1+\frac{x}{N}\right)^k ;$$

en particulier $u_N = \left(1+\frac{x}{N}\right)^N$. Si l'on choisit le pas x/N de plus en plus petit pour "pister" au mieux la fonction f, on voit que le calcul nous conduit naturellement vers la construction de la fonction exponentielle. On attribue au mathématicien suisse Leonhard Euler (1707-1783) cette méthode devenue le prototype de la méthode numérique dite *des éléments finis* permettant la résolution des équations différentielles dont $y' = y$ est le prototype.

La fonction exponentielle est une fonction strictement croissante sur $\mathbb{R}$ car si $x_2 > x_1$,

$$e^{x_2} - e^{x_1} = e^{x_1}(e^{x_2-x_1} - 1) \geq (x_2 - x_1)e^{x_2-x_1} > 0$$

et, comme fonction bijective de $\mathbb{R}$ dans $]0, +\infty[$, admet donc une fonction réciproque (dite logarithme népérien) :

Définition 3.12. *La fonction logarithme népérien est la fonction bijective de* $]0, +\infty[$ *dans* $\mathbb{R}$ *définie comme la fonction réciproque de la fonction exponentielle, c'est-à-dire :*

$$\Big((x \in \mathbb{R}) \wedge (y = \exp x)\Big) \Longleftrightarrow \Big((y \in]0, +\infty[) \wedge (x = \log y)\Big)\,.$$

La fonction logarithme $y \longmapsto \log y$ est une fonction strictement croissante sur $]0, +\infty[$, dérivable d'après la proposition 3.11, et de dérivée la fonction

$$y \to \frac{1}{y}\,.$$

Plus généralement, à cause de la règle de Leibniz, si a est un nombre réel, la fonction composée

$$y \to \log|y - a|$$

est une fonction dérivable sur $\mathbb{R} \setminus \{a\}$, de dérivée en tout point de cet ensemble ouvert la fonction

$$y \to \frac{1}{y-a}$$

(on fera la vérification pour $x > a$ et $x < a$). Il est très important de remarquer que si a est une nombre réel et n un entier relatif, les fonctions f dérivables sur $\mathbb{R} \setminus \{a\}$ et telles que

$$\forall t \in \mathbb{R} \setminus \{a\}\,,\ f'(x) = (x-a)^n$$

sont les fonctions du type :

$$f(x) = \frac{(x-a)^{n+1}}{n+1} + C\,,\ C \in \mathbb{R}\,,$$

si $n \neq -1$ (ce sont donc des fractions rationnelles, comme $x \to (x-a)^n$) tandis que ce sont les fonctions du type

$$f(x) = \log|x-a| + C\,,\ C \in \mathbb{R},,$$

si $n = -1$; une fonction qui n'a rien de rationnel, la fonction logarithme, vient donc subrepticement s'immiscer dans un monde algébrique, celui des fractions rationnelles ; ce fait est un fait capital en mathématiques ; c'est aussi une des raisons pour lesquelles le logarithme a joué très tôt un rôle essentiel en mathématiques car il s'agit d'une fonction intervenant à la croisée de l'algèbre et de l'analyse. L'intérêt de l'exponentielle réside dans son rôle crucial en analyse et en modélisation (phénomènes d'évolution) mais le fait qu'elle échange addition et multiplication fait d'elle (comme d'ailleurs le logarithme), un outil majeur de calcul.

Par réciprocité avec la formule (†) vérifiée par la fonction exponentielle, la fonction logarithme satisfait sur $]0,+\infty[$ la formule majeure suivante :

$$\forall y_1 \in]0,+\infty[\,,\ \forall y_2 \in]0,+\infty[\,,\ \log(y_1 y_2) = \log y_1 + \log y_2\,. \qquad (\dagger\dagger)$$

Au niveau des limites en $+\infty$ et en 0, on a cette fois :

$$\lim_{y\to 0} |y|\log|y| \quad = \quad 0$$
$$\lim_{y\to+\infty} \frac{\log y}{y} = 0\,.$$

Le fait que la fonction logarithme soit dérivable en 1 et de dérivée 1 assure que pour tout $x \in \mathbb{R}$, on a, pour n assez grand

$$\log(1+\frac{x}{n}) = \frac{x}{n} + o_x(1/n)\,;$$

en multipliant par n et en prenant l'exponentielle, on a

$$\left(1+\frac{x}{n}\right)^n = \exp(x + n o_x(1/n))\,,$$

ce qui prouve la formule déjà annoncée $(*)$

$$\lim_{n\to+\infty} \left(1+\frac{x}{n}\right)^n = e^x$$

pour tout $x \in \mathbb{R}$ et donne ainsi une autre manière d'approcher l'exponentielle par des fonctions polynômes (si $x > 0$, cette approximation se fait d'ailleurs de manière croissante "par en dessous").

Couplées avec la fonction exponentielle, on introduit, pour $a > 0$ la fonction

$$x \in \mathbb{R} \longmapsto a^x := e^{x\log a}\,,$$

fonction dérivable sur $\mathbb{R}$ et de dérivée

$$\frac{d}{dx}a^x = \log a \times a^x\,.$$

On a bien sûr les formules

$$\begin{aligned}(a^{x_1})^{x_2} &= a^{x_1 x_2} \quad (a>0, x_1 \in \mathbb{R}, x_2 \in \mathbb{R})\\ a^{x_1+x_2} &= a^{x_1} \times a^{x_2} \quad (a>0, x_1 \in \mathbb{R}, x_2 \in \mathbb{R})\\ (ab)^x &= a^x \times b^x \quad (a>0, b>0, x \in \mathbb{R})\\ a^{-x} &= (1/a)^x \quad (a>0,\ x \in \mathbb{R})\,.\end{aligned}$$

3.6.2 Fonctions trigonométriques et leurs inverses

On a déjà rencontré en fait les fonctions cos et sin et vous savez qu'il s'agit de fonctions périodiques de période 2π. Mais nous allons ici introduire ces deux fonctions autrement, ainsi d'ailleurs que le nombre π. Oublions un instant la "géométrie" et le fait que x soit la mesure en radians d'un angle.

La fonction $x \to \cos x$ sera la fonction paire définie pour $x \geq 0$ par

$$\cos x := \lim_{n\to+\infty} \Big(\sum_{k=0}^{n} (-1)^k \frac{x^{2k}}{(2k)!} \Big) = \lim_{n\to+\infty} A_n(x)$$

(en prenant le deux suites $(A_{2p})_{p\geq 0}$ et $(A_{2p+1})_{p\geq 0}$, on a affaire à deux suites adjacentes ayant même limite). La fonction cos est ainsi dérivable en $x = 0$ et de dérivée 0.

La fonction $x \to \sin x$ sera, elle, la fonction impaire définie pour $x \geq 0$ par

$$\sin x := \lim_{n\to+\infty} \Big(\sum_{k=0}^{n} (-1)^k \frac{x^{2k+1}}{(2k+1)!} \Big) = \lim_{n\to+\infty} B_n(x)$$

(en prenant les deux suites $(B_{2p})_{p\geq 0}$ et $(B_{2p+1})_{p\geq 0}$, on a affaire à deux suites adjacentes ayant même limite). La fonction sin est ainsi dérivable en $x = 0$ et de dérivée 1.

La formule du binôme (c'est un peu plus compliqué que pour l'exponentielle et on l'admettra) implique les deux relations clef

$$\begin{aligned}\cos(x_1+x_2) &= \cos x_1 \cos x_2 - \sin x_1 \sin x_2\\ \sin(x_1+x_2) &= \cos x_1 \sin x_2 + \sin x_1 \cos x_2\,.\end{aligned} \tag{3.1}$$

Il en résulte que la fonction cos est dérivable sur $\mathbb{R}$, avec

$$\cos' = -\sin$$

et que la fonction sin est dérivable sur $\mathbb{R}$, de dérivée

$$\sin' = \cos\,.$$

On constate (du fait que $\cos x$ est atteint comme limite de suites adjacentes) que

$$\begin{aligned}\cos 0 &= 1 > 0\\ \cos 2 &\leq 1 - \frac{2^2}{2!} + \frac{2^4}{4!} = 1 - 2 + \frac{16}{24} = -1/3 < 0\,;\end{aligned}$$

d'après le théorème des valeurs intermédiaires

$$\{x \in [0,2]\,;\, \cos x = 0\} \neq \emptyset$$

et l'on défini le nombre π de la manière suivante :

$$\frac{\pi}{2} = \inf\{x \in [0,2]\,;\, \cos x = 0\}\,.$$

On a en particulier $\cos(\pi/2) = 0$.

En dérivant la fonction

$$x \longmapsto \cos^2 x + \sin^2 x$$

comme produit de fonctions dérivables, on voit que la dérivée est identiquement nulle et que la fonction f est donc constante (d'après la proposition 3.12) égale à 1 ; la relation

$$\cos^2 + \sin^2 \equiv 1$$

nous assure donc, puisque $\cos(\pi/2) = 0$, que $\sin(\pi/2) = \pm 1$; mais comme la fonction cos est croissante sur $[0, \pi/2]$ (car de dérivée sin positive sur $[0, \pi/2]$), on a nécessairement $\sin(\pi/2) = 1$.

Armés des deux relations (3.1), nous constatons que les fonctions cos et sin sont périodiques de période 2π, ce qui signifie

$$\begin{aligned} \forall x \in \mathbb{R}\,,\ \cos(x + 2\pi) &= \cos(x) \\ \forall x \in \mathbb{R}\,,\ \sin(x + 2\pi) &= \sin(x)\,. \end{aligned}$$

On constate aussi (si l'on regarde le sens de variations des deux fonctions cos et sin) que l'application

$$x \in [0, 2\pi[\longmapsto (\cos x, \sin x)$$

permet de paramétrer de manière bijective le cercle unité de $\mathbb{R}^2$, les deux fonctions coordonnées ainsi définies correspondant aux prises de lignes trigonométriques cos et sin d'un angle (exprimé en radians) au sens de la trigonométrie usuelle.

Les fonctions $\cos x$ et $\sin x$ représentent donc bien les lignes trigonométriques du nombre réel x.

Le nombre $\pi/2$ (premier zéro strictement positif de la fonction continue $x \to \cos x$) correspond à la mesure de l'angle droit orienté dans le sens trigonométrique, et donc, on le verra plus loin, à la longueur du quart de périmètre de cercle unité, la longueur étant calculée par approximations du quart de circonférence par des lignes polygonales. Les fonctions cos et sin que l'on vient d'introduire, ainsi d'ailleurs que le nombre π sont bien celles que l'on connaissait déjà !

On a vu que la fonction sin est strictement croissante de $[-\pi/2, \pi/2]$ dans $[-1, 1]$ et admet une fonction inverse strictement croissante $\mathrm{Arcsin} : [-1, 1] \longmapsto [-\pi/2, \pi/2]$. D'après la proposition 3.11, la fonction

$$y \longmapsto \mathrm{Arcsin}\, y$$

est dérivable sur $]-1, 1[$, de dérivée

$$y \to \frac{1}{\cos(\mathrm{Arcsin}\, y)} = \frac{1}{\sqrt{1 - y^2}}\,.$$

La fonction cos est strictement décroissante de $[0, \pi]$ dans $[-1, 1]$ et admet une fonction inverse strictement décroissante $\mathrm{Arcos} : [-1, 1] \longmapsto [0, \pi]$. D'après la proposition 3.11, la fonction

$$y \longmapsto \mathrm{Arcos}\, y$$

est dérivable sur $]-1,1[$, de dérivée

$$y \to -\frac{1}{\sin(\mathrm{Arcos}\, y)} = -\frac{1}{\sqrt{1-y^2}}\,.$$

Le fait que les dérivées de Arcsin et Arcos soient opposées résulte de la formule

$$\mathrm{Arcsin}\, y + \mathrm{Arcos}\, y \equiv \frac{\pi}{2} \;\; \forall y \in [-1,1]\,.$$

Une troisième fonction trigonométrique est importante : c'est la fonction tangente, définie par

$$\tan x := \frac{\sin x}{\cos x}\,,\;\; x \in]\pi/2, \pi/2[\,,$$

fonction strictement croissante de $]-\pi/2,\pi/2[$ dans $\mathbb{R}$ (avec

$$\begin{aligned} \lim_{x\to\pi/2_-} \tan x &= +\infty \\ \lim_{x\to-\pi/2_+} \tan x &= -\infty\,). \end{aligned}$$

La fonction tangente se prolonge par π-périodicité en une fonction continue (surjective, bien sûr non injective) de $E = \mathbb{R} \setminus \{(2k+1)\pi/2\,;\, k\mathbb{Z}\}$ dans $\mathbb{R}$. La fonction $x \longmapsto \tan x$ ainsi prolongée est dérivable sur E, de dérivée

$$x \to \tan'(x) = 1 + \tan^2 x\,.$$

Elle admet sur $]-\pi/2,\pi/2[$ une fonction réciproque, strictement croissante de $\mathbb{R}$ dans $]-\pi/2,\pi/2[$, la fonction

$$y \longmapsto \mathrm{Arctan}\, y$$

définie par l'équivalence

$$\Big((x \in]-\pi/2,\pi/2[) \wedge (y = \tan x)\Big) \Longleftrightarrow \Big((y \in \mathbb{R}) \wedge (x = \mathrm{Arctan}\, y)\Big)\,.$$

On a

$$\begin{aligned} \lim_{x\to-\infty} \mathrm{Arctan}\, x &= -\frac{\pi}{2} \\ \lim_{x\to+\infty} \mathrm{Arctan}\, x &= \frac{\pi}{2} \end{aligned}$$

et la dérivée de cette fonction est

$$\mathrm{Arctan}\,'(y) = \frac{1}{1+\tan^2(\mathrm{Arctan}\, y)} = \frac{1}{1+y^2}\,.$$

Le graphe de la fonction arc-tangente est représenté sur la figure ci-dessous :

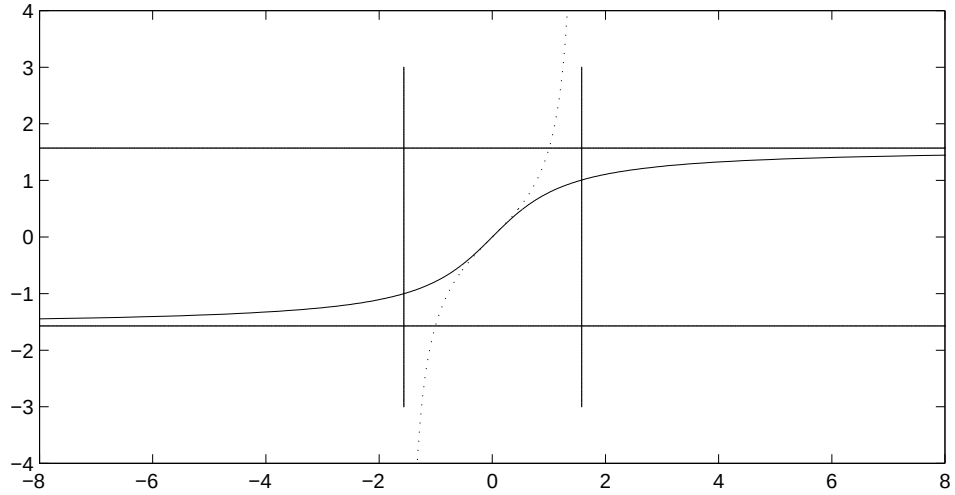

FIGURE 3.9 – les graphes de la fonction tan (en pointillé) et Arctan (en plein)

La fonction Arctan peut être utilisée aux fins de trouver un paramétrage cette fois rationnel du cercle unité du plan $\mathbb{R}^2$. En effet, si $t \in \mathbb{R}$, on a les formules de trigonométrie

$$\begin{aligned} \cos t &= \cos^2(t/2) - \sin^2(t/2) = 2\cos^2(t/2) - 1 \\ \sin t &= 2\sin(t/2)\cos(t/2)\,; \end{aligned}$$

pour $t \in]-\pi, \pi[$, on peut transformer ces formules en introduisant $u = \tan(t/2)$, c'est-à-dire en posant $t = 2\mathrm{Arctan}\, u$. Comme

$$1 + \tan^2(t/2) = \frac{\cos^2(t/2) + \sin^2(t/2)}{\cos^2(t/2)} = \frac{1}{\cos^2(t/2)}\,,$$

les relations deviennent :

$$\begin{aligned} \cos t &= \frac{2}{1+u^2} - 1 = \frac{1-u^2}{1+u^2} \\ \sin t &= \frac{2u}{1+u^2} \end{aligned}$$

et l'application

$$\Phi \ : u \in \mathbb{R} \longmapsto \Big(\frac{1-u^2}{1+u^2}\,,\, \frac{2u}{1+u^2}\Big)$$

est une application bijective entre $\mathbb{R}$ et le cercle unité privé du point $(-1, 0)$; on remarque d'ailleurs que

$$\lim_{u \to \pm\infty} \Big(\frac{1-u^2}{1+u^2}\,,\, \frac{2u}{1+u^2}\Big) = (-1, 0)\,,$$

ce qui montre que l'application Φ peut être prolongée en une application de $\mathbb{R} \cup \{-\infty, +\infty\}$ dans le cercle unité, cette fois surjective, injective sur $\mathbb{R}$ et surtout, ce qui est important, rationnelle (les applications coordonnées sont des applications rationnelles). Les deux paramètres t et u sont liés par la relation

$$t = 2\mathrm{Arctan}\,(u)\,;$$

le paramètre t est une fonction dérivable (et strictement monotone) du paramètre u et l'on a

$$\frac{dt}{du} = \frac{2}{1+u^2}\,.$$

On utilisera ce paramétrage du cercle unité pour les calculs à venir d'intégrales de fonctions rationnelles des lignes trigonométriques cos et sin.

3.6.3 Les fonctions hyperboliques et leurs inverses

De même que les fonctions trigonométriques cos et sin permettent de paramétrer le cercle unité, d'équation cartésienne $x^2 + y^2 = 1$, deux fonctions importantes en relation étroite elles aussi avec la fonction exponentielle permettent de paramétrer une branche de l'hyperbole d'équation $x^2 - y^2 = 1$, autre conique importante (on appelle conique toute section plane d'un cône, les coniques se classant en trois catégories, ellipses, paraboles et hyperboles) ; la branche ainsi paramétrée de l'hyperbole $\{x^2 - y^2 = 1\}$ est celle située dans le demi-plan droit $\{x > 0\}$.

La fonction cosh (*cosinus hyperbolique*) est la fonction paire définie sur $\mathbb{R}$ par

$$\cosh x := \frac{e^x + e^{-x}}{2} = \lim_{n\to+\infty} \sum_{k=0}^{n} \frac{x^{2k}}{(2k)!}\,.$$

C'est une fonction paire, positive ($\cosh x \geq 1$ pour tout $x \in \mathbb{R}$), dérivable sur $\mathbb{R}$ et de dérivée la fonction sinh (*sinus hyperbolique*) définie par

$$\sinh x := \frac{e^x - e^{-x}}{2}\,.$$

La fonction sinh est une fonction impaire, strictement monotone et dérivable sur $\mathbb{R}$.

Les graphes de ces deux fonctions sont représentés sur la figure ci-dessous.

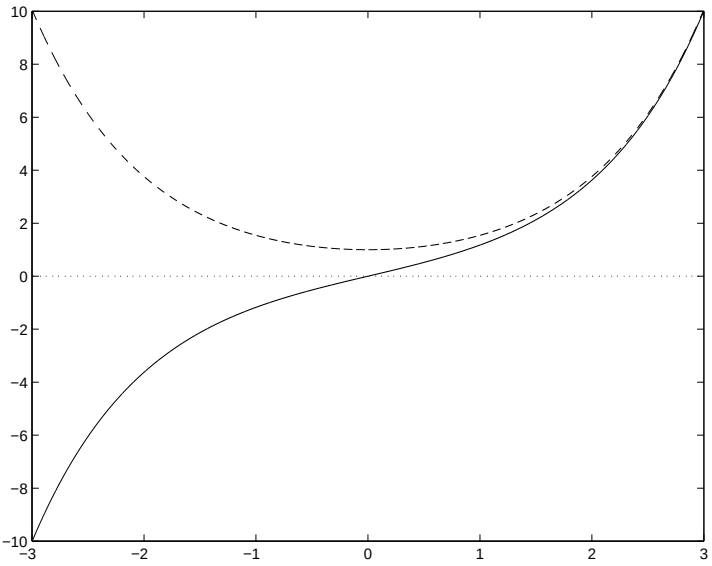

FIGURE 3.10 – les graphes de la fonction cosh (en pointillé) et sinh (en plein)

La fonction cosh tend vers $+\infty$ en $\pm\infty$ plus vite que toutes les fonctions $x \to |x|^\alpha$ avec α strictement positif ; la fonction sinh tend vers $-\infty$ en $-\infty$ et $+\infty$ en $+\infty$ (toujours en valeur absolue plus rapidement que toutes les fonctions puissances). Les deux graphes sont asymptotiques lorsque x tend vers $+\infty$.

Les deux fonctions sont liées par la relation

$$\forall x \in \mathbb{R}\,,\ \cosh^2 x - \sinh^2 x = 1\,,$$

ce qui correspond au fait que

$$t \longmapsto (\cosh t, \sinh t)$$

correspond à une application bijective entre $\mathbb{R}$ et la branche de l'hyperbole du plan d'équation cartésienne $x^2 - y^2 = 1$ qui est incluse dans le demi-plan $\{x > 0\}$ (il existe une branche symétrique par rapport à l'axe $y'Oy$ dans le demi-plan $\{x < 0\}$ qui elle est paramétrée par $t \longmapsto (-\cosh t, \sinh t)$).

La fonction sinh $: \mathbb{R} \longrightarrow \mathbb{R}$ strictement monotone et de dérivée cosh ne s'annulant pas sur $\mathbb{R}$ admet une fonction inverse argsinh ("*Argument sinus hyperbolique*") : $\mathbb{R} \longrightarrow \mathbb{R}$ définie par

$$\Big((x \in \mathbb{R}) \wedge (y = \sinh x)\Big) \Longleftrightarrow \Big((y \in \mathbb{R}) \wedge (x = \operatorname{argsinh} y)\Big)\,.$$

Cette fonction est (d'après la proposition 3.11) dérivable sur $\mathbb{R}$, de dérivée :

$$\operatorname{argsinh}'(y) = \frac{1}{\cosh(\operatorname{argsinh}(y))} = \frac{1}{\sqrt{1+y^2}}\,.$$

En fait, la fonction argsinh se calcule en remarquant que si y est un nombre réel, l'équation

$$\frac{e^x - e^{-x}}{2} = y$$

équivaut au système

$$\begin{aligned} X &= e^x \\ X^2 - 2yX - 1 &= 0\,, \end{aligned}$$

dont la solution est

$$x = \operatorname{argsinh} y = \log(y + \sqrt{1+y^2})\,.$$

On a donc la formule

$$\operatorname{argsinh} y = \log(y + \sqrt{1+y^2})$$

(on vérifiera que la dérivée est bien ce qu'elle doit être).

La fonction cosh $: [0,+\infty[\longrightarrow [1,+\infty[$ est strictement monotone et a pour dérivée sur $]0,+\infty[$ la fonction sinh ne s'annulant pas sur $]0,+\infty[$; elle admet donc une fonction inverse argcosh ("*Argument cosinus hyperbolique*" : $[1,+\infty[\longmapsto [0,+\infty[$ définie par

$$\Big((x \in [0,+\infty[) \wedge (y = \cosh x)\Big) \Longleftrightarrow \Big((y \in [1,+\infty[) \wedge (x = \operatorname{argcosh} y)\Big)\,.$$

Cette fonction est (d'après la proposition 3.11) dérivable sur $]1,+\infty[$, de dérivée :

$$\operatorname{argcosh}'(y) = \frac{1}{\sinh(\operatorname{argcosh}(y))} = \frac{1}{\sqrt{y^2-1}}\,.$$

En fait, la fonction argcosh se calcule en remarquant que si y est un nombre réel supérieur où égal à 1, l'équation

$$\frac{e^x + e^{-x}}{2} = y$$

équivaut au système

$$\begin{aligned} x &\geq 0 \\ X &= e^x \\ X^2 - 2yX + 1 &= 0\,, \end{aligned}$$

dont la solution est

$$x = \operatorname{argcosh} y = \log(y + \sqrt{y^2-1})\,.$$

On a donc la formule

$$\operatorname{argcosh} y = \log(y + \sqrt{y^2-1})$$

(on vérifiera que la dérivée est bien ce qu'elle doit être).

Les graphes de ces deux fonctions argsinh et argcosh sont représentés sur la figure ci-dessous :

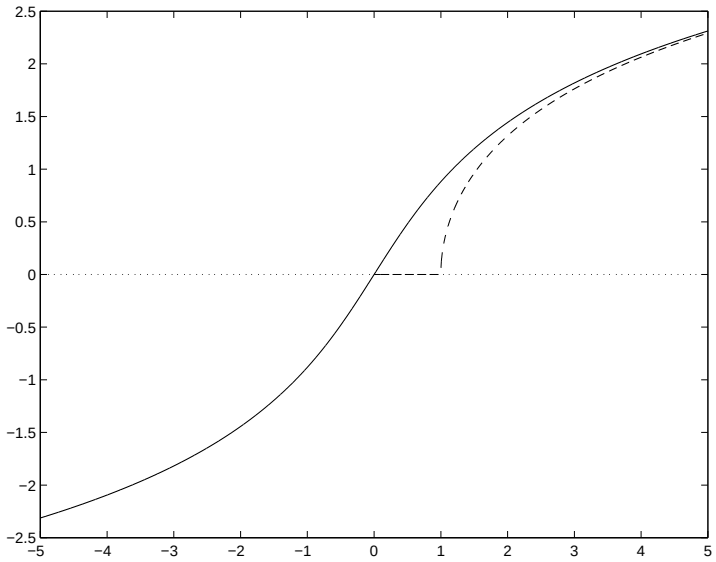

FIGURE 3.11 – les graphes de la fonction argcosh (en pointillé) et argsinh (en plein)

Autre fonction hyperbolique importante, la fonction

$$\tanh\ : x \in \mathbb{R} \longmapsto \frac{\sinh x}{\cosh x}\,;$$

cette fonction (dite *fonction tangente hyperbolique*) est une fonction dérivable sur $\mathbb{R}$, de dérivée

$$(\tanh)'(x) = \frac{\cosh^2 x - \sinh^2 x}{\cosh^2 x} = \frac{1}{\cosh^2 x} = 1 - \tanh^2(x)\,;$$

c'est donc (d'après la proposition 3.11) une fonction strictement croissante de $\mathbb{R}$ dans $]-1,1[$, admettant une fonction réciproque elle aussi strictement croissante de $]-1,1[$ dans $\mathbb{R}$. La fonction réciproque est appelée *fonction argument tangente hyperbolique* et est notée et définie par la règle

$$\Big((x \in \mathbb{R}) \wedge (y = \tanh x)\Big) \Longleftrightarrow \Big((y \in]-1,1[) \wedge (x = \operatorname{argtanh} y)\Big)\,.$$

Elle a pour dérivée

$$\operatorname{argtanh}' y = \frac{1}{1-y^2} = \frac{1}{2}\Big(\frac{1}{y+1} - \frac{1}{y-1}\Big)$$

et l'on peut montrer la formule

$$\forall y \in]-1,1[\,,\ \operatorname{argtanh} y = \log\sqrt{\frac{|y+1|}{|y-1|}}$$

en remarquant que la différence des fonctions figurant aux deux membres est dérivable et de dérivée nulle, donc que cette différence est constante (proposition 3.12) et égale à sa valeur en 0 (soit ici 0).

3.7 Fonctions de deux ou trois variables : une initiation

3.7.1 Le plan $\mathbb{R}^2$ et l'espace $\mathbb{R}^3$

Plan vectoriel, plan affine

L'ensemble produit

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} := \{(x, y)\,;\, x \in \mathbb{R}\,,\ y \in \mathbb{R}\}$$

des couples (x, y) de nombres réels hérite naturellement d'une structure de $\mathbb{R}$-*espace vectoriel*, structure que nous avons présenté dans la section 2.3.1, profitant de l'introduction aux nombres complexes. Muni de cette structure d'espace vectoriel, on dit que $\mathbb{R}^2$ est le *plan vectoriel*, les éléments étant appelés *vecteurs plans* et notés $\vec{V}$.

La loi d'addition interne entre vecteurs est, rappelons le, définie par

$$\Big((x_1, y_1)\,,\,(x_2, y_2)\Big) \longmapsto (x_1, y_1) + (x_2, y_2) := (x_1 + x_2, y_1 + y_2)\,.$$

L'action externe de $\mathbb{R}$ sur $\mathbb{R}^2$ est, elle, définie par

$$(\lambda, (x, y)) \in \mathbb{R} \times \mathbb{R}^2 \longmapsto \lambda \cdot (x, y) := (\lambda x, \lambda y)\,.$$

Ces deux opérations se plient aux règles suivantes :

- l'addition est *associative*, ce qui signifie que

$$(x_1, y_1) + \Big[(x_2, y_2) + (x_3, y_3)\Big] = \Big[(x_1, y_1) + (x_2, y_2)\Big] + (x_3, y_3)$$

  quelque soient (x_1, y_1), (x_2, y_2), (x_3, y_3) ;
- elle est *commutative*, soit

$$(x_1, y_1) + (x_2, y_2) = (x_2, y_2) + (x_1, y_1)$$

  pour tout choix de $(x_1, y_1), (x_2, y_2)$ dans $\mathbb{R}^2$;
- le vecteur nul $(0, 0)$ est élément neutre pour l'addition et tout vecteur (x, y) admet un *opposé* pour l'addition, c'est-à-dire un vecteur (x', y') (en l'occurrence ici $(-x, -y)$) tel que $(x, y) + (x', y') = (0, 0)$;
- l'opération externe est *distributive* par rapport à l'addition, ce qui signifie

$$\lambda \cdot \Big[(x_1, y_1) + (x_2, y_2)\Big] = \lambda \cdot (x_1, y_1) + \lambda \cdot (x_2, y_2)\,;$$

- enfin, on a

$$\lambda \cdot \Big[\mu \cdot (x, y)\Big] = \lambda\mu \cdot (x, y)$$

  et $1 \cdot (x, y) = (x, y)$.

Les trois premières propriétés confèrent à $\mathbb{R}^2$ muni de l'addition une structure de *groupe commutatif* ou *groupe abélien* [1]. La structure de $\mathbb{R}$-espace vectoriel combine, elle, l'opération interne d'addition des vecteurs et l'opération externe de multiplication par un scalaire, ce de manière à ce que les cinq clauses mentionnées ci-dessus soient remplies.

1. Du nom du mathématicien norvégien Nils Henrik Abel, 1802-1829, qui, en même temps qu'Evariste Galois en France, initia la théorie des groupes et mit en lumière la relation avec la résolution des équations algébriques.

Les deux vecteurs $\vec{i} := (1,0)$ et $\vec{j} := (0,1)$ constituent la *base canonique* de $\mathbb{R}^2$ et les deux nombres réels x et y constituent les *coordonnées cartésiennes*[2]. Les coordonnées cartésiennes permettent le repérage des points du plan et la mise en équations des problèmes géométriques posés dans le plan ; c'est sur ce principe que repose la *géométrie cartésienne*.

On parlera indifféremment de *vecteur de* $\mathbb{R}^2$ ou de *point de* $\mathbb{R}^2$; cependant, il y a une distinction subtile : le vecteur (x,y) doit être en fait considéré comme le *bipoint ordonné* $(0,0) \to (x,y)$ (en toute rigueur la classe de tous les bipoints ordonnés $(x_0,y_0) \to (x_0+x, y_0+y)$), tandis que le point (x,y) est, lui, simplement le couple (toujours ordonné) des nombres réels x et y. Si M est le point (x,y) et O le point $(0,0)$ et que l'on veuille différentier ces deux notions de point et de vecteur, on notera $\overrightarrow{OM}$ le vecteur (x,y) et l'on gardera la notation (x,y) pour le point M. Si $M_1 = (x_1,y_1)$ et $M_2 = (x_2,y_2)$ sont deux points du plan, on note aussi $\overrightarrow{M_1M_2}$ le vecteur $\overrightarrow{OM}$, avec $M = (x_2-x_1, y_2-y_1)$ et l'on peut donc formellement écrire la relation $M_2 = M_1 + \overrightarrow{M_1M_2}$. L'ensemble des vecteurs de $\mathbb{R}^2$ constitue le *plan vectoriel*, $\mathbb{R}$-espace vectoriel de référence de dimension 2 (toutes les bases, c'est-à-dire les familles maximales de vecteurs engendrant l'espace vectoriel, sont de cardinal 2, comme la base canonique $\{\vec{i},\vec{j}\}$), tandis que $\mathbb{R}^2$ pensé comme ensemble de points est le *plan affine*. Travailler avec le point de vue consistant à considérer les couples (x,y) de $\mathbb{R}^2$ comme des vecteurs consiste à faire de la *géométrie vectorielle*, travailler avec le point de vue consistant à les considérer comme des points consiste à faire de la *géométrie affine*. Un couple (x,y) de $\mathbb{R}^2$ se visualise donc géométriquement en le point du plan repéré par les coordonnées (cartésiennes) x et y dans le repère obtenu en choisissant arbitrairement une origine dans $\mathbb{R}^2$ et des unités de longueur sur les axes horizontaux et verticaux permettant la matérialisation des vecteurs $(1,0)$ et $(0,1)$ de la base canonique. Le choix d'unités de longueur sur les axes de coordonnées conditionne évidemment la visualisation des points.

Comme le couple (x,y) peut aussi être repéré par son *affixe*, à savoir le nombre complexe $x+iy$, le plan $\mathbb{R}^2$ s'identifie à $\mathbb{C}$ et, sous cet angle, on parle encore de *plan complexe* à propos de $\mathbb{R}^2$. Les vecteurs $\vec{V} \neq (0,0)$ du plan vectoriel $\mathbb{R}^2$ peuvent être ainsi repérés non seulement par leurs coordonnées cartésiennes, mais aussi par leurs *coordonnées polaires* :

$$(x,y) = (r\cos\theta, r\sin\theta)\,,$$

où $r = \sqrt{x^2+y^2}$ et où l'unique nombre réel $\theta \in [0,2\pi[$ est défini par les deux conditions

$$\cos\theta := \frac{x}{\sqrt{x^2+y^2}} \quad \text{et} \quad \sin\theta := \frac{y}{\sqrt{x^2+y^2}}\,.$$

Le nombre positif r est le *module* du vecteur $\vec{V}$ et l'on dit que les nombres $\theta + 2k\pi$ ($k \in \mathbb{Z}$) sont les déterminations de l'*argument* de ce même vecteur (x,y). De même, un point $M \neq (0,0)$ de $\mathbb{R}^2$ peut être repéré par sa distance $r = \|\overrightarrow{OM}\|$ à l'origine $(0,0)$ du plan et par l'angle de vecteurs $\theta := (\vec{i}, \overrightarrow{OM})$ (défini modulo 2π), ce qui donne

$$M = O + r(\cos\theta\,\vec{i} + \sin\theta\,\vec{j})\,,$$

ou encore

$$\overrightarrow{OM} = r(\cos\theta\,\vec{i} + \sin\theta\,\vec{j})\,.$$

2. Ce qualificatif est emprunté au philosophe et mathématicien français René Descartes, 1596-1650.

Espace vectoriel $\mathbb{R}^3$, espace affine $\mathbb{R}^3$

L'ensemble $\mathbb{R}^3$ (espace de la mécanique Newtonienne [3]) est défini comme l'ensemble

$$\mathbb{R}\times\mathbb{R}\times\mathbb{R} := \{(x,y,z)\,;\, x,y,z\in\mathbb{R}\}\,.$$

Les éléments sont dits *vecteurs de l'espace* et notés encore $\vec{V}$.

Addition et multiplication externe sur $\mathbb{R}^3$ sont définies de manière analogue et l'ensemble $\mathbb{R}^3$ muni de ces deux opérations hérite d'une structure de $\mathbb{R}$-espace vectoriel ; c'est *l'espace vectoriel* $\mathbb{R}^3$, dont les points constituent *l'espace affine* $\mathbb{R}^3$. La formule $M_2 = M_1 + \overrightarrow{M_1M_2}$ relie encore points et vecteurs (avec les mêmes conventions de notation que dans la sous-section précédente).

La base canonique de $\mathbb{R}^3$ est la base constituée des trois vecteurs $\vec{\imath}=(1,0,0)$, $\vec{\jmath}=(0,1,0)$ et $\vec{k}=(0,0,1)$ et les trois nombres réels x,y,z sont par définition les *coordonnées cartésiennes* du vecteur (x,y,z) dans cette base.

Un point $M=(x,y,z)\neq(0,0,0)$ de l'espace affine $\mathbb{R}^3$ peut être repéré en *coordonnées sphériques*. La première de ces coordonnées est la quantité

$$r = r(M) = \sqrt{x^2+y^2+z^2}$$

qui, on le verra dans la section suivante, s'interprète comme la distance de M à l'origine $(0,0,0)$ de l'espace affine. L'unique nombre réel $\theta\in[0,2\pi[$ défini par les deux conditions

$$\cos\theta := \frac{x}{\sqrt{x^2+y^2}} \quad\text{et}\quad \sin\theta := \frac{y}{\sqrt{x^2+y^2}}$$

s'interprète comme une détermination de la « longitude » du point M sur la sphère de centre $(0,0,0)$ et de rayon r, le méridien de référence sur cette sphère étant le grand cercle tracé sur cette sphère dans le plan vertical $\{y=0\}$. Le nombre $\varphi\in[0,\pi]$ défini par

$$\varphi := \mathrm{Arcos}\,\frac{z}{\sqrt{x^2+y^2+z^2}}$$

s'interprète, lui, comme la « colatitude » du point M sur la sphère de centre $(0,0,0)$ et de rayon r, cette colatitude étant mesurée depuis le « pôle nord » $(0,0,r(M))$. Les deux angles $\theta\in[0,2\pi[$ et $\varphi\in[0,\pi]$ permettant, avec en plus la connaissance de $r=(M)$, le repérage du point M, sont dits *angles d'Euler* et on a

$$\overrightarrow{OM} = x\,\vec{\imath}+y\,\vec{\jmath}+z\,\vec{k} = r\sin\varphi(\cos\theta\,\vec{\imath}+\sin\theta\,\vec{\jmath})+r\cos\varphi\,\vec{k}\,. \tag{3.2}$$

Un point $M=(x,y,z)\neq(0,0,0)$ de l'espace affine $\mathbb{R}^3$ peut être repéré en *coordonnées cylindriques*. La première de ces coordonnées est la quantité

$$\rho = \rho(M) = \sqrt{x^2+y^2}\,;$$

si $\rho>0$, la seconde coordonnée cylindrique de M est encore l'unique nombre réel $\theta\in[0,2\pi[$ défini par les deux conditions

$$\cos\theta := \frac{x}{\sqrt{x^2+y^2}} \quad\text{et}\quad \sin\theta := \frac{y}{\sqrt{x^2+y^2}}\,;$$

la troisième est l'altitude z de M. On a donc

$$\overrightarrow{OM} = x\,\vec{\imath}+y\,\vec{\jmath}+z\,\vec{k} = \rho(\cos\theta\,\vec{\imath}+\sin\theta\,\vec{\jmath})+z\,\vec{k}\,. \tag{3.3}$$

3. C'est à l'anglais Isaac Newton (1643-727), à la fois physicien, philosophe et mathématicien, et à ses « *Principes* » qui ont fondé le calcul vectoriel, que la terminologie fait ici référence ; l'espace $\mathbb{R}^4$ (les quatre variables étant les trois variables d'espace accompagnées de la variable temps) est, lui, le cadre de la mécanique relativiste (dont A. Einstein posa les premiers jalons théoriques à l'aube du XX-ème siècle).

3.7.2 Produit scalaire, produit vectoriel, angles

Produit scalaire, produit vectoriel dans $\mathbb{R}^2$, distance euclidienne et angles

Dans le plan $\mathbb{R}^2$, on définit une notion d'orthogonalité attachée à un produit scalaire. Le *produit scalaire* des deux vecteurs (x_1, y_1) et (x_2, y_2) est par définition le nombre réel

$$\langle (x_1, y_1)\,,\,(x_2, y_2)\rangle := x_1x_2 + y_1y_2\,.$$

Les deux vecteurs (x_1, y_1) et (x_2, y_2) sont dits *orthogonaux* si ce produit scalaire est nul.

La *distance euclidienne*[4] entre deux points $M_1 := (x_1, y_1)$ et $M_2 := (x_2, y_2)$ du plan est par définition

$$d(M_1, M_2) := \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = \sqrt{\langle \overrightarrow{M_1M_2}\,,\,\overrightarrow{M_1M_2}\rangle}\,.$$

Cette distance obéit aux quatre impératifs exigés d'une distance :
- c'est une fonction positive sur $\mathbb{R}^2 \times \mathbb{R}^2$;
- la distance entre deux points est nulle si et seulement si les deux points sont confondus (on dit que c'est une fonction *définie*) ;
- $d(M_1, M_2) = d(M_2, M_1)$ (*symétrie*) ;
- $d(M_1, M_3) \leq d(M_1, M_2) + d(M_2, M_3)$ (*inégalité triangulaire*) .

Bien sûr, il y a d'autres distances dans $\mathbb{R}^2$ que cette distance euclidienne. Ce que la distance euclidiennne a de particulier est d'être intimement liée au produit scalaire, par la formule

$$d(M_1, M_2) = \sqrt{\Big\langle (x_2 - x_1, y_2 - y_1)\,,\,(x_2 - x_1, y_2 - y_1)\Big\rangle}\,,$$

et permet donc de profiter de l'importante *formule de Pythagore*[5] : si $M_1 = (x_1, y_1)$, $M_2 = (x_2, y_2)$, et $M_3 = (x_3, y_3)$ sont trois points du plan, alors

$$d^2(M_1, M_3) = d^2(M_1, M_2) + d^2(M_2, M_3) + 2\Big\langle (x_2 - x_1, y_2 - y_1)\,,\,(x_3 - x_2, y_3 - y_2)\Big\rangle\,;$$

si en particulier les deux vecteurs $\overrightarrow{M_1M_2}$ et $\overrightarrow{M_2M_3}$ sont orthogonaux, on a

$$d^2(M_1, M_3) = d^2(M_1, M_2) + d^2(M_2, M_3)\,,$$

ce qui signifie que *le carré de l'hypoténuse d'un triangle rectangle est égal à la somme des carrés des côtés adjacents à l'angle droit.*

Si (x_1, y_1) et (x_2, y_2) sont deux vecteurs non nuls, on remarque que l'on a la formule algébrique

$$(x_1y_2 - x_2y_1)^2 + (x_1x_2 + y_1y_2)^2 = (x_1^2 + x_2^2)(y_1^2 + y_2^2)\,,$$

ou encore

$$\left(\frac{\big\langle (x_1, y_1)\,,\,(x_2, y_2)\big\rangle}{\sqrt{x_1^2 + y_1^2}\,\sqrt{x_2^2 + y_2^2}}\right)^2 + \left(\frac{x_1y_2 - x_2y_1}{\sqrt{x_1^2 + y_1^2}\,\sqrt{x_2^2 + y_2^2}}\right)^2 = 1\,.$$

4. C'est à Euclide (environ 330-260 A.C) et à ses *Eléments* que cette terminologie fait ici référence.

5. Pythagore a-t'il réellement existé comme individu ? Était-ce une secte ou un groupe de personnes ? Le mystère demeure sur le mathématicien grec installé à Crotone (Calabre, autrefois Magna Grecia) autour de 500 A.C.

Il existe donc, grâce au fait que tout point du cercle unité s'écrive de manière unique $(\cos\theta, \sin\theta)$ avec $\theta \in [0, 2\pi[$, un unique réel $\theta \in [0, 2\pi[$ tel que

$$\begin{aligned} \cos\theta &= \frac{\langle (x_1, y_1)\,,\, (x_2, y_2)\rangle}{\sqrt{x_1^2+y_1^2}\sqrt{x_2^2+y_2^2}} \\ \sin\theta &= \frac{x_1y_2 - x_2y_1}{\sqrt{x_1^2+y_1^2}\sqrt{x_2^2+y_2^2}}\,. \end{aligned}$$

Ce nombre $\theta \in [0, 2\pi[$ est, par définition, la mesure (en radians) de l'angle orienté formé par les vecteurs (x_1, y_1) et (x_2, y_2).

Si $\vec{V}_1 := (x_1, y_1)$ et $\vec{V}_2 := (x_2, y_2)$ sont deux vecteurs indépendants du plan, la quantité $x_1y_2 - x_2y_1$ est égale à la surface du parallélogramme construit à partir de (x_1, y_1) et (x_2, y_2) si le repère $\{(0,0), \vec{V}_1, \vec{V}_2\}$ est direct, ou à l'opposé de cette surface si le repère $\{(0,0), \vec{V}_1, \vec{V}_2\}$ est un repère indirect dans le plan lorsque celui ci est orienté de manière à ce que la base canonique $\{(1,0), (0,1)\}$ soit une base directe.

On peut d'ailleurs plonger $\mathbb{R}^2$ dans $\mathbb{R}^3$ en identifiant les points (x, y) et $(x, y, 0)$ et définir le vecteur $\vec{V}_1 \wedge \vec{V}_2$ (dit *produit extérieur* de $\vec{V}_1$, $\vec{V}_2$) par

$$\vec{V}_1 \wedge \vec{V}_2 := (0, 0, x_1y_2 - x_2y_1) = (x_1y_2 - x_2y_1)\,\vec{k}\,. \tag{3.4}$$

On a représenté ce vecteur (dont la longueur vaut l'aire du parallélogramme construit sur $\vec{V}_1$ et $\vec{V}_2$) sur la figure suivante :

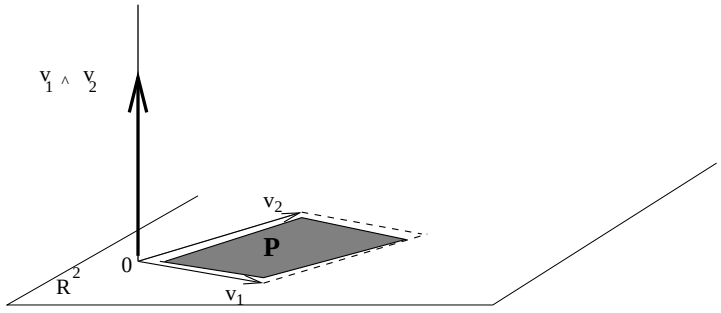

FIGURE 3.12 – Le produit extérieur et l'aire d'un parallélogramme

Un sous-ensemble U du plan est dit *ouvert* si, étant donné un point quelconque (x_0, y_0) de U, il existe un disque

$$D_{(x_0,y_0)}(\epsilon) := \{(x, y)\,;\, d((x, y), (x_0, y_0)) < \epsilon\}$$

tel que

$$(x_0, y_0) \in D_{(x_0,y_0)}(\epsilon) \subset U\,.$$

On dit aussi que U est « voisinage » de chacun de ses points.

Dire qu'une suite de points $(M_n)_{n\geq 0}$ (avec $\overrightarrow{OM_n} = (x_n, y_n)$) converge vers le point M (tel que $\overrightarrow{OM} = (x, y)$) signifie

$$\lim_{n\to+\infty} d((x_n, y_n), (x, y)) = 0\,;$$

ceci équivaut à dire les deux choses

$$\lim_{n\to+\infty} x_n = x \quad \text{et} \quad \lim_{n\to+\infty} y_n = y\,.$$

On peut ainsi définir l'adhérence $\overline{A}$ d'un sous-ensemble A du plan comme l'ensemble des points limites de suites de points de A et la notion de limite d'une fonction $f : A \to \mathbb{R}$ en un point $(a,b) \in \overline{A}$: dire que

$$\lim_{\substack{(x,y)\to(a,b)\\(x,y)\in A}} f(x,y) = l$$

(où $l \in \mathbb{R}$) équivaut à dire que pour toute suite $((x_n,y_n))_{n\geq 0}$ de points de A convergeant vers (a,b), on a

$$\lim_{n\to+\infty} f(x_n,y_n) = l\,,$$

ou encore

$$\forall\,\epsilon>0\,,\ \exists\,\eta>0\,,\ \text{tel que } \Big(d((x,y),(a,b))<\eta \text{ et } (x,y)\in A\Big) \Longrightarrow |f(x,y)-l|<\epsilon\,.$$

Produit scalaire, produit vectoriel dans $\mathbb{R}^3$, distance euclidienne et cosinus d'angle

Dans l'espace $\mathbb{R}^3$, on définit aussi une notion d'orthogonalité attachée à un produit scalaire. Le *produit scalaire* des deux vecteurs (x_1,y_1,z_1) et (x_2,y_2,z_2) est par définition le nombre réel

$$\langle (x_1,y_1,z_1)\,,\,(x_2,y_2,z_2)\rangle := x_1x_2+y_1y_2+z_1z_2\,.$$

Les deux vecteurs (x_1,y_1,z_1) et (x_2,y_2,z_2) sont dits *orthogonaux* si ce produit scalaire est nul.

La *distance euclidienne* entre deux points $M_1 := (x_1,y_1,z_1)$ et $M_2 := (x_2,y_2,z_2)$ de l'espace est par définition

$$d(M_1,M_2) := \sqrt{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2} = \sqrt{\langle \overrightarrow{M_1M_2}\,,\,\overrightarrow{M_1M_2}\rangle}\,.$$

Cette distance obéit (comme dans le cas du plan $\mathbb{R}^2$) aux quatre impératifs exigés d'une distance. Elle est encore intimement liée au produit scalaire, par la formule

$$d(M_1,M_2) = \sqrt{\Big\langle (x_2-x_1,y_2-y_1,,z_2-z_1)\,,\,(x_2-x_1,y_2-y_1,,z_2-z_1)\Big\rangle}\,,$$

et permet donc encore de profiter de l'importante *formule de Pythagore* : si $M_1=(x_1,y_1,z_1)$, $M_2=(x_2,y_2,z_2)$, et $M_3=(x_3,y_3,z_3)$ sont trois points de l'espace, alors

$$\begin{aligned} d^2(M_1,M_3) &= d^2(M_1,M_2)+d^2(M_2,M_3)\\ &\quad +2\Big\langle (x_2-x_1,y_2-y_1,z_2-z_1)\,,\,(x_3-x_2,y_3-y_2,z_3-z_2)\Big\rangle; \end{aligned}$$

si en particulier les deux vecteurs $\overrightarrow{M_1M_2}$ et $\overrightarrow{M_2M_3}$ sont orthogonaux, on a

$$d^2(M_1,M_3) = d^2(M_1,M_2)+d^2(M_2,M_3)\,,$$

ce qui signifie encore que *le carré de l'hypoténuse d'un triangle rectangle (de l'espace cette fois) est égal à la somme des carrés des côtés adjacents à l'angle droit.*

Si $\vec{V_1} = (x_1, y_1, z_1)$ et $\vec{V_2} = (x_2, y_2, z_2)$ sont deux vecteurs non nuls de $\mathbb{R}^3$, le nombre

$$\frac{\langle \vec{V_1}\,,\,\vec{V_2}\rangle}{\sqrt{\langle \vec{V_1}\,,\,\vec{V_1}\rangle}\,\sqrt{\langle \vec{V_2}\,,\,\vec{V_2}\rangle}}$$

est un nombre appartenant à $[-1, 1]$, que l'on peut donc écrire de manière unique $\cos\theta$ avec $\theta \in [0, \pi]$; on note donc

$$\cos(\vec{V_1}, \vec{V_2}) := \frac{\langle \vec{V_1}\,,\,\vec{V_2}\rangle}{\sqrt{\langle \vec{V_1}\,,\,\vec{V_1}\rangle}\,\sqrt{\langle \vec{V_2}\,,\,\vec{V_2}\rangle}}\,;$$

si ce nombre est < 0 (c'est-à-dire si $\theta \in]\pi/2, \pi]$), on dit que l'angle des vecteurs $\vec{V_1}$ et $\vec{V_2}$ est *obtus* ; si ce nombre est > 0 (c'est-à-dire si $\theta \in [0, \pi/2[$), on dit que l'angle des vecteurs $\vec{V_1}$ et $\vec{V_2}$ est *aigu* ; si ce nombre est nul, les vecteurs $\vec{V_1}$ et $\vec{V_2}$ sont orthogonaux.

Le produit extérieur $\vec{V_1} \wedge \vec{V_2}$ de deux vecteurs $\vec{V_1} = (x_1, y_1, z_1)$ et $\vec{V_2} = (x_2, y_2, z_2)$ est défini par les règles « tournantes »

$$\begin{array}{rcl} \vec{i} \wedge \vec{j} & = & \vec{k} = -\vec{j} \wedge \vec{i} \\ \vec{j} \wedge \vec{k} & = & \vec{i} = -\vec{k} \wedge \vec{j} \\ \vec{k} \wedge \vec{i} & = & \vec{j} = -\vec{i} \wedge \vec{k} \end{array}$$

et par les deux clauses de linéarité

$$\begin{array}{rcl} (\lambda\vec{V_1} + \mu\vec{W_1}) \wedge \vec{V_2} & = & \lambda\vec{V_1} \wedge \vec{V_2} + \mu\vec{W_1} \wedge \vec{V_2} \\ \vec{V_1} \wedge (\lambda\vec{V_2} + \mu\vec{W_2}) & = & \lambda\vec{V_1} \wedge \vec{V_2} + \mu\vec{V_1} \wedge \vec{W_2} \end{array}$$

pour tout choix de vecteurs $\vec{V_1}, \vec{W_1}, \vec{V_2}, \vec{W_2}$ et de scalaires λ, μ. Cela donne donc

$$\vec{V_1} \wedge \vec{V_2} := (y_1 z_2 - y_2 z_1)\,\vec{i} + (z_1 x_2 - x_1 z_2)\,\vec{j} + (x_1 y_2 - x_2 y_1)\,\vec{k} \tag{3.5}$$

et cette définition est cohérente avec la définition du produit extérieur de deux vecteurs (x_1, y_1) et (x_2, y_2) du plan proposée en (3.4) si l'on identifie ces vecteurs plans respectivement avec $(x_1, y_1, 0)$ et $(x_2, y_2, 0)$.

Le nombre

$$\|\vec{V_1} \wedge \vec{V_2}\| := \sqrt{(y_1 z_2 - y_2 z_1)^2 + (z_1 x_2 - x_1 z_2)^2 + (x_1 y_2 - x_2 y_1)^2}$$

mesure exactement la surface du parallélogramme plan construit sur les vecteurs $\vec{V_1}, \vec{V_2}$ (voir la figure 3.13). Le produit extérieur de $\vec{V_1}$ et $\vec{V_2}$ est nul si l'un de ces deux vecteurs est un multiple de l'autre (on dit que les vecteurs sont *liés*) ; si ce n'est pas le cas, le vecteur $\vec{V_1} \wedge \vec{V_2}$ est orthogonal à la fois à $\vec{V_1}$ et $\vec{V_2}$ et pointe dans $\mathbb{R}^3$ de manière à ce que le repère $(\vec{V_1}, \vec{V_2}, \vec{V_1} \wedge \vec{V_2})$ soit direct [6] ; si $\vec{V_3}$ est un troisième vecteur de l'espace $\mathbb{R}^3$, le nombre

$$\left|\left\langle \vec{V_1} \wedge \vec{V_2}\,,\,\vec{V_3}\right\rangle\right|$$

mesure exactement le volume du parallélépipède construit à partir des trois vecteurs $\vec{V_1}, \vec{V_2}, \vec{V_3}$ (ce volume étant nul dès que ce parallélépipède se trouve « aplati ») ; le nombre

$$\left\langle \vec{V_1} \wedge \vec{V_2}\,,\,\vec{V_3}\right\rangle$$

6. Se référer à la classique *règle des trois doigts* ou du *bonhomme d'Ampère* que vous connaissez et appliquez en physique.

et, lui, appelé *produit mixte* des trois vecteurs $\vec{V}_1, \vec{V}_2, \vec{V}_3$ et on a

$$\left\langle \vec{V}_1 \wedge \vec{V}_2 \, , \, \vec{V}_3 \right\rangle = \left\langle \vec{V}_2 \wedge \vec{V}_3 \, , \, \vec{V}_1 \right\rangle = \left\langle \vec{V}_3 \wedge \vec{V}_1 \, , \, \vec{V}_2 \right\rangle$$

(remarquez encore le caractère « tournant » de ces formules).

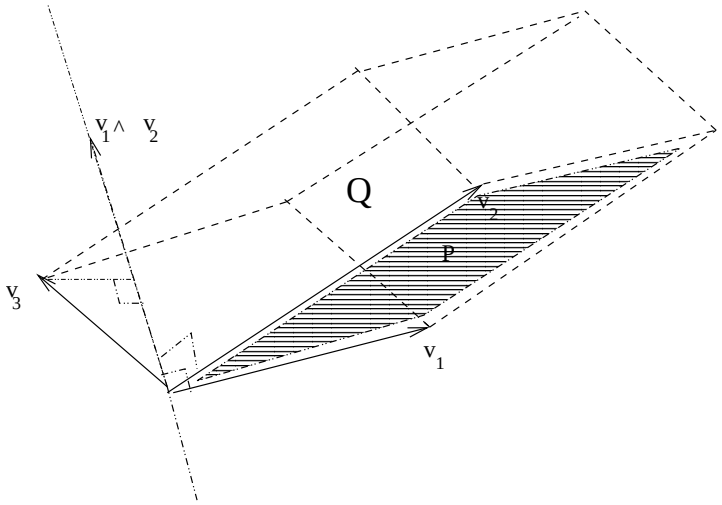

FIGURE 3.13 – Le produit extérieur, l'aire d'un parallélogramme, le volume d'un parallélépipède

Le fait d'avoir muni $\mathbb{R}^3$ d'une distance permet de définir la notion d'ouvert U de $\mathbb{R}^3$ ainsi que la notion de limite d'une fonction réelle en un point adhérent à son domaine de définition. Un sous-ensemble U de l'espace est dit *ouvert* si, étant donné un point quelconque (x_0, y_0, z_0) de U, il existe une sphère pleine ouverte

$$S_{(x_0,y_0,z_0)}(\epsilon) := \{(x, y, z) \, ; \, d((x, y, z), (x_0, y_0, z_0)) < \epsilon\}$$

telle que

$$(x_0, y_0, z_0) \in S_{(x_0,y_0,z_0)}(\epsilon) \subset U \, .$$

On dit aussi que U est « voisinage » de chacun de ses points.

Dire qu'une suite de points $(M_n)_{n\geq 0}$ (avec $\overrightarrow{OM_n} = (x_n, y_n, z_n)$) converge vers le point M (tel que $\overrightarrow{OM} = (x, y, z)$) signifie

$$\lim_{n\to+\infty} d((x_n, y_n, z_n), (x, y, z)) = 0 \, ;$$

ceci équivaut à dire les trois choses

$$\lim_{n\to+\infty} x_n = x \quad \text{et} \quad \lim_{n\to+\infty} y_n = y \quad \lim_{n\to+\infty} z_n = z \, .$$

On peut ainsi définir l'adhérence $\overline{A}$ d'un sous-ensemble A de l'espace comme l'ensemble des points limites de suites de points de A et la notion de limite d'une fonction $f : A \to \mathbb{R}$ en un point $(a, b, c) \in \overline{A}$: dire que

$$\lim_{\substack{(x,y,z)\to(a,b,c) \\ (x,y,z)\in A}} f(x, y, z) = l$$

(où $l \in \mathbb{R}$) équivaut à dire que pour toute suite $((x_n, y_n, z_n))_{n\geq 0}$ de points de A convergeant vers (a, b, c), on a

$$\lim_{n\to+\infty} f(x_n, y_n, z_n) = l$$

ou encore

$$\forall\, \epsilon > 0\,,\ \exists\, \eta > 0\,,\ \text{tel que } \Big(d((x, y, z), (a, b, c)) < \eta \text{ et } (x, y, z) \in A\Big) \Longrightarrow |f(x, y, z) - l| < \epsilon\,.$$

3.7.3 Continuité et différentiabilité en un point d'une fonction de deux ou trois variables ; graphe et gradient

Le cas des fonctions de deux variables

Si f est une fonction définie dans un sous-ensemble A du plan $\mathbb{R}^2$ et à valeurs réelles, on dit que f est continue en un point (x_0, y_0) de A si et seulement si

$$\lim_{\substack{(x,y)\to(x_0,y_0)\\(x,y)\in A}} f(x, y) = f(x_0, y_0)\,.$$

Si maintenant f est définie dans un voisinage ouvert U d'un point (x_0, y_0) du plan (et est à valeurs réelles), on dit que f est *différentiable* en (x_0, y_0) s'il existe une application $\mathbb{R}$-linéaire

$$df_{(x_0,y_0)} \ : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

telle que

$$\lim_{\substack{(h,k)\to(0,0)\\(h,k)\neq(0,0)}} \frac{f(x_0 + h, y_0 + k) - f(x_0, y_0) - df_{(x_0,y_0)}(h, k)}{\sqrt{h^2 + k^2}} = 0\,,$$

autrement dit, on peut écrire, si la « perturbation » (h, k) est « petite »,

$$f(x_0 + h, y_0 + k) \simeq f(x_0, y_0) + df_{(x_0,y_0)}(h, k)\,,$$

l'erreur dans cette approximation étant (en valeur absolue) négligeable devant $\sqrt{h^2 + k^2}$. On peut alors écrire

$$df_{(x_0,y_0)}(h, k) = a_{(x_0,y_0)}h + b_{(x_0,y_0)}k\,,$$

où les deux nombres $a_{(x_0,y_0)}$ et $b_{(x_0,y_0)}$ sont définis par

$$\begin{aligned} a_{(x_0,y_0)} &= \frac{\partial f}{\partial x}(x_0, y_0) := \lim_{\substack{h\to 0\\h\in\mathbb{R}^*}} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h} \\ b_{(x_0,y_0)} &= \frac{\partial f}{\partial y}(x_0, y_0) := \lim_{\substack{k\to 0\\k\in\mathbb{R}^*}} \frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k}\,. \end{aligned}$$

Ces deux nombres sont appelés respectivement *dérivées partielles* de f par rapport à x et y au point (x_0, y_0) et il n'y a rien de plus facile pour les calculer que, dans l'expression de $f(x, y)$, de « colorier » en rouge une des deux variables $\mathbf{x}$ ou $\mathbf{y}$, de considérer ensuite

$$\begin{aligned} \frac{d}{dx}[f(x, \mathbf{y})] &= \frac{\partial f}{\partial x}(x, y) \\ \frac{d}{dy}[f(\mathbf{x}, y)] &= \frac{\partial f}{\partial y}(x, y)\,, \end{aligned}$$

les dérivations se faisant par rapport aux variables « non coloriées » (« colorier » une variable revient donc en quelque sorte à la « geler »).

Exemple. On a

$$\begin{aligned}\frac{\partial}{\partial x}\Big[\frac{1}{1+x^2+\mathbf{y}^4}\Big] &= \frac{-2x}{(1+x^2+\mathbf{y}^4)^2}\\ \frac{\partial}{\partial y}\Big[\frac{1}{1+\mathbf{x}^2+y^4}\Big] &= \frac{-4y^3}{(1+\mathbf{x}^2+y^4)^2}\,.\end{aligned}$$

Remarque. Une application différentiable en un point (x_0, y_0) d'un ouvert U de $\mathbb{R}^2$ est continue en ce point, mais l'assertion réciproque est fausse (puisqu'elle est fausse, on la vu, dans le cas des fonctions d'une variable).

Important ! *On admettra ici que si f est définie dans un ouvert U de $\mathbb{R}^2$, si les dérivées partielles*

$$\frac{\partial f}{\partial x}(x,y) \quad \text{et} \quad \frac{\partial f}{\partial y}(x,y)$$

existent (au sens sont calculables comme ci-dessus en tout point (x,y) de U) et que les fonctions

$$\begin{aligned}(x,y)\in U \quad &\longrightarrow \quad \frac{\partial f}{\partial x}(x,y)\\ (x,y)\in U \quad &\longrightarrow \quad \frac{\partial f}{\partial y}(x,y)\end{aligned}$$

sont continues sur U, alors f est différentiable en tout point de U et l'on a alors, bien sûr, pour tout $(x,y)\in U$,

$$\forall\,(h,k)\in\mathbb{R}^2\,,\ df_{(x,y)}(h,k) = \frac{\partial f}{\partial x}(x,y)\,h + \frac{\partial f}{\partial x}(x,y)\,k\,.$$

Le *graphe* d'une fonction $f\ : A\subset\mathbb{R}^2 \longrightarrow \mathbb{R}$ est par définition le sous-ensemble $\Gamma(f)$ de $\mathbb{R}^3$ donné par

$$\Gamma(f) := \{(x,y,z)\in\mathbb{R}^3\,;\,(x,y)\in A \text{ et } z = f(x,y)\}\,.$$

Par exemple, sur la figure 3.14 ci-dessous, on a figuré le graphe (représenté en trois dimensions) de l'application

$$(x,y)\in[-10,10]\times[-10,10]\longmapsto 3x^2-2y^2\,.$$

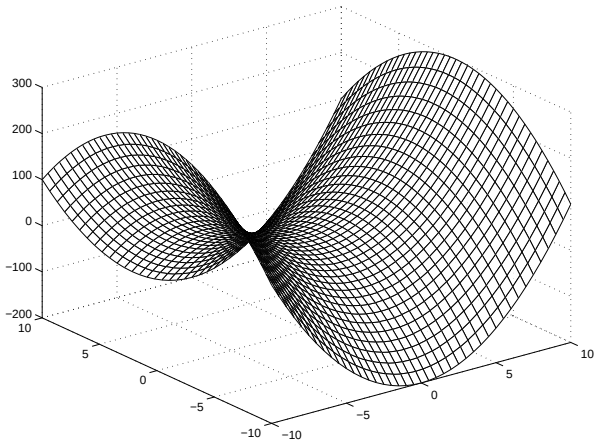

FIGURE 3.14 – Le graphe de $(x,y)\longmapsto 3x^2-2y^2$

Si f est une fonction définie et différentiable en tout point d'un ouvert U de $\mathbb{R}^2$, le *gradient* de f est par définition l'application

$$\overrightarrow{\nabla f} \ : (x,y) \in U \longmapsto \Big(\frac{\partial f}{\partial x}(x,y)\,,\ \frac{\partial f}{\partial y}(x,y) \Big) \in \mathbb{R}^2\,.$$

Le gradient de f est donc une application de U dans $\mathbb{R}^2$ qui à tout point de U associe donc un vecteur de $\mathbb{R}^2$; une telle application est dite *champ de vecteurs* sur U. En un point (x,y) de U où $\overrightarrow{\nabla f}(x,y) \neq (0,0)$, la « ligne de plus grande pente » sur le graphe de f au départ du point $(x_0,y_0,f(x_0,y_0))$ est la courbe ainsi paramétrée par le paramètre t (voisin de 0) :

$$\begin{aligned} (x,y) &= (x_0,y_0) + t\,\overrightarrow{\nabla f}(x_0,y_0) \\ z &= f\big((x_0,y_0) + t\,\overrightarrow{\nabla f}(x_0,y_0)\big)\,; \end{aligned}$$

(on remarque d'ailleurs que $t \longmapsto z(t)$ est une fonction croissante de t au voisinage de $t=0$) ; cela résulte de ce que, pour (h,k) voisin de $(0,0)$,

$$f(x_0+h,y_0+k) \simeq f(x_0,y_0) + \big\langle \overrightarrow{\nabla f}(x_0,y_0)\,,\,(h,k) \big\rangle\,,$$

tandis que le maximum des nombres

$$\frac{\big|\big\langle \overrightarrow{\nabla f}(x_0,y_0)\,,\,(h,k)\big\rangle\big|}{\sqrt{h^2+k^2}}\,, \quad (h,k) \neq (0,0)$$

est réalisé précisément lorsque

$$(h,k) = \lambda \overrightarrow{\nabla f}\,,\ \lambda \in \mathbb{R}^*\,.$$

Ainsi, c'est en se déplaçant sur le graphe de f suivant (en (x,y)) la direction que nous indique à chaque instant le gradient de f (*resp.* son opposé) que l'on peut espérer se rapprocher au plus vite des points où f présente un maximum (*resp.* un minimum) local. C'est là l'origine d'une très importante méthode en mathématiques appliquées, la *méthode du gradient*.

Le cas des fonctions de trois variables

Si f est une fonction définie dans un sous-ensemble A de l'espace $\mathbb{R}^3$ et à valeurs réelles, on dit que f est continue en un point (x_0,y_0,z_0) de A si et seulement si

$$\lim_{\substack{(x,y,z)\to(x_0,y_0,z_0)\\(x,y,z)\in A}} f(x,y,z) = f(x_0,y_0,z_0)\,.$$

Si maintenant f est définie dans un voisinage ouvert U d'un point (x_0,y_0,z_0) du plan (et est à valeurs réelles), on dit que f est *différentiable* en (x_0,y_0,z_0) s'il existe une application $\mathbb{R}$-linéaire

$$df_{(x_0,y_0,z_0)} \ : \mathbb{R}^3 \longrightarrow \mathbb{R}$$

telle que

$$\lim_{\substack{(h,k,l)\to(0,0)\\(h,k,l)\neq(0,0)}} \frac{f(x_0+h,y_0+k,z_0+l) - f(x_0,y_0,z_0) - df_{(x_0,y_0,z_0)}(h,k,l)}{\sqrt{h^2+k^2+l^2}} = 0\,,$$

autrement dit, on peut écrire, si la « perturbation » (h,k,l) est « petite »,

$$f(x_0+h,y_0+k,z_0+l) \simeq f(x_0,y_0,z_0) + df_{(x_0,y_0,z_0)}(h,k,l)\,,$$

l'erreur dans cette approximation étant (en valeur absolue) négligeable devant $\sqrt{h^2+k^2+l^2}$. On peut alors écrire

$$df_{(x_0,y_0,z_0)}(h,k,l) = a_{(x_0,y_0,z_0)}h + b_{(x_0,y_0,z_0)}k + c_{(x_0,y_0,z_0)}\,l\,,$$

où les trois nombres $a_{(x_0,y_0,z_0)}$, $b_{(x_0,y_0,z_0)}$ et $c_{(x_0,y_0,z_0)}$ sont définis par

$$\begin{aligned}
a_{(x_0,y_0,z_0)} &= \frac{\partial f}{\partial x}(x_0,y_0,z_0) := \lim_{\substack{h\to 0\\ h\in\mathbb{R}^*}} \frac{f(x_0+h,y_0,z_0)-f(x_0,y_0,z_0)}{h}\\
b_{(x_0,y_0,z_0)} &= \frac{\partial f}{\partial y}(x_0,y_0,z_0) := \lim_{\substack{k\to 0\\ k\in\mathbb{R}^*}} \frac{f(x_0,y_0+k,z_0)-f(x_0,y_0,z_0)}{k}\\
c_{(x_0,y_0,z_0)} &= \frac{\partial f}{\partial z}(x_0,y_0,z_0) := \lim_{\substack{l\to 0\\ l\in\mathbb{R}^*}} \frac{f(x_0,y_0,z_0+l)-f(x_0,y_0,z_0)}{l}\,.
\end{aligned}$$

Ces trois nombres sont appelés respectivement *dérivées partielles* de f par rapport à x, y, z au point (x_0,y_0,z_0) et il n'y a rien de plus facile pour les calculer que, dans l'expression de $f(x,y,z)$, de « colorier » en rouge deux des trois variables $\{\mathbf{y},\mathbf{z}\}$, $\{\mathbf{x},\mathbf{z}\}$, $\{\mathbf{x},\mathbf{y}\}$, de considérer ensuite

$$\begin{aligned}
\frac{d}{dx}[f(x,\mathbf{y},\mathbf{z})] &= \frac{\partial f}{\partial x}(x,y,z)\\
\frac{d}{dy}[f(\mathbf{x},y,\mathbf{z})] &= \frac{\partial f}{\partial y}(x,y,z)\\
\frac{d}{dz}[f(\mathbf{x},\mathbf{y},z)] &= \frac{\partial f}{\partial z}(x,y,z)\,,
\end{aligned}$$

les dérivations se faisant par rapport aux variables « non coloriées » (« colorier » une variable revient donc en quelque sorte à la « geler »).

Exemple. On a

$$\begin{aligned}
\frac{\partial}{\partial x}\Big[\frac{1}{1+x^2+\mathbf{y}^4+\mathbf{z}^6}\Big] &= \frac{-2x}{(1+x^2+\mathbf{y}^4+\mathbf{z}^6)^2}\\
\frac{\partial}{\partial y}\Big[\frac{1}{1+\mathbf{x}^2+y^4+\mathbf{z}^6}\Big] &= \frac{-4y^3}{(1+\mathbf{x}^2+y^4+\mathbf{z}^6)^2}\\
\frac{\partial}{\partial z}\Big[\frac{1}{1+\mathbf{x}^2+\mathbf{y}^4+z^6}\Big] &= \frac{-6z^5}{(1+\mathbf{x}^2+\mathbf{y}^4+z^6)^2}
\end{aligned}$$

Remarque. Une application différentiable en un point (x_0,y_0,z_0) d'un ouvert U de $\mathbb{R}^3$ est continue en ce point, mais l'assertion réciproque est fausse (puisqu'elle est fausse, on la vu, dans le cas des fonctions d'une ou de deux variables).

Important ! *On admettra ici que si f est définie dans un ouvert U de $\mathbb{R}^3$, si les dérivées partielles*

$$\frac{\partial f}{\partial x}(x,y,z)\,,\quad \frac{\partial f}{\partial y}(x,y,z)\quad\text{et}\quad \frac{\partial f}{\partial z}(x,y,z)$$

existent (au sens sont calculables comme ci-dessus en tout point (x,y,z) de U) et que les fonctions

$$\begin{aligned}
(x,y,z)\in U &\longrightarrow \frac{\partial f}{\partial x}(x,y,z)\\
(x,y,z)\in U &\longrightarrow \frac{\partial f}{\partial y}(x,y,z)\\
(x,y,z)\in U &\longrightarrow \frac{\partial f}{\partial z}(x,y,z)
\end{aligned}$$

sont continues sur U, alors f est différentiable en tout point de U et l'on a alors, bien sûr, pour tout $(x,y,z) \in U$,

$$\forall\,(h,k,l) \in \mathbb{R}^3\,,\ df_{(x,y,z)}(h,k,l) = \frac{\partial f}{\partial x}(x,y,z)\,h + \frac{\partial f}{\partial x}(x,y,z)\,k + \frac{\partial f}{\partial x}(x,y,z)\,l\,.$$

Le *graphe* d'une fonction $f\ : A \subset \mathbb{R}^3 \longrightarrow \mathbb{R}$ est par définition le sous-ensemble $\Gamma(f)$ de $\mathbb{R}^4$ donné par

$$\Gamma(f) := \{(x,y,z,w) \in \mathbb{R}^4\,;\,(x,y,z) \in A \text{ et } w = f(x,y,z)\}\,.$$

La représentation graphique de $\Gamma(f)$ n'est cette fois plus possible comme elle l'était pour une fonction de deux variables (on ne peut visualiser les objets en quatre dimensions).

Si f est une fonction définie et différentiable en tout point d'un ouvert U de $\mathbb{R}^3$, le *gradient* de f est par définition l'application

$$\overrightarrow{\nabla f}\ : (x,y,z) \in U \longmapsto \Big(\frac{\partial f}{\partial x}(x,y,z)\,,\,\frac{\partial f}{\partial y}(x,y,z)\,,\,\frac{\partial f}{\partial z}(x,y,z)\Big) \in \mathbb{R}^3\,.$$

Le gradient de f est donc une application de U dans $\mathbb{R}^3$ qui à tout point de U associe donc un vecteur de $\mathbb{R}^3$; une telle application est dite encore *champ de vecteurs* sur U. En un point (x,y,z) de U où $\overrightarrow{\nabla f}(x,y,z) \neq (0,0,0)$, la « ligne de plus grande pente » sur le graphe de f au départ du point $(x_0,y_0,z_0,f(x_0,y_0,z_0))$ est la courbe ainsi paramétrée par le paramètre t (voisin de 0) :

$$\begin{aligned}(x,y,z) &= (x_0,y_0,z_0) + t\,\overrightarrow{\nabla f}(x_0,y_0,z_0)\\ w &= f\big((x_0,y_0,z_0) + t\,\overrightarrow{\nabla f}(x_0,y_0,z_0)\big)\,;\end{aligned}$$

(on remarque d'ailleurs que $t \longmapsto w(t)$ est une fonction croissante de t au voisinage de $t=0$); cela résulte de ce que, pour (h,k,l) voisin de $(0,0,0)$,

$$f(x_0+h,y_0+k,z_0+l) \simeq f(x_0,y_0,z_0) + \big\langle \overrightarrow{\nabla f}(x_0,y_0,z_0)\,,\,(h,k,l)\big\rangle\,,$$

tandis que le maximum des nombres

$$\frac{\big|\big\langle \overrightarrow{\nabla f}(x_0,y_0,z_0)\,,\,(h,k,l)\big\rangle\big|}{\sqrt{h^2+k^2+l^2}}\,,\quad (h,k,l) \neq (0,0,0)$$

est réalisé précisément lorsque

$$(h,k,l) = \lambda \overrightarrow{\nabla f}(x_0,y_0,z_0)\,,\ \lambda \in \mathbb{R}^*\,.$$

Ainsi, c'est en se déplaçant sur le graphe de f suivant (en (x,y,z)) la direction que nous indique à chaque instant le gradient de f (*resp.* son opposé) que l'on peut espérer se rapprocher au plus vite des points où f présente un maximum (*resp.* un minimum) local. On retrouve ici encore, pour les fonctions de trois variables cette fois, la *méthode du gradient*.

3.7.4 La « *chain rule* » du calcul différentiel : quelques exemples

La règle de Leibniz (ou aussi « *chain rule* » dans la terminologie anglo-saxonne) présentée dans la section 3.5 se transpose au cadre des fonctions composées de plusieurs variables. Nous ne mentionnerons ici que quatre exemples importants.

Exemple 1

Supposons que

$$\gamma \ : I \subset \mathbb{R} \longmapsto \gamma(t) = (x(t), y(t))$$

soit une application différentiable d'un intervalle I de $\mathbb{R}$, à valeurs dans $\mathbb{R}^2$, ce qui signifie que les fonctions $t \longmapsto x(t)$ et $t \longmapsto y(t)$ sont dérivables en tout point de I (une telle application est appelée un arc paramétré plan différentiable). Supposons que F soit une application définie et différentiable sur un ouvert U de $\mathbb{R}^2$, avec $\gamma(I) \subset U$.

La fonction $F \circ \gamma$ est alors dérivable en tout point de I, de dérivée

$$(F \circ \gamma)'(t) = \frac{\partial F}{\partial x}(x(t), y(t))\, x'(t) + \frac{\partial F}{\partial y}(x(t), y(t))\, y'(t) = \left\langle \overrightarrow{\nabla F}(x(t), y(t))\, ,\, (x'(t), y'(t)) \right\rangle .$$

Pour voir cela, il suffit de remarquer que

$$\begin{aligned} F(\gamma(t+\tau)) &= F(x(t+\tau), y(t+\tau)) = F(x(t) + \tau x'(t) + o(\tau), y(t) + \tau y'(t) + o(\tau)) \\ &= F(x(t), y(t)) + \tau \left\langle \overrightarrow{\nabla F}(x(t), y(t))\, ,\, (x'(t), y'(t)) \right\rangle + o(\tau) . \end{aligned}$$

Exemple 2

Supposons que

$$\gamma \ : I \subset \mathbb{R} \longmapsto \gamma(t) = (x(t), y(t), z(t))$$

soit une application différentiable d'un intervalle I de $\mathbb{R}$, à valeurs dans $\mathbb{R}^3$, ce qui signifie que les fonctions $t \longmapsto x(t)$, $t \longmapsto y(t)$ et $t \longmapsto z(t)$ sont dérivables en tout point de I (une telle application est appelée un arc paramétré gauche différentiable). Supposons que F soit une application définie et différentiable sur un ouvert U de $\mathbb{R}^3$, avec $\gamma(I) \subset U$.

La fonction $F \circ \gamma$ est alors dérivable en tout point de I, de dérivée

$$\begin{aligned} (F \circ \gamma)'(t) &= \frac{\partial F}{\partial x}(x(t), y(t), z(t))\, x'(t) + \frac{\partial F}{\partial y}(x(t), y(t), z(t))\, y'(t) + \frac{\partial F}{\partial z}(x(t), y(t), z(t))\, z'(t) \\ &= \left\langle \overrightarrow{\nabla F}(x(t), y(t), z(t))\, ,\, (x'(t), y'(t), z'(t)) \right\rangle . \end{aligned}$$

Pour voir cela, il suffit de remarquer que

$$\begin{aligned} F(\gamma(t+\tau)) &= F(x(t+\tau), y(t+\tau), z(t+\tau)) \\ &= F(x(t) + \tau x'(t) + o(\tau), y(t) + \tau y'(t) + o(\tau), z(t) + \tau z'(t) + o(\tau)) \\ &= F(x(t), y(t), z(t)) + \tau \left\langle \overrightarrow{\nabla F}(x(t), y(t), z(t))\, ,\, (x'(t), y'(t), z'(t)) \right\rangle + o(\tau) . \end{aligned}$$

Exemple 3

Soit $F \ : (u, v) \longmapsto (x(u, v), y(u, v))$ une application d'un ouvert U de $\mathbb{R}^2$, à valeurs dans $\mathbb{R}^2$, telle que les applications coordonnées x et y soient toutes les deux différentiables en tout point de U. Soit V un ouvert de $\mathbb{R}^2$ contenant $F(U)$ et G une application différentiable de V dans $\mathbb{R}$.

L'application $G \circ F \ : U \longrightarrow \mathbb{R}$ est différentiable en tout point (u, v) de U et on a

$$\begin{aligned} \frac{\partial [G \circ F]}{\partial u}(u, v) &= \frac{\partial G}{\partial x}(F(u, v)) \frac{\partial x}{\partial u} + \frac{\partial G}{\partial y}(F(u, v)) \frac{\partial y}{\partial u} \\ \frac{\partial [G \circ F]}{\partial v}(u, v) &= \frac{\partial G}{\partial x}(F(u, v)) \frac{\partial x}{\partial v} + \frac{\partial G}{\partial y}(F(u, v)) \frac{\partial y}{\partial v} . \end{aligned}$$

Pour voir ce résultat, il suffit d'écrire explicitement ce que signifie la différentiabilité de G en $F(u,v)$, soit

$$\begin{aligned} G(x(u,v)+H, y(u,v)+K) &= G(F(u,v)+(H,K)) \\ &= G(F(u,v)) + dG_{F(u,v)}(H,K) + o(\sqrt{H^2+K^2}) \end{aligned} \tag{3.6}$$

et de remarquer que

$$\begin{aligned} x(u+h,v+k) &= x(u,v) + dx_{(u,v)}(h,k) + o(\sqrt{h^2+k^2}) = x(u,v)+H \\ y(u+h,v+k) &= y(u,v) + dy_{(u,v)}(h,k) + o(\sqrt{h^2+k^2}) = y(u,v)+K \end{aligned}$$

avant de substituer dans (3.6) et de conclure à la différentiabilité de $G \circ F$ en (t,s).

Exemple (passage aux coordonnées polaires dans le plan). Si F est l'application

$$(r,\theta) \in]r_1,r_2[\times]\theta_1,\theta_2[\longmapsto (r\cos\theta\,,\, r\sin\theta)\,,$$

et si G est une application différentiable dans le secteur ouvert $F(]r_1,r_2[\times]\theta_1,\theta_2[)$ du plan et à valeurs dans $\mathbb{R}$, la fonction

$$g\ : (r,\theta) \longmapsto G(r\cos\theta, r\sin\theta)$$

est différentiable dans $]r_1,r_2[\times]\theta_1,\theta_2[$, avec

$$\begin{aligned} \frac{\partial g}{\partial r}(r,\theta) &= \frac{\partial G}{\partial x}(r\cos\theta, r\sin\theta)\,\cos\theta + \frac{\partial G}{\partial y}(r\cos\theta, r\sin\theta)\,\sin\theta \\ \frac{\partial g}{\partial \theta}(r,\theta) &= -r\frac{\partial G}{\partial x}(r\cos\theta, r\sin\theta)\,\sin\theta + r\frac{\partial G}{\partial y}(r\cos\theta, r\sin\theta)\,\cos\theta\,. \end{aligned}$$

Exemple 4

Soit $F\ : (u,v,w) \longmapsto (x(u,v), y(u,v))$ une application d'un ouvert U de $\mathbb{R}^3$, à valeurs dans $\mathbb{R}^3$, telle que les applications coordonnées x, y, z soient toutes les trois différentiables en tout point de U. Soit V un ouvert de $\mathbb{R}^3$ contenant $F(U)$ et G une application différentiable de V dans $\mathbb{R}$.

L'application $G \circ F\ : U \longrightarrow \mathbb{R}$ est différentiable en tout point (u,v,w) de U et on a

$$\begin{aligned} \frac{\partial[G\circ F]}{\partial u}(u,v,w) &= \frac{\partial G}{\partial x}(F(u,v,w))\frac{\partial x}{\partial u} + \frac{\partial G}{\partial y}(F(u,v,w))\frac{\partial y}{\partial u} + \frac{\partial G}{\partial z}(F(u,v,w))\frac{\partial z}{\partial u} \\ \frac{\partial[G\circ F]}{\partial v}(u,v,w) &= \frac{\partial G}{\partial x}(F(u,v,w))\frac{\partial x}{\partial v} + \frac{\partial G}{\partial y}(F(u,v,w))\frac{\partial y}{\partial v} + \frac{\partial G}{\partial z}(F(u,v,w))\frac{\partial z}{\partial v} \\ \frac{\partial[G\circ F]}{\partial w}(u,v,w) &= \frac{\partial G}{\partial x}(F(u,v,w))\frac{\partial x}{\partial w} + \frac{\partial G}{\partial y}(F(u,v,w))\frac{\partial y}{\partial w} + \frac{\partial G}{\partial z}(F(u,v,w))\frac{\partial z}{\partial w}\,. \end{aligned}$$

Pour voir ce résultat, il suffit d'écrire explicitement ce que signifie la différentiabilité de G en $F(u,v,w)$, soit

$$\begin{aligned} G(x(u,v,w)+H, y(u,v,w)+K, z(u,v,w)+L) &= G(F(u,v,w)+(H,K,L)) \\ &= G(F(u,v,w)) + dG_{F(u,v,w)}(H,K,L) \\ &\quad + o(\sqrt{H^2+K^2+L^2}) \end{aligned} \tag{3.7}$$

et de remarquer que

$$\begin{aligned} x(u+h,v+k,w+l) &= x(u,v,w)+dx_{(u,v,w)}(h,k,l)+o(\sqrt{h^2+k^2+l^2})=x(u,v,w)+H \\ y(u+h,v+k,w+l) &= y(u,v,w)+dy_{(u,v,w)}(h,k,l)+o(\sqrt{h^2+k^2+l^2})=y(u,v,w)+K \\ z(u+h,v+k,w+l) &= z(u,v,w)+dz_{(u,v,w)}(h,k,l)+o(\sqrt{h^2+k^2+l^2})=z(u,v,w)+L \end{aligned}$$

avant de substituer dans (3.7) et de conclure à la différentiabilité de $G\circ F$ en (t,s).

Exemple (passage aux coordonnées sphériques dans l'espace). Si F est l'application

$$(r,\theta,\varphi)\in]r_1,r_2[\times]\theta_1,\theta_2[\times]\varphi_1,\varphi_2[\longmapsto (r\sin\varphi\,\cos\theta\,,\ r\sin\varphi\,\sin\theta,r\cos\varphi)\,,$$

et si G est une application différentiable dans le secteur ouvert $F(]r_1,r_2[\times]\theta_1,\theta_2[\times]\varphi_1,\varphi_2[)$ de l'espace et à valeurs dans $\mathbb{R}$, la fonction

$$g\ :(r,\theta,\varphi)\longmapsto G(r\sin\varphi\,\cos\theta,r\sin\varphi\,\sin\theta,r\cos\varphi)$$

est différentiable dans $]r_1,r_2[\times]\theta_1,\theta_2[\times]\varphi_1,\varphi_2[$, avec

$$\begin{aligned} \frac{\partial g}{\partial r}(r,\theta,\varphi) &= \frac{\partial G}{\partial x}(F(r,\theta,\varphi))\,\sin\varphi\,\cos\theta+\frac{\partial G}{\partial y}(F(r,\theta,\varphi))\,\sin\varphi\sin\theta+\frac{\partial G}{\partial y}(F(r,\theta,\varphi))\,\cos\varphi \\ \frac{\partial g}{\partial \theta}(r,\theta,\varphi) &= -r\frac{\partial G}{\partial x}(F(r,\theta,\varphi))\,\sin\varphi\,\sin\theta+r\frac{\partial G}{\partial y}(F(r,\theta,\varphi))\,\sin\varphi\cos\theta \\ \frac{\partial g}{\partial \varphi}(r,\theta,\varphi) &= -r\frac{\partial G}{\partial x}(F(r,\theta,\varphi))\,\cos\varphi\,\cos\theta+r\frac{\partial G}{\partial y}(F(r,\theta,\varphi))\,\cos\varphi\sin\theta-r\frac{\partial G}{\partial z}(F(r,\theta,\varphi))\sin\varphi\,. \end{aligned}$$

3.7.5 Dérivées d'ordre supérieur ; laplacien

Soit f une fonction différentiable de deux (*resp.* trois) variables dans un ouvert U de $\mathbb{R}^2$ (*resp.* de $\mathbb{R}^3$), à valeurs dans $\mathbb{R}$. On dit que f est *deux fois différentiable* en un point (x_0,y_0) (*resp.* (x_0,y_0,z_0)) de U si les deux fonctions $\partial f/\partial x$, $\partial f/\partial y$ (*resp.* les trois fonctions $\partial f/\partial x$, $\partial f/\partial y$, $\partial f/\partial z$) sont différentiables en (x_0,y_0) (*resp.* en (x_0,y_0,z_0)).

Le fait que f soit deux fois différentiable en un point de son ouvert de définition implique donc l'existence de dérivées partielles à l'ordre 2 en ce point.

Pour une fonction de deux variables, ces dérivées partielles à l'ordre 2 sont :

$$\begin{aligned} \frac{\partial^2 f}{\partial x^2}(x_0,y_0) &= \frac{d}{dx}\Big[\frac{\partial f}{\partial x}(x,\mathbf{y})\Big](x_0,y_0) \\ \frac{\partial^2 f}{\partial y\partial x}(x_0,y_0) &= \frac{d}{dy}\Big[\frac{\partial f}{\partial x}(\mathbf{x},y)\Big](x_0,y_0) \\ \frac{\partial^2 f}{\partial x\partial y}(x_0,y_0) &= \frac{d}{dx}\Big[\frac{\partial f}{\partial y}(x,\mathbf{y})\Big](x_0,y_0) \\ \frac{\partial^2 f}{\partial y^2}(x_0,y_0) &= \frac{d}{dy}\Big[\frac{\partial f}{\partial y}(\mathbf{x},y)\Big](x_0,y_0) \end{aligned}$$

(mêmes règles de « coloriage » des variables que précédemment pour effectuer ces calculs de dérivées partielles à l'ordre 2 du point de vue pratique). En fait, un lemme important de calcul différentiel, le lemme de Schwarz [7], assure, lorsque f est deux fois différentiable au point (x_0,y_0) l'égalité automatique des dérivées croisées

$$\frac{\partial^2 f}{\partial y\partial x}(x_0,y_0)=\frac{\partial^2 f}{\partial x\partial y}(x_0,y_0)\,.$$

7. On admettra ici ce lemme attribué au mathématicien allemand Hermann Schwarz (1843-1921), constamment utilisé en analyse.

Formellement, on peut même écrire, si f est deux fois différentiable en (x_0, y_0),

$$f(x_0+h, y_0+k) = f(x_0,y_0) + \Big(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\Big)[f](x_0,y_0) + \frac{1}{2}\Big(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\Big)^2[f](x_0,y_0) + o(h^2+k^2)\,,$$

ce qui permet d'approcher les variations de f au voisinage de (x_0, y_0) à l'ordre deux en fonction de la petite perturbation (h, k).

Pour une fonction de trois variables, les dérivées partielles à l'ordre 2 sont :

$$\begin{aligned}
\frac{\partial^2 f}{\partial x^2}(x_0,y_0,z_0) &= \frac{d}{dx}\Big[\frac{\partial f}{\partial x}(x,\mathbf{y},\mathbf{z})\Big](x_0,y_0,z_0)\\
\frac{\partial^2 f}{\partial y\partial x}(x_0,y_0,z_0) &= \frac{d}{dy}\Big[\frac{\partial f}{\partial x}(\mathbf{x},y,\mathbf{z})\Big](x_0,y_0,z_0) = \frac{\partial^2 f}{\partial x\partial y}(x_0,y_0,z_0) = \frac{d}{dx}\Big[\frac{\partial f}{\partial y}(x,\mathbf{y},\mathbf{z})\Big](x_0,y_0,z_0)\\
\frac{\partial^2 f}{\partial z\partial x}(x_0,y_0,z_0) &= \frac{d}{dz}\Big[\frac{\partial f}{\partial x}(\mathbf{x},\mathbf{y},z)\Big](x_0,y_0,z_0) = \frac{\partial^2 f}{\partial x\partial z}(x_0,y_0,z_0) = \frac{d}{dx}\Big[\frac{\partial f}{\partial z}(x,\mathbf{y},\mathbf{z})\Big](x_0,y_0,z_0)\\
\frac{\partial^2 f}{\partial y^2}(x_0,y_0,z_0) &= \frac{d}{dy}\Big[\frac{\partial f}{\partial y}(\mathbf{x},y,\mathbf{z})\Big](x_0,y_0,z_0)\\
\frac{\partial^2 f}{\partial z\partial y}(x_0,y_0,z_0) &= \frac{d}{dz}\Big[\frac{\partial f}{\partial y}(\mathbf{x},\mathbf{y},z)\Big](x_0,y_0,z_0) = \frac{\partial^2 f}{\partial y\partial z}(x_0,y_0,z_0) = \frac{d}{dy}\Big[\frac{\partial f}{\partial z}(\mathbf{x},y,\mathbf{z})\Big](x_0,y_0,z_0)\\
\frac{\partial^2 f}{\partial z^2}(x_0,y_0,z_0) &= \frac{d}{dz}\Big[\frac{\partial f}{\partial z}(\mathbf{x},\mathbf{y},z)\Big](x_0,y_0,z_0)
\end{aligned}$$

(mêmes règles de « coloriage » des variables que précédemment pour effectuer ces calculs de dérivées partielles à l'ordre 2 du point de vue pratique). Les égalités entre dérivées « croisées » viennent encore du lemme de Schwarz. Formellement, on peut même écrire, si f est deux fois différentiable en (x_0, y_0, z_0),

$$\begin{aligned}
f(x_0+h, y_0+k, z_0+l) &= f(x_0,y_0) + \Big(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} + l\frac{\partial}{\partial z}\Big)[f](x_0,y_0)\\
&\quad + \frac{1}{2}\Big(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} + l\frac{\partial}{\partial z}\Big)^2[f](x_0,y_0,z_0) + o(h^2+k^2+l^2)\,,
\end{aligned}$$

ce qui permet d'approcher les variations de f au voisinage de (x_0, y_0, z_0) à l'ordre deux en fonction de la petite perturbation (h, k, l).

Une fonction f de deux ou trois variables, à valeurs réelles, qui admet des dérivées partielles jusqu'à l'ordre 2 continues sur un ouvert U est deux fois dfférentiable en tout point de cet ouvert ; pour rendre compte de la continuité des dérivées partielles à l'ordre 2 sur U, on dit que f est *de classe* C^2 dans l'ouvert U.

Pour une fonction de classe C^2 dans un ouvert U de $\mathbb{R}^2$, on définit le *laplacien*[8] de f comme la fonction

$$\Delta[f] := \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}\,;$$

c'est donc une fonction continue de U dans $\mathbb{R}$. Même chose pour une fonction de classe C^2 dans un ouvert U de $\mathbb{R}^3$, dont le *laplacien* est défini comme la fonction

$$\Delta[f] := \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}\,.$$

8. Mathématicien, astronome, probabiliste, mais aussi homme politique sous le consulat et l'empire, Pierre-Simon Laplace (1749-1827) a posé les jalons de la mécanique et de l'analyse moderne.

Exemple. Le laplacien de la fonction

$$(x,y) \in \mathbb{R}^2 \setminus \{(0,0)\} \longmapsto \log \sqrt{x^2+y^2}$$

est identiquement nul, comme celui du potentiel newtonien

$$(x,y,z) \in \mathbb{R}^3 \setminus \{(0,0,0)\} \longmapsto -\frac{1}{\sqrt{x^2+y^2+z^2}\,.}$$

Le laplacien est un opérateur différentiel du second ordre très important du point de vue pratique : pour les fonctions de plusieurs variables, c'est lui qui permet de mettre en évidence les variations d'une fonction (par exemple les « lignes de rupture » et les « contours » des objets dans une image 2-dimensionnelle), ce que fait la dérivée pour les fonctions d'une variable.

3.7.6 Champs de vecteurs dans un ouvert du plan ou de l'espace

Un *champ de vecteurs* dans un ouvert U du plan est la donnée d'une application de U dans $\mathbb{R}^2$.

Exemples. Le gradient d'une fonction $f : U \longrightarrow \mathbb{R}$ différentiable dans U est un exemple de champ de vecteurs. Les champs magnétiques ou électromagnétiques $\vec{F}$ en physique peuvent par exemple être visualisés grâce à des paticules de limaille de fer aimantées, chaque particule se positionnant au point (x,y) suivant la direction dans laquelle pointe le vecteur $\vec{F}(x,y)$. Une carte météorologique (par exemple une carte des vents) illustre également la visualisation d'un champ de vecteurs.

De même, un *champ de vecteurs* dans un ouvert U de l'espace est la donnée d'une application de U dans $\mathbb{R}^3$.

Si $(x,y,z) \longmapsto \vec{F}(x,y,z)$ est un champ de vecteurs dans un ouvert U de $\mathbb{R}^3$ et si les applications coordonnées P,Q,R de $\vec{F} = P\,\vec{i} + Q\,\vec{j} + R\,\vec{k}$ sont différentiables dans U, on associe au champ de vecteurs $\vec{F}$ un nouveau champ de vecteurs dans l'ouvert U, le *rotationnel* de $\vec{F}$, défini formellement par

$$\begin{aligned} \overrightarrow{\mathrm{rot}}\,\vec{F} : &= \Big(\frac{\partial}{\partial x}\,\vec{i} + \frac{\partial}{\partial y}\,\vec{j} + \frac{\partial}{\partial z}\,\vec{k}\Big) \wedge (P\,\vec{i} + Q\,\vec{j} + R\,\vec{k}) \\ &= \Big(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\Big)\,\vec{i} + \Big(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\Big)\,\vec{j} + \Big(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\Big)\,\vec{k} \end{aligned}$$

(le produit extérieur de deux vecteurs de l'espace ayant été défini en (3.5), remarquer encore l'aspect « tournant » des formules pour les retenir).

Lorsque $\vec{F}$ s'écrit $\overrightarrow{\nabla f}$, où f est une fonction de classe C^2 dans U (on dit aussi dans ce cas que le champ de vecteurs $\vec{F}$ *dérive du potentiel scalaire* f), alors on a

$$\overrightarrow{\mathrm{rot}}\,\vec{F} = \overrightarrow{\mathrm{rot}}\,(\overrightarrow{\nabla f}) \equiv \vec{0}\,,$$

autrement dit le rotationnel d'un champ dérivant d'un potentiel est nul (il suffit de faire le calcul et d'appliquer le lemme de Schwarz assurant l'égalité des dérivées secondes croisées).

Lorsque U est un ouvert de $\mathbb{R}^2$ et $\vec{F} = (P,Q)$ un champ de vecteurs différentiable dans U, on peut considérer le champ de vecteurs

$$(x,y,z) \in U \times \mathbb{R} \longmapsto \vec{\mathbf{F}}(x,y,z) := P(x,y)\,\vec{i} + Q(x,y)\,\vec{j}$$

dans l'ouvert $U \times \mathbb{R}$ de $\mathbb{R}^3$; on a

$$\overrightarrow{\mathrm{rot}}\,\vec{\mathbf{F}} = \Big(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\Big)\,\vec{k}\,.$$

Ce rotationnel est nul si $\vec{F}$ s'écrit $\overrightarrow{\nabla f}$, où f est une fonction de classe C^2 de U dans $\mathbb{R}$. Le rotationnel de $\vec{\mathbf{F}}$ ainsi associé au champ de vecteurs $\vec{F} = (P, Q)$ dans l'ouvert U de $\mathbb{R}^2$ est un champ de vecteurs mettant en évidence les aspects « tourbillonaires » du champ $\vec{F}$, d'où la terminologie de *rotationnel*; voici par exemple (sur la figure 3.15 ci-dessous) l'image de

$$(x, y) \longmapsto \Big|\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\Big| = \|\overrightarrow{\mathrm{rot}}\,\vec{\mathbf{F}}\|$$

lorsque $\vec{F} = (P, Q)$ est le champ de vitesse des particules dans un écoulement turbulent (on note la mise en évidence des tourbillons).

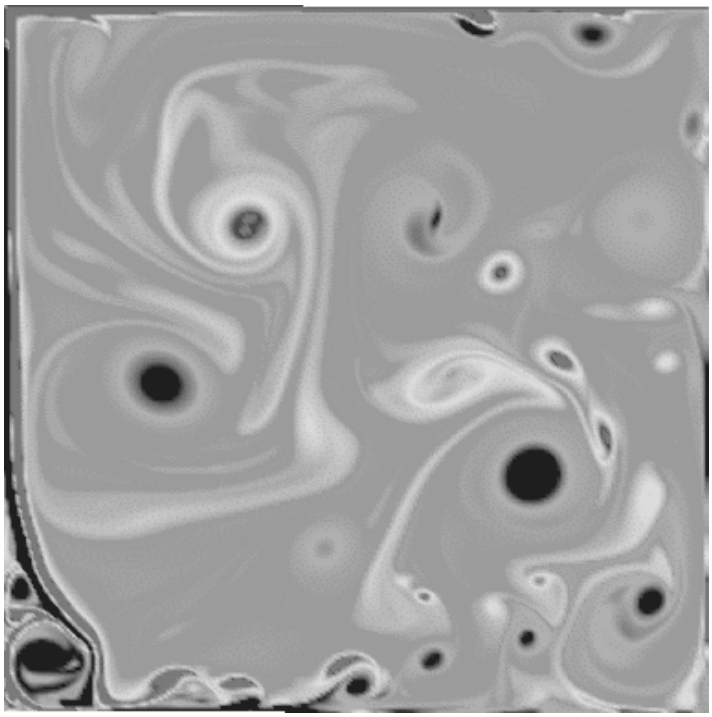

FIGURE 3.15 – Visualisation du rotationnel du champ de vitesses dans un écoulement turbulent

Si $\vec{F} = (P, Q, R)$ est un champ de vecteurs différentiable dans un ouvert U de $\mathbb{R}^3$, on peut associer au champ $\vec{F}$ une fonction scalaire importante, la *divergence* du champ $\vec{F}$ définie (dans l'ouvert U) commme la fonction

$$(x, y, z) \in U \longmapsto \mathrm{div}\,\vec{F}(x, y, z) := \frac{\partial P}{\partial x}(x, y, z) + \frac{\partial Q}{\partial y}(x, y, z) + \frac{\partial R}{\partial z}(x, y, z)\,.$$

Exemples. Si $\vec{F} = \overrightarrow{\nabla f}$, où f est une fonction de classe C^2 d'un ouvert U de $\mathbb{R}^3$ dans $\mathbb{R}$, on a

$$\mathrm{div}\,(\overrightarrow{\nabla f}) = \frac{\partial}{\partial x}\Big[\frac{\partial f}{\partial x}\Big] + \frac{\partial}{\partial y}\Big[\frac{\partial f}{\partial y}\Big] + \frac{\partial}{\partial z}\Big[\frac{\partial f}{\partial z}\Big] = \Delta\, f\,.$$

Si $\vec{F} = \overrightarrow{\mathrm{rot}}\,\vec{G}$, où $\vec{G} = (P, Q, R)$ est un champ de vecteurs de classe C^2 dans un ouvert U de $\mathbb{R}^3$, on a

$$\mathrm{div}\,(\overrightarrow{\mathrm{rot}}\,\vec{G}) = \frac{\partial}{\partial x}\Big(\frac{\partial R}{\partial y} - \frac{\partial Q}{\partial z}\Big) + \frac{\partial}{\partial y}\Big(\frac{\partial P}{\partial z} - \frac{\partial R}{\partial x}\Big) + \frac{\partial}{\partial z}\Big(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\Big) \equiv 0$$

grâce encore au lemme de Schwarz sur l'égalité des dérivées croisées.

Voici un petit formulaire récapitulatif des opérations sur les champs de vecteurs ou les fonctions scalaires.

$$\overrightarrow{\mathrm{rot}}\,[\overrightarrow{\nabla F}] = \vec{0}$$

$$\overrightarrow{\mathrm{rot}}\,[f\,\vec{F}] = \overrightarrow{\nabla f} \wedge \vec{F} + f\,\overrightarrow{\mathrm{rot}}\,\vec{f}$$

$$\mathrm{div}\,[\overrightarrow{\nabla f}] = \Delta f$$

$$\mathrm{div}\,[\overrightarrow{\mathrm{rot}}\,\vec{G}] = 0$$

$$\mathrm{div}\,[f\,\vec{F}] = \langle \overrightarrow{\nabla f}\,,\,\vec{F}\rangle + f\,\mathrm{div}\,\vec{F}$$

$$\mathrm{div}\,[f_1\,\overrightarrow{\nabla f_2}] = \langle \overrightarrow{\nabla f_1}\,,\,\overrightarrow{\nabla f_2}\rangle + f_1\,\Delta\,f_2$$

$$\mathrm{div}\,[\vec{F}_1 \wedge \vec{F}_2] = \langle \overrightarrow{\mathrm{rot}}\,\vec{F}_1\,,\,\vec{F}_2\rangle - \langle \overrightarrow{\mathrm{rot}}\,\vec{F}_2\,,\,\vec{F}_1\rangle$$

$$\Delta\left[\log\sqrt{x^2+y^2}\right] = 0 \quad \forall\,(x,y) \in \mathbb{R}^2 \setminus \{(0,0)\}$$

$$\Delta\Big[-\frac{1}{\sqrt{x^2+y^2+z^2}}\Big] = 0 \quad \forall\,(x,y,z) \in \mathbb{R}^3 \setminus \{(0,0,0)\}$$

(on pourra s'entraîner à faire toutes les vérifications en exercices formellement).

3.8 Aires, intégration, primitives

3.8.1 La notion d'aire d'un domaine plan

Dans le plan $\mathbb{R}^2$, on sait définir l'aire d'un "pavé"

$$P := [a_1, b_1] \times [a_2, b_2]$$

(avec $-\infty < a_j < b_j < +\infty$, $j = 1, 2$) ; cette aire vaut par définition

$$\mathrm{aire}\,(P) := (b_1 - a_1) \times (b_2 - a_2)\,.$$

On sait donc aussi calculer l'aire d'une union finie R de pavés en découpant cette union comme une mosaïque de pavés P' dont les intérieurs $]a_1, b_1[\times]a_2, b_2[$ sont disjoints ; l'aire de l'union est dans ce cas la somme des aires des pavés P' (voir figure 3.16 ci-dessous).

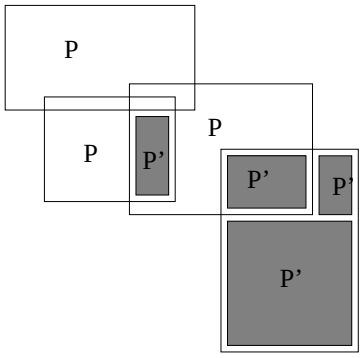

FIGURE 3.16 – L'aire d'une union R de pavés

La physique fait apparaître bien souvent des structures (qualifiées de fractales) s'auto-reproduisant à toutes les échelles, comme par exemple le flocon de neige (on a déjà rencontré le flocon de von Koch) ; si l'on imagine un tel flocon planaire ou une figure fractale telle l'ensemble de Mandelbrojt représenté en clair sur la figure 3.17 suivante, il est difficile d'emblée de décider quelle est l'aire de ce domaine ! Dans l'espace, la situation est encore plus compliquée : quelle est par exemple l'aire (ramenée à une aire plane) d'un piton rocheux tel l'aiguille du Midi à Chamonix ?

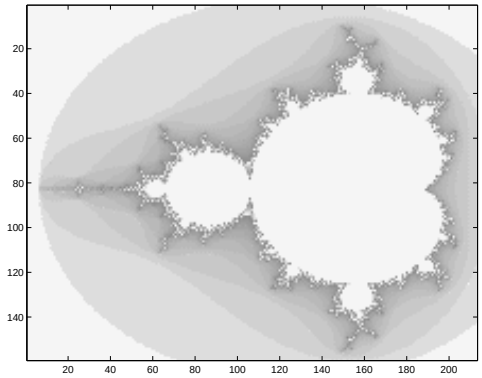

FIGURE 3.17 – Une structure planaire "fractale"

Le paradoxe de Banach-Tarski évoqué au chapitre 1 (on peut découper une sphère pleine comme un puzzle et réaliser avec les morceaux une sphère pleine de volume double) nous laisse pressentir qu'il existe des sous-ensembles de l'espace $\mathbb{R}^3$ que l'on sera toujours incapable de mesurer ; il en est de même dans le plan.

La méthode la plus naïve de tenter de calculer une aire est inspirée des probabilités et est d'origine très ancienne : supposons par exemple que le "cadre" de la figure 3.17 soit d'aire normalisée égale à 1. On jette des points sur ce cadre, tous les points de chute étant équiprobables. On fait cela un nombre N (très très grand de fois), puis on divise par N le nombre de jets (N_{fav}) où le point est tombé sur la cible (le domaine A dont on veut calculer l'aire). Dans les bons cas (c'est-à-dire lorsque parler de cette aire a un sens), la loi des grands nombres (outil clef du principe sur lequel se fonde le raisonnement statistique, comme dans les sondages d'opinion par exemple) permet d'assurer que lorsque N (le nombre de jets,

ceux ci étant supposés indépendants) tend vers $+\infty$, alors le quotient N_{fav}/N tend vers l'aire de A. Cette méthode très intuitive est connue comme la *méthode de Monte Carlo* et permet, lentement, c'est vrai, de calculer l'aire d'un domaine si cela est possible.

Nous nous proposons maintenant de décrire succintement, après les méthodes d'inspiration "probabiliste", les méthodes numériques de calcul d'aire ; ce sont bien sûr elles qui, dans 99% des cas, sont les seules utilisables pour calculer une aire (et donc une intégrale). Ce sont ces méthodes qui ont guidé la genèse de la théorie de l'intégration, depuis Riemann au XIX-ème siècle jusqu'à Henri Lebesgue au début du XX-ème siècle. C'est le point de vue de Lebesgue qui soutend ici en filigramme notre esquisse de présentation. Les méthodes passant par la recherche de formules exactes ne sont utilisables que dans des cas très particulier que nous décrirons ultérieurement (sections 3.7.2 et 3.7.3 à venir).

Ce qu'il est *a priori* facile de faire est de tenter de mesurer un ensemble borné A du plan "de l'extérieur" en définissant son *aire supérieure* comme la borne inférieure (c'est-à-dire le plus grand minorant) de l'ensemble

$$\{\text{aire}\,(R)\,;\ R \text{ union de pavés contenant } A\}\ ;$$

on note cette "aire supérieure" $\text{aire}^*(A)$ (voir la figure 3.18).

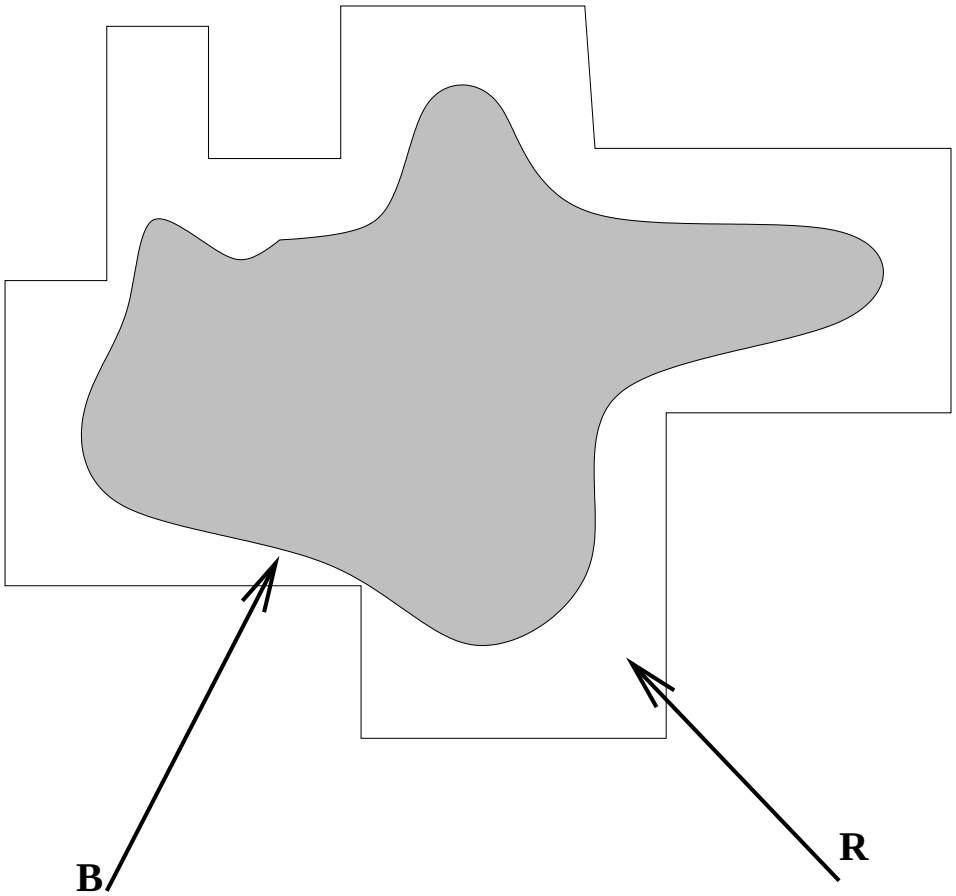

FIGURE 3.18 – Le calcul de l'aire supérieure d'un sous-ensemble borné B du plan

Si maintenant A est un sous-ensemble borné du plan, on dit que l'on peut le mesurer si, pour tout $\epsilon > 0$, on peut trouver une union finie de pavés R_ϵ telle que l'aire supérieure de l'ensemble

$$\begin{aligned}\Delta(A, R_\epsilon) &:= \{(x,y) \in A \text{ tels que } (x,y) \notin R_\epsilon\} \\ &\qquad \cup \{(x,y) \in R_\epsilon \text{ tels que } (x,y) \notin A\} \\ &= A \,\Delta\, R_\epsilon\end{aligned}$$

soit au plus ϵ ; sur la figure 3.19, on a représenté un tel ensemble $\Delta(A, R_\epsilon)$. On peut mesurer l'ensemble A si l'on peut trouver des unions finies de rectangles qui "collent" au mieux à A, excepté sur un ensemble d'aire supérieure arbitrairement petite (c'est-à-dire intuitivement occupant un espace arbitrairement petit dans le plan).

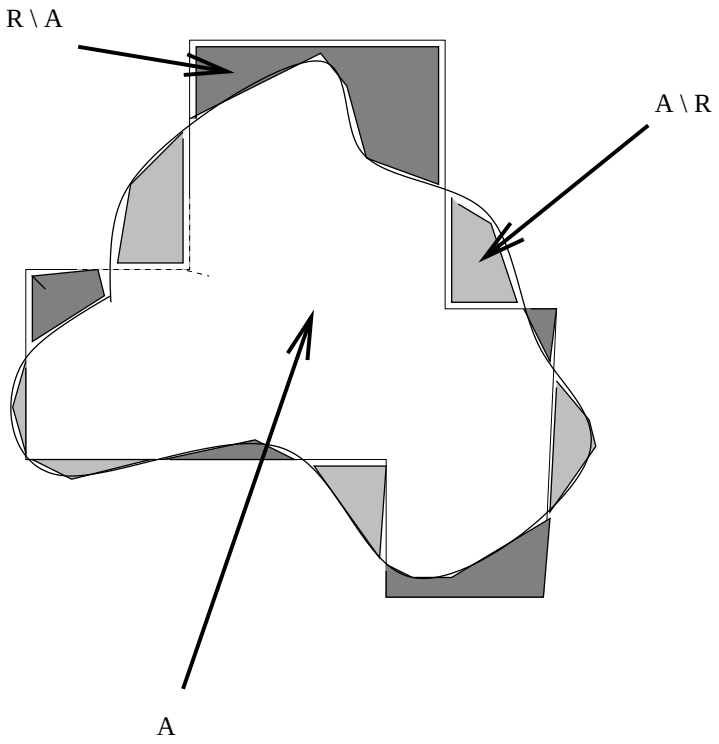

FIGURE 3.19 – La différence symétrique $\Delta(A, R)$ de A et d'une union finie de pavés R

Si A est un sous-ensemble borné du plan que l'on est capable de mesurer, on appelle *aire de A* son aire supérieure ; si P_0 est un grand pavé contenant A, on voit que cette aire supérieure est aussi égale au nombre $\text{aire}\,(P_0) - \text{aire}^*(P_0 \setminus A)$, ce qui signifie essentiellement que l'on obtient le même nombre en approchant A "de l'extérieur" ou "de l'intérieur" ; c'est là précisément ce qui caractérise les sous-ensembles du plan que l'on est capable de mesurer et qui permet d'en définir sans aucune ambigüité l'aire.

Si f est une fonction définie sur D et à valeurs dans $[0, +\infty[$, on appelle *sous-graphe* de f l'ensemble

$$\text{SG}\,(f) := \{(x,y) \in D \times \mathbb{R}\,;\, 0 \leq y \leq f(x)\}\,.$$

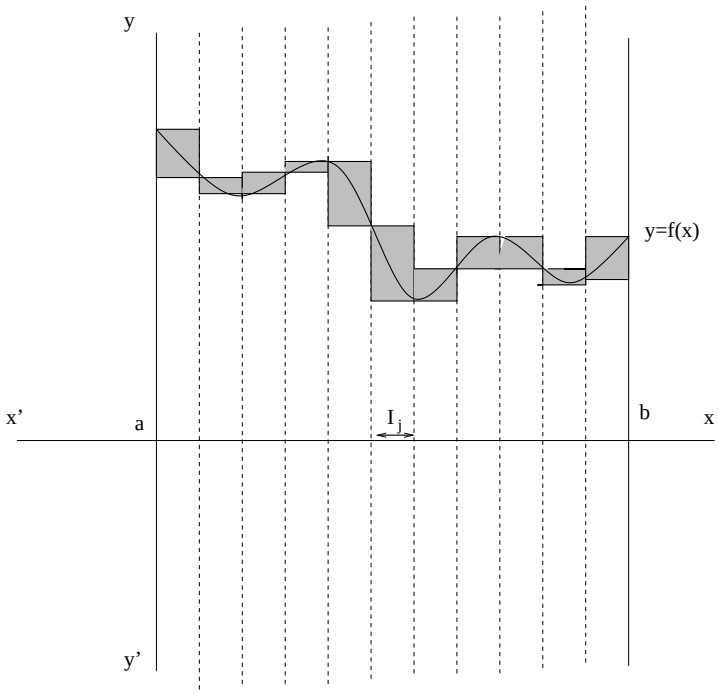

FIGURE 3.20 – Le sous-graphe de f et l'ensemble Δ_N

Le sous-graphe A d'une fonction continue positive sur un segment $[a,b]$ est un exemple très important d'ensemble que l'on sait mesurer. Pour voir cela, on découpe l'intervalle $[a,b]$ en N segments égaux I_k (de longueur $(b-a)/N$), $k = 1, ..., N$ puis on introduit les deux ensembles $R_N^{\sup}$ et $R_N^{\inf}$ définis respectivement par

$$R_N^{\sup} := \bigcup_{k=1}^{N} (I_k \times [0, \sup_{I_k} f])$$
$$R_N^{\inf} := \bigcup_{k=1}^{N} (I_k \times [0, \inf_{I_k} f]) .$$

(voir la figure 3.20 ci-dessous). L'ensemble $A \,\Delta\, R_N^{\sup}$ est inclus dans $R_N^{\sup} \setminus R_N^{\inf}$, donc dans l'union Δ_N des pavés $I_k \times [\inf_{I_k} f, \sup_{I_k} f]$ et son aire supérieure est donc majorée par

$$\frac{b-a}{N} \sum_{k=1}^{N} (\sup_{I_k} f - \inf_{I_k} f) ;$$

comme f est continue, on admettra que pour N assez grand, on peut faire en sorte que pour tout intervalle I_k de la subdivision, on ait

$$\sup_{I_k} f - \inf_{I_k} f < \epsilon .$$

L'aire supérieure de la différence symétrique $A \,\Delta\, R_N^{\sup}$ peut donc être rendue arbitrairement petite pour un choix convenable d'union de pavés $R_N^{\sup}$ et le graphe de f est un ensemble que l'on peut mesurer.

Toute fonction réelle continue f sur un intervalle $[a,b]$ s'écrit comme la différence de deux fonctions positives continues, à savoir $f^+ := \sup(f,0)$ et $f^- := \sup(-f,0)$. Ceci nous permet de définir la notion d'intégrale de f sur $[a,b]$ (en remarquant que f s'écrit $f = f^+ - f^-$ sur $[a,b]$).

Définition 3.11. Si f est une fonction continue de $[a,b]$ dans $\mathbb{R}$, on appelle *intégrale de* f sur $[a,b]$ le nombre réel

$$\int_{[a,b]} f(t), dt = \int_a^b f(t)dt = \text{Aire}\left(\text{SG}\left(f^+\right)\right) - \text{Aire}\left(\text{SG}\left(f^-\right)\right).$$

On a les formules évidentes suivantes, si f et g sont deux fonctions continues sur $[a,b]$, avec $a < b$:

$$\begin{aligned}
\int_a^b f(t)\,dt + \int_a^b g(t)\,dt &= \int_a^b (f(t)+g(t))\,dt \\
\int_a^b f(t)\,dt &= \int_a^c f(t)\,dt + \int_c^b f(t)\,dt\,,\ \forall c \in]a,b[\\
(f \geq 0 \text{ sur } [a,b]) &\implies \int_a^b f(t)\,dt \geq 0 \\
(f \leq g \text{ sur } [a,b]) &\implies \int_a^b f(t)\,dt \leq \int_a^b g(t)\,dt \\
\Big|\int_a^b f(t)\,dt\Big| &\leq \int_a^b |f(t)|\,dt\,.
\end{aligned}$$

La seconde relation est dite *relation de Chasles*; la quatrième relation traduit la *monotonie de la prise d'intégrale*; la dernière inégalité joue un rôle majeur et s'interprète comme la "version continue" de l'inégalité triangulaire.

On convient, si $a < b$, de noter

$$\int_b^a f(t)\,dt := -\int_a^b f(t)\,dt$$

et, si $a = b$,

$$\int_a^a f(t)\,dt = 0\,;$$

avec ces conventions, la relation de Chasles reste valable si a, b, c sont trois points quelconques et dans un ordre quelconque d'un intervalle I de $\mathbb{R}$ sur lequel la fonction f est continue. Ceci sera important pour la suite. Si a et b sont deux nombres réels et f une fonction continue sur $[a,b]$, on conviendra de noter

$$\int_{[a,b]} f(t) = \epsilon \int_a^b f(t)dt$$

avec $\epsilon = 1$ si $a \leq b$ et $\epsilon = -1$ si a est strictement supérieur à b.

3.8.2 Primitive d'une fonction continue sur un intervalle ouvert de $\mathbb{R}$

On peut maintenant énoncer un résultat majeur, dit parfois *théorème fondamental de l'analyse*, transcription en dimension 1 d'une formule bien connue des physiciens en dimension supérieure, la formule de Stokes :

Théorème fondamental de l'analyse *Soit f une fonction continue sur un intervalle ouvert I de $\mathbb{R}$, à valeurs réelles, et a un point de I. La fonction*

$$F\ : x \in I \longmapsto \int_a^x f(t)\,dt$$

est dérivable sur I, de dérivée $F' \equiv f$. On dit que F est une primitive de f ; c'est d'ailleurs la primitive de f qui s'annule en a (une telle primitive est unique).

Preuve. Avec la relation de Chasles et les propriétés de l'intégrale, on a, si $x_0 \in I$ et h est assez petit (pour que l'intervalle de bornes x_0 et $x_0 + h$ soit dans I)

$$F(x_0 + h) = F(x_0) + f(x_0)h + \int_{x_0}^{x_0+h} (f(t) - f(x_0))\, dt\,.$$

Comme f est continue en x_0,

$$\forall \epsilon > 0 \ \exists \eta > 0 \text{ tel que } : \Big((|h| < \eta) \wedge (t \in [x_0, x_0 + h])\Big) \Longrightarrow |f(t) - f(t_0)| \leq \epsilon\,.$$

On a donc, pour $|h| < \eta$,

$$|F(x_0 + h) - F(x_0) - hf(x_0)| \leq \int_{[x_0,x_0+h]} |f(t) - f(x_0)|\, dt \leq \int_{[x_0,x_0+h]} \epsilon\, dt = \epsilon|h|\,,$$

donc

$$F(x_0 + h) = F(x_0) + hf(x_0) + o(h)\,,$$

ce qui démontre la première partie de notre résultat.

Supposons maintenant que G soit une autre primitive de f s'annulant en a. La fonction $G - F$ est une fonction dérivable sur I et de dérivée identiquement nulle ; d'après la proposition 3.12 (admise), la fonction $G - F$ est constante ; comme elle est nulle en $x = a$, elle est identiquement nulle et l'on a $G \equiv F$, ce qui prouve la seconde partie de notre théorème (la clause d'unicité de la primitive s'annulant en a). □

Le théorème fondamental de l'analyse est souvent confondu avec son corollaire immédiat :

Corollaire. *Si f est une fonction à valeurs réelles définie et dérivable sur un intervalle ouvert I et si $[a, b] \subset I$, on a*

$$\int_a^b f'(t)\, dt = f(b) - f(a)\,.$$

Une application majeure de ce résultat est la très célèbre car très utile formule suivante :

Formule d'intégration par parties. *Soient f et g deux fonctions à valeurs réelles dérivables sur un intervalle ouvert I de $\mathbb{R}$, à dérivées continues sur I ; on a, si $[a, b] \subset I$,*

$$\int_a^b f'(t)g(t)\, dt = -\int_a^b f(t)g'(t)\, dt + f(b)g(b) - f(a)g(a) = -\int_a^b f'(t)g(t)\, dt + \Big[fg\Big]_a^b\,.$$

Preuve. On applique le corollaire à la fonction $h = fg$ en remarquant que $h' = f'g + fg'$ d'après la règle de Leibniz. □

Autre application très importante pour les applications, la formule de changement de variables.

Formule de changement de variables. *Soient I et J deux intervalles ouverts de $\mathbb{R}$, $[a, b] \subset I$, et u une application de I dans $\mathbb{R}$, dérivable sur I, strictement monotone sur $[a, b]$, de dérivée u' continue sur $[a, b]$, avec $u(I) \subset J$. Soit $c = u(a)$ et $d = u(b)$; alors, pour toute fonction f à valeurs réelles continue sur J*

$$\int_c^d f(t)dt = \int_a^b f(u(x))u'(x)dx\,. \qquad (\dagger)$$

On a donc

$$\int_{[c,d]} f(t)dt = \int_{[a,b]} f(u(x))u'(x)dx$$

si u est strictement monotone croissante et

$$\int_{[c,d]} f(t)dt = -\int_{[a,b]} f(u(x))u'(x)dx$$

si u est strictement monotone décroissante, ce que l'on peut résumer en la seconde formule

$$\int_{[c,d]} f(t)dt = \int_{[a,b]} f(u(x))|u'(x)|\,dx \qquad (\dagger\dagger)$$

Preuve. On prouve la première formule (†) (la seconde en résulte immédiatement). Soit F une primitive de f sur J ; la fonction $F \circ u$ est alors une primitive de $f \circ u$ sur I d'après la règle de Leibniz. On a donc

$$\int_a^b f(u(x))u'(x)dx = F(u(b)) - F(u(a)) = F(d) - F(c) = \int_c^d f(t)\,dt\,,$$

ce qui est la formule (†). □

Remarque. On remarque que l'on a pas utilisé le fait que u soit strictement monotone sur I, mais simplement le fait que $u(I) \subset J$, donc que l'on peut composer u avec une primitive de f sur J. Dans la formulation que nous avons donné ensuite (formule (††)), cette monotonie intervient.

3.8.3 Calcul d'intégrales et calcul de primitives

Le calcul de l'intégrale d'une fonction continue sur un intervalle se fait de manière numérique ; pour une fonction positive sur $[a,b]$, on reprend la subdivision de $[a,b]$ en N intervalles I_j, $j = 1, ..., N$ consécutifs de même longueur et on retient l'encadrement de l'intégrale

$$\frac{b-a}{N}\sum_{j=1}^{N} \inf_{I_j} f \le \int_a^b f(t)\,dt \le \frac{b-a}{N}\sum_{j=1}^{N} \sup_{I_j} f\,;$$

on calcule l'intégrale d'une fonction de signe quelconque en écrivant la fonction sous la forme $f = f^+ - f^- = \sup(f,0) - \sup(-f,0)$ sur $[a,b]$. C'est évidemment là le moyen universel de calculer une intégrale.

Souvent on a besoin en physique de calculer l'intégrale d'une fonction continue sur $[a,b]$, mais à valeurs complexes (ce qui signifie que les fonctions $\mathrm{Re}\, f$ et $\mathrm{Im}\, f$ (parties réelle et imaginaire de f) sont continues. On définit alors l'intégrale de f sur le segment $[a,b]$ par

$$\int_a^b f(t)\,dt := \int_a^b \mathrm{Re}\, f(t)\,dt + i\int_a^b \mathrm{Im}\, f(t)\,dt$$

et le calcul se fait en général numériquement.

Cependant l"'herbier" de fonctions que nous avons construit nous permet de connaître parfois une primitive F de la fonction f, auquel cas le calcul exact de l'intégrale de f sur $[a,b]$ est possible *via* la formule

$$\int_a^b f(t)\,dt = F(b) - F(a)$$

(théorème fondamental de l'analyse).

Les formules d'intégration par parties ou de changement de variables fournissent des méthodes pour ramener une intégrale à celle d'une fonction plus simple. Les logiciels de calcul formel (comme Maple ou Mathematica, que nous avons exploité en cours) intègrent le formalisme algébrique qui soutend, lorsque cela est possible, ces calculs ; leur efficacité montre, si besoin est, combien il est difficile aujourd'hui de concevoir les mathématiques en se passant de l'aide qu'ils sont susceptibles de nous apporter.

Exemples.

On verra dans la section suivante comment se calcule la primitive d'une fraction rationnelle ; nous laissons donc cet exemple de côté pour l'instant et indiquons comment exploiter la formule de changement de variables ou la formule d'intégration par parties pour ramener le calcul de primitive ou le calcul d'intégrales à ce cadre.

1. Les expressions rationnelles en les lignes trigonométriques

Si $[a,b]$ est un segment inclus dans $]-\pi,\pi[$ et si P et Q sont deux polynômes en deux variables X,Y (à coefficients réels) tels que

$$Q(\cos\theta,\sin\theta)\neq 0\,,\ \ \theta\in[a,b]\,,$$

alors

$$\int_a^b \frac{P(\cos\theta,\sin\theta)}{Q(\cos\theta,\sin\theta)}\,d\theta=\int_{\tan(a/2)}^{\tan(b/2)} \frac{P\Big(\frac{1-u^2}{1+u^2},\frac{2u}{1+u^2}\Big)}{Q\Big(\frac{1-u^2}{1+u^2},\frac{2u}{1+u^2}\Big)}\,\frac{2du}{1+u^2}\,,$$

ce qui ramène le calcul à celui de l'intégrale d'une fraction rationnelle dont nous reparlerons plus loin. On utilise pour voir cela le changement de variables $u=\tan(\theta/2)$, soit $\theta=2\mathrm{Arctan}\,u$, vu précédemment. On se ramène ainsi au cadre des fractions rationnelles : si G est une primitive de la fraction rationnelle

$$u\longmapsto \frac{P\Big(\frac{1-u^2}{1+u^2},\frac{2u}{1+u^2}\Big)}{Q\Big(\frac{1-u^2}{1+u^2},\frac{2u}{1+u^2}\Big)}\,\frac{2}{1+u^2}\,,$$

alors

$$\theta\longmapsto F(\theta):=G(\tan(\theta/2))$$

est une primitive de

$$\theta\longmapsto \frac{P(\cos\theta,\sin\theta)}{Q(\cos\theta,\sin\theta)}\,.$$

Remarquons toutefois que les choses sont plus simples lorsque la fonction à intégrer est simplement

$$\theta\longmapsto P(\cos\theta,\sin\theta)\,;$$

dans ce cas, on peut profiter du fait que tout polynôme trigonométrique $P(\cos\theta,\sin\theta)$ à coefficients réels s'exprime grâce aux formules d'Euler comme une somme d'expressions du type

$$P(\cos\theta,\sin\sin\theta)=a_0+\sum_{j=1}^{N}(a_j\cos(k_j\theta)+b_j\sin(k_j\theta))$$

où les a_j sont des nombres réels et les k_j des entiers non nuls. Une primitive de la fonction donnée par $\theta\longmapsto P(\cos\theta,\sin\theta)$ est alors

$$F\ :\ \theta\longmapsto a_0\theta+\sum_{j=1}^{N}\frac{a_j\sin(k_j\theta)-b_j\cos(k_j\theta)}{k_j}$$

et le calcul exact de l'intégrale

$$\int_a^b P(\cos\theta, \sin\theta)\, d\theta = F(b) - F(a)$$

en résulte.

Les calculs s'avèrent plus simples si la fonction à intégrer s'exprime sous l'une des formes

$$H(\sin\theta)\, \cos\theta\,,\ H(\cos\theta)\, \sin\theta\,,$$

où H est une fraction rationnelle ; dans le premier cas, le changement de variable à effectuer est $u = \sin\theta$; dans le second cas, c'est $u = \cos\theta$ qui est le changement de variable adéquat.

2. Les expressions rationnelles en les lignes trigonométriques hyperboliques

Si I est un intervalle ouvert de $\mathbb{R}$ et si

$$x \longmapsto \frac{P(\cosh x, \sinh x)}{Q(\cosh x, \sinh x)}$$

est une fraction rationnelle en $\cosh x$ et $\sinh x$ telle que

$$P(\cosh x, \sinh x) \neq 0$$

sur I, on peut utiliser le changement de variables $e^x = u$ pour remarquer (en utilisant la formule de changement de variables) que, si $a, b \in I$,

$$\int_a^b \frac{P(\cosh x, \sinh x)}{Q(\cosh x, \sinh x)}\, dx = \int_{e^a}^{e^b} \frac{P(\frac{u+u^{-1}}{2}, \frac{u-u^{-1}}{2})}{Q(\frac{u+u^{-1}}{2}, \frac{u-u^{-1}}{2})}\, \frac{du}{u}$$

et que par conséquent, si G est une primitive de la fraction rationnelle

$$u \longmapsto \frac{P(\frac{u+u^{-1}}{2}, \frac{u-u^{-1}}{2})}{Q(\frac{u+u^{-1}}{2}, \frac{u-u^{-1}}{2})}\, \frac{1}{u}$$

sur $]0, +\infty[$, une primitive F de

$$x \longmapsto \frac{P(\cosh x, \sinh x)}{Q(\cosh x, \sinh x)}$$

sur I est donnée par

$$F\ :\ x \longmapsto G(e^x)\,.$$

3. Les fonctions dont une dérivée à un certain ordre est une fraction rationnelle.

Dans ce cas, c'est l'intégration par parties que l'on exploite en cherchant à faire apparaître la dérivée de la fonction à intégrer : par exemple, pour $x > 0$,

$$\int_1^x \log t\, dt = [t \log t]_1^x - \int_1^x t\, \frac{dt}{t} = x \log x - x$$

et, pour $x \in \mathbb{R}$,

$$\int_0^x \mathrm{Arctan}\, t\, dt = [t\mathrm{Arctan}\, t]_0^x - \int_0^x \frac{t\, dt}{1+t^2} = x\mathrm{Arctan}\, x - \frac{\log(1+x^2)}{2}\,.$$

4. Les fonctions du type $t \longmapsto t^n e^{\lambda t}$, où $n \in \mathbb{N}$, $\lambda \in \mathbb{R}^*$.

On utilise l'intégration par parties pour remarquer que, pour tout $x \in \mathbb{R}$,

$$\int_0^x e^{\lambda t} t^n dt = \Big[\frac{e^{\lambda t} t^n}{\lambda}\Big]_0^x - \frac{n}{\lambda}\int_0^x e^{\lambda t}\, t^{n-1}\, dt\,,$$

ce qui permet de faire "descendre" l'exposant n jusqu'à $n = 0$.

5. Les fonctions du type $t \longmapsto t^n \cos(\mu t)$ ou $t \longmapsto t^n \sin(\mu t)$, où $n \in \mathbb{N}$, $\mu \in \mathbb{R}^*$.

Même principe que précédemment, avec encore l'intégration par parties :

$$\begin{aligned}\int_0^x \cos(\mu t)\, t^n\, dt &= \Big[\frac{\sin(\mu t) t^n}{\mu}\Big]_0^x - \frac{n}{\mu}\int_0^x \sin(\mu t)\, t^{n-1}\, dt \\ \int_0^x \sin(\mu t)\, t^n\, dt &= -\Big[\frac{\cos(\mu t) t^n}{\mu}\Big]_0^x + \frac{n}{\mu}\int_0^x \cos(\mu t)\, t^{n-1}\, dt\,,\end{aligned}$$

ce qui permet encore de faire "chuter" l'exposant n jusqu'à $n = 0$. En fait, on pourrait calculer aussi de cette manière les intégrales des fonctions à valeurs complexes du type

$$t \longmapsto e^{(\lambda+i\mu)t} t^n = t^n e^{\lambda t}(\cos(\mu t) + i \sin(\mu t))$$

en remarquant (toujours *via* la formule d'intégration par parties) que

$$\int_0^x e^{(\lambda+i\mu)t} t^n dt = \Big[\frac{e^{(\lambda+i\mu)t} t^n}{\lambda+i\mu}\Big]_0^x - \frac{n}{\lambda+i\mu}\int_0^x e^{(\lambda+i\mu)t}\, t^{n-1}\, dt\,.$$

Isolant les parties réelles et imaginaires des deux membres dans ce type de formule, on peut ainsi calculer en faisant chuter l'exposant n jusqu'à $n = 0$ les primitives des fonctions

$$\begin{aligned} t &\longmapsto t^n e^{\lambda t}\cos(\mu t) \\ t &\longmapsto t^n e^{\lambda t}\sin(\mu t)\,.\end{aligned}$$

Ces calculs (qui englobent les exemples 4 et 5) s'avèrent très importants en mathématiques de l'ingénieur.

On envisagera à titre d'exercice le dernier cas suivant :

6. Intégrales de fonctions du type $F(x, \sqrt{ax^2+bx+c})$ (F fraction rationnelle en deux variables).

Introduites par N. Abel et liées au paramétrage des coniques, les intégrales de telles expressions se rencontrent couramment. On transforme le trinôme

$$ax^2 + bx + c = a(x + \frac{b}{2a})^2 + c - \frac{b^2}{4a} = a(X^2 - \frac{\Delta}{4a^2})\,,$$

où $X := x + \dfrac{b}{2a}$ et $\Delta := b^2 - 4ac$. On envisage ensuite tous les cas (le cas $\Delta = 0$ se ramène au cas des fractions rationnelles en x ; on peut en effet éliminer la prise de racine carrée car nous avons affaire sous le radical à un carré parfait) :

- $a > 0$ et $\Delta = -\delta^2 < 0$: on pose comme changement de variable

$$x = -\frac{b}{2a} + \frac{\delta}{2a} \times \sinh u\,,$$

  ce qui nous ramène au calcul de l'intégrale d'une expression du type étudiée dans l'exemple 2 ;
- $a > 0$ et $\Delta = \delta^2 > 0$: on pose comme changement de variable

$$x = -\frac{b}{2a} + \frac{\delta}{2a} \times (\pm\cosh u)\,,$$

  ce qui nous ramène encore au calcul de l'intégrale d'une expression étudiée dans l'exemple 2 ;

– $a < 0$ et $\Delta = \delta^2 > 0$: on pose comme changement de variable

$$x = -\frac{b}{2a} + \frac{\delta}{2|a|} \times \cos t$$

ou

$$x = -\frac{b}{2a} + \frac{\delta}{2|a|} \times \sin t\,,$$

ce qui nous ramène encore au calcul de l'intégrale d'une expression étudiée dans l'exemple 1.

3.8.4 Primitives de fractions rationnelles

Comme les fractions rationnelles

$$x \longmapsto \frac{P(x)}{Q(x)}$$

(avec P et Q à coefficients réels) sont les seules fonctions réelles auxquelles le calcul machine nous donne réellement accès (car elles sont construites à partir d'opérations algébriques), il est très intéressant que l'on sache toujours en exprimer explicitement une primitive (même si l'expression de cette primitive peut s'avérer compliquée et devoir passer d'abord par la recherche des pôles de la fraction rationnelle dans le plan complexe, ce que malheureusement seul le calcul numérique permet d'envisager) ; on donnera ici (au moins si Q est de degré au plus 2) le principe de construction de cette primitive.

On remarque tout d'abord que la division euclidienne de P par Q (on pourra admettre de quoi il s'agit et prendre directement la forme R/Q avec $\deg R < \deg Q$ pour la fraction rationnelle) nous permet d'écrire

$$\frac{P(x)}{Q(x)} = \sum_{k=0}^{N} a_k x^k + \frac{R(x)}{Q(x)}\,,$$

avec $\deg R < \deg Q$; comme une primitive de la fonction polynomiale

$$x \longmapsto \sum_{k=0}^{N} a_k x^k$$

est

$$x \longmapsto \sum_{k=0}^{N} a_k \frac{x^{k+1}}{k+1}\,,$$

on est ramené à chercher une primitive de R/Q (dans $\mathbb{R}$ privé des zéros éventuels de Q).

Si $\deg Q = 1$, alors $\deg R \leq 0$ et il s'agit de trouver une primitive de

$$x \longmapsto \frac{1}{ax+b}$$

($a \neq 0$) sur $\mathbb{R} \setminus \{-b/a\}$; une telle primitive est, on l'a vu, la fonction

$$x \longmapsto \frac{1}{a} \log |ax+b|$$

et nous fait sortir du monde des fonctions rationnelles.

Si $\deg Q = 2$, on est ramené à chercher une primitive pour une fraction rationnelle

$$x \longmapsto \frac{\alpha x + \beta}{ax^2 + bx + c}\,.$$

Trois cas sont à envisager :

– Si le discriminant du trinôme aX^2+bX+c est strictement positif, ce trinôme se factorise en

$$aX^2+bX+c = a(X-x_1)(X-x_2)$$

avec x_1, x_2 réels distincts. La fraction R/Q s'écrit sous la forme

$$\frac{\alpha x+\beta}{a(x-x_1)(x-x_2)} = \frac{u_1}{x-x_1}+\frac{u_2}{x-x_2}\,,$$

où u_1 et u_2 sont deux nombres réels (pour calculer u_1 multiplier R/Q par $x-u_1$ et faire $u_1=0$, même principe pour calculer u_2) ; une primitive de $x \to R(x)/Q(x)$ dans $\mathbb{R}\backslash\{x_1,x_2\}$ est la fonction

$$x \longmapsto u_1 \log|x-x_1| + u_2 \log|x-x_2|\,.$$

– Si le discriminant du trinôme aX^2+bX+c est nul, ce trinôme s'écrit $a(x-x_0)^2$ et la fraction R/Q s'écrit sous la forme

$$\frac{\alpha x+\beta}{a(x-x_0)^2} = \frac{\alpha}{a(x-x_0)}+\frac{\alpha x_0+\beta}{a(x-x_0)^2}$$

et une primitive sur $\mathbb{R}\setminus\{0\}$ est

$$x \longmapsto \frac{\alpha}{a}\,\log|x-x_0| - \frac{\alpha x_0+\beta}{a}\,\frac{1}{x-x_0}\,.$$

– Si $\Delta=-\delta^2<0$, le trinôme aX^2+bX+c s'écrit $a\Big((X+(b/2a))^2+\dfrac{\delta^2}{4a^2}\Big)$ et la fraction rationnelle s'écrit sous la forme

$$\frac{\alpha x+\beta}{ax^2+bx+c} = u\frac{x+(b/2a)}{(x+b/2a)^2+\delta^2/4a^2}+v\frac{1}{(x+b/2a)^2+\delta^2/4a^2}\,,$$

où u et v sont deux coefficients réels que l'on calculera. Une primitive sur $\mathbb{R}$ est alors la fonction

$$x \longmapsto \frac{u}{2}\log((x+b/2a)^2+\delta^2/4a^2)+\frac{2av}{\delta}\mathrm{Arctan}\left[2a\frac{\Big(x+(b/2a)\Big)}{\delta}\right].$$

On voit surgir ici la fonction Arctan qui d'ailleurs est encore un avatar de la fonction logarithme.

3.9 Équations différentielles

Les phénomènes physiques y fonction d'un paramètre (qui est en général du temps) obéissent à des contraintes qui imposent une relation entre les dérivées successives de y (supposées exister sur un intervalle temporel I). Une telle relation, par exemple

$$F(t,y(t),y'(t)) \equiv 0$$

(seuls t, y et y' y sont ici impliqués) ou

$$F(t,y(t),y'(t),y''(t)) \equiv 0$$

(cette fois, le temps, y et ses deux premières dérivées sont impliqués) est dite *équation différentielle*, dans le premier cas *d'ordre* 1, dans le second cas *d'ordre 2*.

On s'intéressera ici aux équations différentielles dites linéaires, d'ordre 1 et de plus résolubles en y', c'est-à-dire du type

$$y'(t) = a(t)y(t) + b(t)\,,$$

ou d'ordre 2 et résolubles en y'' (dérivée de y'), c'est-à-dire du type

$$y''(t) = a(t)y'(t) + b(t)y(t) + c(t).$$

Souvent y' est interprété comme la vitesse, y'' comme l'accélération. Souvent aussi en physique ou en mécanique, y est une fonction à deux composantes et il peut être utile de penser que y est une fonction définie sur un intervalle I de $\mathbb{R}$ et à valeurs dans $\mathbb{C}$.

3.9.1 Équations linéaires du premier ordre et problème de Cauchy associé

Soit I un intervalle de $\mathbb{R}$, a et b deux fonctions continues sur I, t_0 un point de I, et y_0 un nombre réel.

Nous nous intéressons ici à la recherche des courbes intégrales de l'équation différentielle

$$y'(t) = a(t)y(t) + b(t) \qquad (*)$$

passant par le point (t_0, y_0), c'est à dire des graphes des solutions y de l'équation $y'(t) = a(t)y(t) + b(t)$ définies au voisinage de t_0 et telles que $y(t_0) = y_0$ (ce qui signifie que le graphe de la solution passe par le point (t_0, y_0) du plan).

Avant d'attaquer ce problème, essayons d'examiner numériquement les choses et de "simuler" l'évolution d'un phénomène physique y (à partir de l'instant $t = t_0$) lorsque ce phénomène est régi par une équation différentielle

$$y' = a(t)y + b(t)\,,$$

où a et b sont des fonctions continues ; pour cela, on décide d'un pas d'échantillonnage τ du temps et l'on s'intéresse à la suite de terme général

$$u_{\tau,n} := y(t_0 + n\tau)\,,\ n = 0, 1, ...$$

La dérivée au point $t_0 + n\tau$ peut être approchée par

$$y'(t_0 + n\tau) \simeq \frac{y(t_0 + (n+1)\tau) - y(t_0 + n\tau)}{\tau} = \frac{u_{\tau,n+1} - u_{\tau,n}}{\tau}$$

et l'on voit ainsi que la suite $(u_{\tau,n})_n$ est régie (de manière approchée) par la relation de récurrence :

$$u_{\tau,n+1} = u_{\tau,n} + \tau(a(t_0 + n\tau)u_{\tau,n} + b(t_0 + n\tau))\,.$$

On peut calculer les termes de cette suite récurrente et, en affichant sur un graphique les points $(t_0 + n\tau, u_{\tau,n})$, $n \in \mathbb{N}$, on obtient une approximation de l'évolution du phénomène physique (d'autant meilleure que τ est petit). Cette méthode est la méthode d'Euler ; on l'a déjà rencontré à propos de l'exponentielle (solution de l'équation différentielle très simple $y' = y$) et c'est le prototype de toutes les méthodes numériques utilisées dans la modélisation des phénomènes physiques régis par des équations différentielles. Il semble se dégager en tout cas le fait qu'il n'y ait, pour des phénomènes régis par une équation différentielle du type (*) qu'une seule courbe intégrale passant par (t_0, y_0).

Revenons donc au problème de la recherche des courbes intégrales de l'équation (*) (passant par (t_0, y_0)) du point de vue théorique. La réponse à cette question est fournie par le théorème de Cauchy (Augustin

Cauchy, 1789-1857, est un mathématicien français du XIX-ème siècle à l'origine du développement extensif de l'analyse mathématique à partir de préoccupations d'origine physique) :

Théorème de Cauchy pour les équations linéaires d'ordre 1. *Sous les hypothèses précédentes, il existe une unique courbe intégrale*

$$t \in I \longmapsto (t, y(t))$$

de l'équation différentielle (*) passant par le point (t_0, y_0)*. Cette solution est donnée par la formule*

$$y(t) = \left(y_0 + \int_{t_0}^{t} b(u) \exp\Big(- \int_{t_0}^{u} a(v)\, dv \Big) du \right) \exp\Big(\int_{t_0}^{t} a(u) du \Big) . \qquad (**)$$

Preuve. Avant de montrer qu'il existe une solution, on va prouver qu'elle est unique. Supposons qu'il y ait deux fonctions y_1, y_2 dérivables sur I et telles que

$$\forall t \in I\,,\ \Big(y_1'(t) = a(t) y_1(t) + b(t) \Big) \wedge \Big(y_2'(t) = a(t) y_2(t) + b(t) \Big)$$

avec de plus $y_1(t_0) = y_2(t_0)$. Considérons la fonction $z = y_1 - y_2$; cette fonction est dérivable sur I, nulle en t_0, et telle que

$$z'(t) = a(t) z(t)\,.$$

Soit A la primitive de a sur I s'annulant en t_0, soit

$$A\ :\ t \in I \longmapsto \int_{t_0}^{t} a(u)\, du\,.$$

Posons, pour tout $t \in I$,

$$Z(t) := z(t) \exp(-A(t))\,;$$

La fonction Z est dérivable sur I et vérifie

$$Z'(t) = (z'(t) - a(t) z(t)) \exp(-A(t)) = 0\,,\ \forall t \in I\,.$$

D'après la proposition 3.12, la fonction Z est constante ; comme $Z(t_0) = z(t_0) = 0$, on a $Z(t) = 0$ pour tout t dans I, donc $z \equiv 0$. L'unicité de notre solution est ainsi prouvée.

Pour achever la preuve du théorème de Cauchy, on pourrait se contenter de vérifier que la fonction y définie par $(**)$ satisfait l'équation différentielle $y'(t) = a(t) y(t) + b(t)$ sur I en même temps que la "condition initiale" $y(t_0) = y_0$. C'est un peu pénible, mais cela se fait (faites le en exercice).

Plutôt que de faire cette vérification, on va tenter de résoudre $(*)$ avec la condition initiale $y(t_0) = y_0$ en trois temps :

1. D'abord en cherchant toutes les solutions de l'équation différentielle dite *homogène* ou *sans second membre*

$$y'(t) = a(t) y(t)\,,\ \forall t \in I\,. \qquad (\dagger)$$

Si l'on pose $Y(t) = y(t) \exp(-A(t))$, chercher y solution de $(\dagger)$ sur I revient à chercher Y solution de

$$(Y'(t) + a(t) Y(t)) \exp(A(t)) = a(t) Y(t) \exp(A(t))\,,\ \forall t \in I\,,$$

soit

$$Y'(t) = 0\,,\ \forall t \in I\,;$$

d'après la proposition 3.12, dire que y est solution de (†) équivaut à dire que Y est une fonction constante ; la solution générale de l'équation (†) sur I est donc

$$y(t) = C\exp(A(t))\,,$$

où C est une constante complexe arbitraire. Les solutions de l'équation (†) sur I dépendent donc d'un degré de liberté (C). Ce sont les multiples de la fonction

$$t \in I \longmapsto \exp(A(t))\,.$$

L'ensemble $\mathcal{E}$ des solutions y de l'équation (†) sur I, muni de l'addition et de la multiplication externe

$$(\lambda, f) \in \mathbb{R} \times \mathcal{E} \longmapsto \lambda \cdot y = \lambda y\,,$$

est un $\mathbb{R}$-espace vectoriel, engendré par un élément de base, à savoir la fonction $t \longmapsto \exp(A(t))$.

2. Ensuite, en cherchant une solution particulière de l'équation (∗) sous la forme

$$y(t) = C(t)\exp(A(t))\,,$$

où C est une fonction dérivable de t ; cette méthode est dite *méthode de variation de la constante* et on la retrouve classiquement lorsqu'il s'agit de trouver, une fois résolue l'équation sans second membre, une solution particulière de l'équation avec second membre. écrire que y est solution de (∗) donne ici

$$(C'(t) + a(t)C(t))\exp(A(t)) = C(t)\exp(A(t)) + b(t)\,,$$

soit

$$C'(t) = b(t)\exp(-A(t))\,,$$

dont une solution est la fonction

$$y_{\text{part}} \ : t \in I \longmapsto \exp(A(t))\int_{t_0}^{t} b(u)\exp(-A(u))\,du\,.$$

3. En remarquant que la solution générale de (∗) sur I s'écrit

$$y(t) = y_{\text{part}}(t) + C\exp(A(t)) = \exp(A(t)) \times \left(\int_{t_0}^{t} b(u)\exp(-A(u))\,du + C\right),$$

puis en ajustant le degré de liberté que constitue la constante C pour que la solution $y = y_C$ de (∗) ainsi construite vaille y_0 en t_0. Ici l'ajustement correspond à prendre $C = y_0$ car $A(t_0) = 0$.

On trouve bien la solution (∗∗) proposée. □

3.9.2 Un exemple de problème de Cauchy du premier ordre non linéaire se ramenant au cas linéaire

La résolution du problème de Cauchy avec données initiales (t_0, y_0) pour certaines équations du premier ordre non linéaires peut dans certains cas très particuliers se ramener à la résolution d'un problème de Cauchy pour une équation du premier ordre linéaire. Tel est le cas pour la résolution de

$$y'(t) = a(t)y(t) + b(t)y^{\alpha}(t)$$

avec $\alpha \in \mathbb{R}$, $\alpha \neq 0, 1$, sous la condition initiale

$$y(t_0) = y_0 > 0\,.$$

Ce type d'équation, classique en physique ou en mécanique, est dite *équation de Bernoulli*, ce en relation avec le mathématicien et physicien suisse Jacob Bernoulli (1654-1705) ; cette équation (lorsque $\alpha = n \in \mathbb{N} \setminus \{0, 1\}$) apparaît dans la recherche des courbes de "descente en temps constant", c'est-à-dire des courbes dites "isochrones" (par exemple celles, dites de Leibniz, le long desquelles un objet descendant sous les lois de la gravité descendra avec une composante de vitesse verticale constante, se reporter à ma page web pour d'autres exemples, tels les courbes "brachistochrones" ou les courbes isochrones de Huygens) ; de telles courbes jouent un rôle important par exemple en horlogerie de précision (d'où l'influence de l'école suisse au travers de la dynastie des Bernoulli !).

Si y existe, y reste strictement positive dans un intervalle ouvert J assez petit contenant t_0 et inclus dans I et l'on a, dans cet intervalle ouvert J,

$$\frac{y'(t)}{(y(t))^\alpha} = a(t)(y(t))^{1-\alpha} + b(t)\,,\ \forall t \in J\,. \qquad (***)$$

En posant, pour tout $t \in J$,

$$z(t) := (y(t))^{1-\alpha}\,,$$

on voit que $(***)$ équivaut à

$$z'(t) = (1-\alpha)a(t)\,z(t) + (1-\alpha)b(t)\,.$$

Le problème de la résolution de l'équation $y'(t) = a(t)y(t) + b(t)y^\alpha(t)$ avec condition initiale $y(t_0) = y_0$ se ramène donc à la résolution au voisinage de t_0 de l'équation linéaire

$$z'(t) = (1-\alpha)a(t)z(t) + (1-\alpha)b(t)$$

avec condition initiale $z(t_0) = y_0^{1-\alpha} = z_0$. Une fois z trouvée, on en déduit y par

$$y(t) = (z(t))^{\frac{1}{1-\alpha}}$$

et l'intervalle de vie de y est le plus grand intervalle ouvert sur lequel la solution $z(t)$ définie par la formule

$$z(t) = \left(z_0 + \int_{t_0}^{t} (1-\alpha)b(u) \exp\Big(-\int_{t_0}^{u} (1-\alpha)a(v)\,dv\Big)\,du\right) \exp\Big(\int_{t_0}^{t} (1-\alpha)a(u)du\Big)$$

reste strictement positive.

Remarque. Dans nombre de problèmes de physique, les fonctions a et b intervenant dans le problème de Cauchy $y' = ay + b$, $y(t_0) = y_0$, ainsi que la donnée initiale y_0, sont complexes. La résolution s'effectue de la même manière mais il faut lire la formule $(**)$ comme une formule où les diverses quantités impliquées sont complexes (on a vu ce que signifiait l'intégrale sur un segment $[a, b]$ d'une fonction à valeurs complexes dont les parties réelle et imaginaire sont des fonctions continues sur $[a, b]$).

3.9.3 Les équations différentielles du second ordre à coefficients constants

Une fonction f définie sur un intervalle ouvert J de $\mathbb{R}$ et à valeurs réelles est dite *deux fois dérivable* sur I (ou encore *admet une dérivée seconde* sur J) si f est dérivable sur I et la fonction f' également ; on note alors $(f')' = f''$.

Soit I un intervalle ouvert de $\mathbb{R}$, t_0 un point de I, et a, b, c trois fonctions continues sur I ; chercher une fonction y définie et deux fois dérivable dans un intervalle ouvert $J \subset I$ contenant t_0 et telle que

$$\forall t \in J\,,\ y''(t) = a(t)y'(t) + b(t)y(t) + c(t) \qquad (\dagger\dagger)$$

s'appelle *résoudre l'équation différentielle linéaire du second ordre* $(\dagger\dagger)$ près de t_0. Chercher les solutions y de cette équation obéissant aux deux conditions initiales

$$y(t_0) = y_0\,,\ y'(t_0) = v_0$$

s'appelle *résoudre le problème de Cauchy linéaire du second ordre*

$$\begin{array}{rcl} y''(t) & = & a(t)y'(t) + b(t)y(t) + c(t) \text{ près de } t_0 \\ y(t_0) & = & y_0 \\ y'(t_0) & = & v_0\,, \end{array}$$

les conditions

$$y(t_0) = y_0\,,\ y'(t_0) = v_0$$

étant appelées conditions initiales.

On peut montrer que le problème de Cauchy posé pour une équation différentielle linéaire

$$y''(t) = a(t)y'(t) + b(t)y(t) + c(t)\,,$$

avec a, b, c continues sur un intervalle I, et conditions initiales $y(t_0) = y_0$, $y'(t_0) = v_0$, admet une unique solution y définie et deux fois dérivable sur I tout entier.

C'est encore la méthode d'Euler qui permet de résoudre numériquement ce problème : si τ est un pas d'échantillonnage du temps fixé *a priori* et si $u_{n,\tau}$ est une approximation de $y(t_0 + n\tau)$, la dérivée au point $t_0 + n\tau$ peut être approchée par

$$y'(t_0 + n\tau) \simeq \frac{y(t_0 + (n+1)\tau) - y(t_0 + n\tau)}{\tau} = \frac{u_{\tau,n+1} - u_{\tau,n}}{\tau}$$

tandis que la dérivée seconde peut être approchée, elle, par

$$\begin{array}{rcl} y''(t_0 + n\tau) \simeq \dfrac{y'(t_0 + (n+1)\tau) - y'(t_0 + n\tau)}{\tau} & \simeq & \dfrac{\dfrac{u_{\tau,n+2} - u_{\tau,n+1}}{\tau} - \dfrac{u_{\tau,n+1} - u_{\tau,n}}{\tau}}{\tau} \\ & \simeq & \dfrac{u_{\tau,n+2} - 2u_{\tau,n+1} + u_{\tau,n}}{\tau^2} \end{array}$$

et l'on voit ainsi que la suite $(u_{\tau,n})_n$ est régie (de manière approchée) par la relation de récurrence :

$$u_{\tau,n+2} = u_{\tau,n}\Big(\tau^2 b(t_0 + n\tau) - \tau a(t_0 + n\tau) - 1\Big) + u_{\tau,n+1}\Big(\tau a(t_0 + n\tau) + 2\Big) + c(t_0 + n\tau)\tau^2$$

avec les deux conditions initiales :

$$\begin{array}{rcl} u_{\tau,0} & = & y_0 \\ u_{\tau,1} & = & y_0 + \tau v_0\,. \end{array}$$

Le calcul des $u_{\tau,n}$, $n \geq 0$, à partir des conditions initiales est possible de proche en proche et, en affichant sur un graphique les points $(t_0 + n\tau, u_{\tau,n})$, $n \in \mathbb{N}$, on obtient une approximation de l'évolution

du phénomène physique régi par l'équation différentielle linéaire du second ordre (††) (d'autant meilleure que τ est petit). Cette méthode d'Euler nous permet aussi de deviner que par un point (t_0, y_0) (avec $t_0 \in I$) passe une seule courbe intégrale ayant pour tangente au point (t_0, y_0) la droite de pente imposée v_0.

Nous allons montrer ce résultat de manière constructive (c'est évidemment cela qui est important !) lorsque $t \to a(t), t \to b(t)$ sont des fonctions constantes sur I. Du fait des motivations physiques, nous nous placerons à la fois dans le cadre (général) complexe et dans le cadre (particulier) réel. On rencontre de telles équations différentielles par exemple en électronique ou en mécanique. Par exemple, si l'on a affaire à une cellule électrique comme celle de la figure ci-dessous

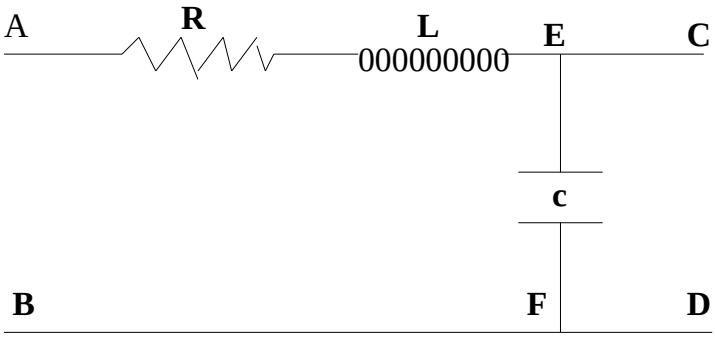

FIGURE 3.21 – Une cellule électrique du second ordre

on voit (le condensateur de capacité c étant supposé non chargé à l'instant $t = 0$), que, si i désigne l'intensité (comptée algébriquement) du courant circulant de A vers C , la loi d'Ohm implique (si R désigne la valeur numérique de la résistance, L l'inductance de la bobine)

$$u_{AB}(t) = R\,i(t) + L\,i'(t) + \frac{1}{c}\int_0^t i(\xi)d\xi = R\,i(t) + L\,i'(t) + U_{CD}(t)\,,\ t \geq 0\,,$$

et

$$u_{CD}(t) = \frac{1}{c}\int_0^t i(\xi)d\xi\,,\ t \geq 0$$

si $U_{AB}(t)$ désigne (à l'instant t) la différence de potentiel entre A et B et $U_{CD}(t)$ celle entre C et D, soit $i(t) = cU'_{CD}(t)$; en regroupant ces relations, on trouve immédiatement, en posant $U_{CD} = y$ et $U_{AB} = f$:

$$Lcy''(t) + Rcy'(t) + y(t) = f(t)\,,$$

soit

$$y''(t) = -\frac{R}{L}\,y'(t) - \frac{1}{Lc}\,y(t) + \frac{f(t)}{Lc}\,,$$

ce qui montre que y est solution d'une équation différentielle du second ordre à coefficients constants.

Théorème de Cauchy pour les équations linéaires du second ordre à coefficients constants. *Soit I un intervalle ouvert de $\mathbb{R}$, $t_0 \in I$, a, b deux nombres complexes, c une fonction continue de I dans $\mathbb{C}$ (la partie réelle et la partie imaginaire de c sont des fonctions continues dans I) ; soit y_0 et v_0 deux nombres complexes. le problème de Cauchy linéaire du second ordre*

$$\begin{aligned} y''(t) &= ay'(t) + by(t) + c(t) \\ y(t_0) &= y_0 \\ y'(t_0) &= v_0 \end{aligned}$$

admet une et une seule solution $y : I \longmapsto \mathbb{C}$ *à valeurs complexes définie sur* I *tout entier et deux fois dérivable sur* I *(ses parties réelle et imaginaire sont deux fois dérivables). Dans le cas particulier où* a, b, y_0, v_0 *sont des nombres réels et où* c *est une fonction de* $I \longmapsto \mathbb{R}$*, le problème de Cauchy linéaire du second ordre ci-dessus admet une unique solution réelle* $y : I \longmapsto \mathbb{R}$*.*

Preuve.

I. Preuve de l'unicité de la solution.

On se place dans le cadre (le plus général) où toutes les données sont *a priori* complexes. Montrons tout d'abord, comme dans le cas du premier ordre, que si une solution $y : I \longmapsto \mathbb{C}$ existe sur I tout entier, elle est unique. Pour cela, prenons deux solutions y_1 et y_2 de notre problème de Cauchy. La différence $z = y_1 - y_2$ est solution du problème de Cauchy

$$\begin{aligned} y''(t) - ay'(t) - by(t) &= 0 \ \forall t \in I \\ y(t_0) &= 0 \\ y'(t_0) &= 0\,. \end{aligned}$$

Le trinôme $X^2 - aX - b$ admet deux racines complexes w_1 et w_2 et on peut le factoriser en

$$X^2 - aX - b = (X - w_1)(X - w_2)\,.$$

Si D désigne l'opération de prise de dérivée et D^2 celle de prise de dérivée deux fois, on a

$$D^2 - aD - b = (D - w_1\mathrm{Id}) \circ (D - w_2\mathrm{Id})\,.$$

Si l'on pose

$$Z = (D - w_2\mathrm{Id})[z] = z' - w_2 z\,,$$

on voit que

$$(D - w_1\mathrm{Id})[Z] = Z' - w_1 Z = 0\,.$$

La fonction Z est donc solution du problème de Cauchy linéaire du premier ordre

$$\begin{aligned} Z'(t) - w_1 Z(t) &= 0\,,\ \forall t \in I \\ Z(t_0) &= 0\,. \end{aligned}$$

En utilisant ce que nous avons fait dans le cas du problème de Cauchy pour des équations linéaires du premier ordre, on voit que

$$Z(t) = 0\,,\ \forall t \in I$$

(la fonction nulle est aussi solution de ce problème !). La fonction z est donc solution du problème de Cauchy linéaire du premier ordre

$$\begin{aligned} z'(t) - w_2 z(t) &= 0\,,\ \forall t \in I \\ z(t_0) &= 0\,; \end{aligned}$$

on en déduit (toujours en utilisant ce que nous avons fait dans le cas du problème de Cauchy pour des équations linéaires du premier ordre) que $z = 0$ sur I, ce qui prouve l'unicité de notre solution au problème de Cauchy linéaire du second ordre posé.

II. Résolution de $y'' = ay' + by$

Avant de construire une solution de l'équation $y''(t) = ay'(t) + by(t) + c(t)$, nous remarquons (comme dans le cas du problème de Cauchy du premier ordre) que ceci est plus facile si c est la fonction

nulle. Traitons d'abord ce cas particulier important (on dit que l'on s'intéresse à l'équation *sans second membre*) tout d'abord dans le cas général où a et b sont des constantes complexes quelconques, puis dans le cas particulier très important où a et b sont des nombres réels.

IIA. *Résolution de l'équation homogène $y'' = ay' + b$ dans le cas où a et b sont des constantes complexes arbitraires*

Notons w_1 et w_2 les deux racines (complexes) de l'équation caractéristique $X^2 - aX - b = 0$. Nous devons distinguer les deux sous-cas $w_1 \neq w_2$ et $w_1 = w_2$.

– Dans le premier sous-cas, les deux fonctions à valeurs complexes

$$\begin{aligned} t \in I &\longmapsto \exp(w_1 t) \\ t \in I &\longmapsto \exp(w_2 t) \end{aligned}$$

sont solutions de $y''(t) - ay'(t) - by(t) = 0$; en effet

$$[D^2 - aD - b][e^{w_j(\cdot)}] = (w_j^2 - aw_j - b)\, e^{w_j(\cdot)} = 0\,,\ j = 1, 2\,.$$

Toutes les fonctions de la forme

$$t \longmapsto C_1 e^{w_1 t} + C_2 e^{w_2 t}\,,\ C_1, C_2 \in \mathbb{C}\,,$$

sont solutions de l'équation différentielle du second ordre linéaire sans second membre

$$y''(t) - ay'(t) - by(t) = 0\,.$$

On a là une famille de solutions dépendant de deux degrés de liberté (ces degrés de liberté ici complexes étant matérialisés par les deux constantes complexes C_1 et C_2). Si nous imposons les conditions initiales

$$\begin{aligned} y(t_0) &= y_0 \\ y'(t_0) &= v_0\,, \end{aligned}$$

où y_0 et v_0 sont deux nombres complexes, les constantes C_1 et C_2 sont parfaitement déterminées par la résolution du système linéaire de deux équations à deux inconnues

$$\begin{aligned} C_1 e^{w_1 t_0} + C_2 e^{w_2 t_0} &= y_0 \\ C_2 w_1 e^{w_1 t_0} + C_2 w_2 e^{w_2 t_0} &= v_0\,; \end{aligned}$$

ce système admet une solution unique (C_1, C_2) car $(w_2 - w_1)\exp(w_1 t_0 + w_2 t_0)$ (qui est le déterminant du système) est non nul.

– Dans le second sous-cas, on constate que

$$t \longmapsto e^{w_1 t}$$

et

$$t \longmapsto t e^{w_1 t}$$

sont solutions de de $y''(t) - ay'(t) - by(t) = 0$. Toutes les fonctions de la forme

$$t \longmapsto (C_1 + tC_2) e^{w_1 t}\,,\ C_1, C_2 \in \mathbb{C}\,,$$

sont solutions de l'équation différentielle du second ordre linéaire sans second membre

$$y''(t) - ay'(t) - by(t) = 0\,.$$

On a là encore une famille de solutions dépendant de deux degrés de liberté (ces degrés de liberté ici complexes étant matérialisés par les deux constantes complexes C_1 et C_2). Si nous imposons les conditions initiales

$$\begin{aligned} y(t_0) &= y_0 \\ y'(t_0) &= v_0\,, \end{aligned}$$

où y_0 et v_0 sont deux nombres complexes, les constantes C_1 et C_2 sont parfaitement déterminées par la résolution du système linéaire de deux équations à deux inconnues

$$\begin{aligned} (C_1 + t_0 C_2)e^{w_1 t_0} &= y_0 \\ (C_1 w_1 + C_2(t_0 w_1 + 1))e^{w_1 t_0} &= v_0\,; \end{aligned}$$

ce système admet encore une solution unique (C_1, C_2) car

$$(1 + t_0 w_1 - t_0 w_1)\exp(w_1 t_0 + w_2 t_0) = \exp(w_1 t_0 + w_2 t_0)$$

(qui est le déterminant du système) est non nul.

L'ensemble des fonctions deux fois dérivables sur I et solutions de l'équation sans second membre $y'' - ay' - by = 0$ est donc (dans les deux sous-cas considérés) un $\mathbb{C}$-espace vectoriel de dimension 2 (les solutions dépendent de deux degrés de liberté exactement).

IIB. *Résolution de l'équation homogène $y'' = ay' + by$ dans le cas particulier où a et b sont des constantes réelles arbitraires*

Dans ce cas, on cherche les solutions réelles $y\ : \mathbb{R} \longmapsto \mathbb{R}$ de l'équation $y'' = ay' + by$.

Si a et b sont réels, le trinôme du second degré $X^2 - aX - b$ (dit *équation caractéristique* de l'équation sans second membre $y'' - ay' - by = 0$) entre dans l'une des trois classes suivantes :

- Si $a^2 + 4b > 0$, le trinôme $X^2 - aX - b$ a deux racines réelles distinctes λ_1 et λ_2 ; dans ce cas (c'est le premier sous cas du cas général étudié précédemment, mais avec cette fois de plus w_1 et w_2 réels), les solutions réelles de l'équation $y'' = ay' + by$ sont les fonctions

$$t \longmapsto y(t) = C_1 \exp(\lambda_1 t) + C_2 \exp(\lambda_2 t)\,,$$

où C_1 et C_2 sont deux constantes réelles ; les solutions réelles de $y'' = ay' + by$ dépendent de deux degrés de liberté réels cette fois (C_1 et C_2). Si $\inf(\lambda_j) > 0$, on constate une explosion exponentielle de toutes les solutions lorsque t tend vers $+\infty$; au contraire, si $\sup(\lambda_j) < 0$, on constate un phénomène d'extinction (toutes les solutions tendent vers 0 lorsque t tend vers $+\infty$).
- Si $a^2 + 4b = 0$, le trinôme $X^2 - aX - b$ a une racine double λ réelle ; dans ce cas (c'est le second sous-cas du cas général étudié précédemment, mais avec cette fois $w_1 = w_2$ réelle), les solutions réelles de l'équation $y'' = ay' + by$ sont les fonctions

$$t \longmapsto y(t) = C_1 \exp(\lambda t) + C_2 t \exp(\lambda t)\,,$$

où C_1 et C_2 sont deux constantes réelles ; les solutions réelles de $y'' = ay' + by$ dépendent encore de deux degrés de liberté réels cette fois (C_1 et C_2). On constate un phénomène d'explosion exponentielle pour toutes les solutions lorsque $\lambda = a/2 > 0$, un phénomène d'extinction lorsque $\lambda = a/2 < 0$.
- Si $a^2 + 4b < 0$, le trinôme $X^2 - aX - b$ a deux racines complexes conjuguées $w_1 = \lambda + i\omega$ et $w_2 = \lambda - i\omega$, où $\lambda = a/2$ et $\omega = \sqrt{|a^2 + 4b|}/2$ sont des nombres réels ; dans ce cas, on voit (en

cherchant les solutions complexes comme dans le cas général puis en en prenant la partie réelle) que les solutions réelles de l'équation $y'' = ay' + by$ sont les fonctions

$$t \longmapsto y(t) = \exp(\lambda t) \times (C_1 \cos(\omega t) + C_2 \sin(\omega t)),$$

où C_1 et C_2 sont deux constantes réelles ; les solutions réelles de $y'' = ay' + by$ dépendent encore de deux degrés de liberté réels cette fois (C_1 et C_2). Si $\lambda > 0$, le phénomène physique "explose" pour t tendant vers $+\infty$ (ce qui correspond à un régime d'oscillation instable car les oscillations sont amplifiées) ; si $\lambda < 0$, on constate que $y(t)$ tend vers 0 lorsque t tend vers $+\infty$ (on a affaire à un régime d'oscillation amorti) ; le cas $\lambda = 0$ correspond, lui, à un régime d'oscillation stable. Sur les deux figures ci-dessous, on a représenté ce qui se passait (dans le cas $a^2 + 4b < 0$) si $\lambda = a/2 > 0$ (oscillations amplifiées) et $\lambda = a/2 < 0$ (oscillations amorties) :

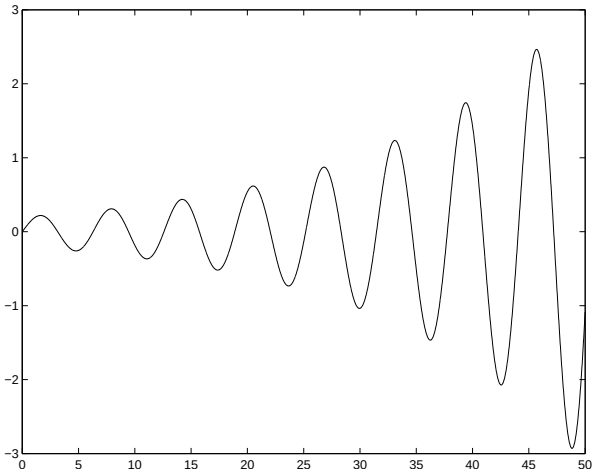

FIGURE 3.22 – Oscillations amplifiées : $a = 0.1$, $b = -1$, $t_0 = 0$, $y_0 = 0$, $y'_0 = 0.2$

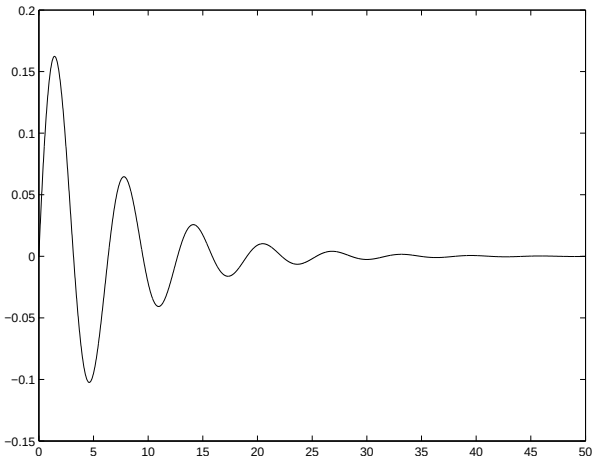

FIGURE 3.23 – Oscillations amorties : $a = -0.3$, $b = -1$, $t_0 = 0$, $y_0 = 0$, $y'_0 = 0.2$

III. Résolution du problème de Cauchy.

IIIA. *Résolution du problème de Cauchy dans le cas général ($a, b, y_0, v_0 \in \mathbb{C}$, $c : I \longmapsto \mathbb{C}$)*

Pour construire la solution (complexe) à notre problème de Cauchy dans ce cas général, nous utilisons les deux solutions indépendantes (pour l'équation homogène) y_1 et y_2 que nous avons construit au **IIA** ($y_1(t) = e^{w_1 t}$, $y_2(t) = e^{w_2 t}$ dans le premier sous-cas, $y_1(t) = e^{w_1 t}$, $y_2(t) = te^{w_1 t}$ dans le second sous-cas), en remarquant que dans ces deux sous-cas, la fonction

$$W : t \longmapsto y_1(t)y_2'(t) - y_2(t)y_1'(t)$$

ne s'annule pas sur I. Cette fonction s'appelle le *wronskien* de la paire de solutions (y_1, y_2) de l'équation linéaire sans second membre $y'' - ay' - by = 0$.

Nous cherchons ensuite une solution particulière de l'équation $y''(t) - ay'(t) - by(t) = c(t)$ de la forme

$$y(t) = C_1(t)y_1(t) + C_2(t)y_2(t)\,,$$

les fonctions C_1 et C_2 étant supposées être deux fois dérivables sur I et satisfaire au système

$$\begin{aligned} C_1'(t)y_1(t) + C_2'(t)y_2(t) &= 0 \quad \forall t \in I \\ C_1'(t)y_1'(t) + C_2'(t)y_2'(t) &= c(t) \quad \forall t \in I \end{aligned}$$

(c'est encore une méthode de variation des constantes comme pour la recherche d'une solution particulière de l'équation du premier ordre $y'(t) = a(t)y(t) + b(t)$). En reportant et en utilisant le fait que y_1 et y_2 sont solutions de $y'' - ay' - by = 0$, on voit qu'une telle fonction y est solution de l'équation de $y''(t) - ay'(t) - by(t) = c(t)$ sur l'intervalle I. On trouve ainsi comme solution particulière de l'équation avec second membre $y''(t) = ay'(t) + by(t) + c(t)$ la fonction

$$t \longmapsto y_{\text{part}}(t) := \int_{t_0}^{t} c(u)\frac{y_1(u)y_2(t) - y_2(u)y_1(t)}{W(u)}\,du\,.$$

La solution générale de l'équation avec second membre $y''(t) = ay'(t) + by(t) + c(t)$ s'écrit donc

$$y(t) = C_1y_1(t) + C_2y_2(t) + \int_{t_0}^{t} c(u)\frac{y_1(u)y_2(t) - y_2(u)y_1(t)}{W(u)}\,du \qquad (\dagger\dagger\dagger)$$

où C_1 et C_2 sont des constantes ; la solution de notre problème de Cauchy (conditions initiales $y(t_0) = y_0$ et $y'(t_0) = v_0$) s'obtient en ajustant les constantes pour que

$$\begin{aligned} C_1y_1(t_0) + C_2y_2(t_0) &= y_0 \\ C_1y_1'(t_0) + C_2y_2'(t_0) &= v_0\,, \end{aligned}$$

ce qui, on l'a vu, est possible en résolvant un système linéaire de deux équations à deux inconnues à coefficients complexes sont le déterminant est non nul. Le théorème de Cauchy est complètement prouvé dans ce cas général (et la solution exhibée).

IIIB. *Résolution du problème de Cauchy lorsque $a, b, y_0, v_0 \in \mathbb{R}$ et $c : I \longmapsto \mathbb{R}$*

Dans ce cas, on peut affirmer que notre problème de Cauchy

$$y'' - ay' - by - c = 0\,,\ y(t_0) = y_0,\ y'(t_0) = v_0\,,$$

admet une solution unique réelle définie dans I.

– Si $a^2+4b>0$ et si λ_1 et λ_2 sont les deux racines (réelles distinctes) du trinôme X^2-aX-b, la solution générale $y\ : I\longmapsto R$ de l'équation $y''(t)=ay'(t)+by(t)+c(t)$ s'écrit sous la forme (†††) avec $y_1(t)=\exp(\lambda_1 t)$ et $y_2=\exp(\lambda_2 t)$, les constantes C_1 et C_2 jouant le rôle des deux degrés de liberté étant ici réelles ; le problème de Cauchy se résout en ajustant ces constantes de manière à ce que

$$\begin{aligned} C_1y_1(t_0)+C_2y_2(t_0) &= y_0\\ C_1y_1'(t_0)+C_2y_2'(t_0) &= v_0\,; \end{aligned}$$

on trouve une unique solution réelle comme le stipule l'énoncé.

– Si $a^2+4b=0$ et si λ est la racine réelle double du trinôme X^2-aX-b, la solution générale $y\ : I\longmapsto R$ de l'équation $y''(t)=ay'(t)+by(t)+c(t)$ s'écrit sous la forme (†††) avec $y_1(t)=\exp(\lambda t)$ et $y_2=t\exp(\lambda t)$, les constantes C_1 et C_2 jouant le rôle des deux degrés de liberté étant ici réelles ; le problème de Cauchy se résout encore en ajustant ces constantes de manière à ce que

$$\begin{aligned} C_1y_1(t_0)+C_2y_2(t_0) &= y_0\\ C_1y_1'(t_0)+C_2y_2'(t_0) &= v_0\,; \end{aligned}$$

on trouve une unique solution réelle comme le stipule l'énoncé.

– Si $a^2+4b<0$ et si $\lambda\pm i\omega$ sont les deux racines complexes conjuguées du trinôme X^2-aX-b, la solution générale $y\ : I\longmapsto R$ de l'équation $y''(t)=ay'(t)+by(t)+c(t)$ s'écrit sous la forme (†††) avec $y_1(t)=\exp(\lambda t)\cos(\omega t)$ et $y_2=\exp(\lambda t)\sin(\omega t)$, les constantes C_1 et C_2 jouant le rôle des deux degrés de liberté étant ici réelles ; le problème de Cauchy se résout encore en ajustant ces constantes de manière à ce que

$$\begin{aligned} C_1y_1(t_0)+C_2y_2(t_0) &= y_0\\ C_1y_1'(t_0)+C_2y_2'(t_0) &= v_0\,; \end{aligned}$$

on trouve une unique solution réelle comme le stipule l'énoncé.

Au terme de cette étude, on a donc résolu de manière effective le problème de Cauchy linéaire du second ordre posé, tant dans le cadre complexe que réel. Le théorème est bien prouvé. □

IV. Remarque finale complémentaire d'ordre pratique

On peut aussi remarquer que si le second membre est de la forme $c\ : t\longmapsto P(t)\exp(wt)$, où w est un nombre complexe et P une fonction polynômiale de degré d, il est plus judicieux de chercher une solution particulière $\tilde{y}_{\text{part}}$ de la forme $z(t)e^{wt}$ (ce n'est pas la même qu'au **III**) et de se ramener ainsi à la résolution d'une équation

$$a_1z''(t)+b_1z'(t)+c_1z(t)=P(t) \qquad (*)$$

où P est une fonction polynômiale de degré d, que d'utiliser la méthode de variation des constantes décrite au **III** ; on cherche en effet une solution polynomiale particulière $t\longmapsto z(t)$ à l'équation $(*)$:

– si $c_1\neq 0$, il est possible de trouver une fonction $z=Q$ vérifiant (*) qui soit polynômiale en t de degré d (on la cherche en identifiant les coefficients pour que cela marche) ;
– si $c_1=0$ et $b_1\neq 0$, on cherche z' sous la forme d'une fonction polynômiale Z de degré d telle que $a_1Z'+b_1Z=P$ (on est ainsi ramené au premier sous-cas), puis on prend ensuite une primitive de Z pour trouver une fonction $z=Q$ solution de (*), polynômiale en t de degré $d+1$;
– si $c_1=b_1=0$, on construit une solution $z=Q$ de (*) en intégrant deux fois de suite P et en divisant par a_1 ; cela donne une solution polynômiale en t de degré $d+2$.

dans tous les cas, la fonction $z=Q$ trouvée est polynômiale en t, de degré au plus $d+2=\deg P+2$. La solution générale de l'équation $y''-ay'-by=c(t)$ s'écrit alors

$$y(t)=C_1y_1(t)+C_2y_2(t)+\tilde{y}_{\text{part}}(t)\,,$$

où y_1 et y_2 sont les fonctions introduites au **II** (dans le cas complexe ou dans le cas réel). Le problème de Cauchy se résout finalement en ajustant les constantes C_1 et C_2 pour que

$$\begin{aligned} C_1 y_1(t_0) + C_2 y_2(t_0) + \tilde{y}_{\text{part}}(t_0) &= y_0 \\ C_1 y_1'(t_0) + C_2 y_2'(t_0) + \tilde{y}_{\text{part}}'(t_0) &= v_0 \,. \end{aligned}$$

Par linéarité, on constate que la remarque s'applique également lorsque le second membre est de la forme $c : t \longmapsto P(t)\cos(\omega t)$ ou $c : t \longmapsto P(t)\sin(\omega t)$, P étant une fonction polynômiale de t et ω un nombre réel : il suffit en effet, pour construire dans ce cas une solution particulière de l'équation $y''(t) = ay'(t) + by(t) + c(t)$, de combiner des solutions particulières pour l'équation $y''(t) = ay'(t) + by(t) + \tilde{c}(t)$, où $\tilde{c}(t) = (P(t)/2)e^{\pm i\omega t}$ ou bien $\tilde{c}(t) = (P(t)/2i)e^{\pm i\omega t}$.

FIN DU CHAPITRE 3 ET DU COURS