Université Bordeaux 1
Département de licence
U.F.R. Mathématiques et Informatique

Mathématiques de base

Semestre d'orientation MISMI, cours MIS101. Version 2, Octobre 2005

Présentation du cours.

Cet ouvrage fournit un support pour un cours d'un semestre d'introduction aux mathématiques telles qu'elles sont enseignées à l'université. Il ne s'agit pas seulement d'un cours de mathématiques de base, il aspire en effet à donner un aperçu plus complet de la discipline en mettant en évidence d'une part ses logiques de développement internes et d'autre part en soulignant la place importante des mathématiques dans l'évolution des sciences et dans la compréhension du monde.

Nous avons essayé de rédiger un texte qui laisse entrevoir la richesse des concepts et des outils mathématiques et nous serions satisfaits si, avec ce cours, nous parvenions à donner une image des mathématiques comme d'une discipline vivante ayant une histoire très intense, encore largement à écrire, et qui est présente dans la vie de tous les jours par ses applications technologiques ou simplement parce que les mathématiques font partie de notre culture (qui n'a jamais entendu les expressions "c'est la quadrature du cercle" ou "les problèmes augmentent avec une croissance exponentielle" ?).

Souvent les mathématiques servent comme outil de sélection et encore au moment du Baccalauréat beaucoup d'élèves ne savent pas vraiment pourquoi on les a "forcés à subir" autant de mathématiques. Nous n'avons que très peu de moyens d'action pour changer cette situation, que nous déplorons. Par contre nous sommes fermement convaincus de l'utilité d'un cours comme le nôtre pour tout étudiant se destinant à faire des études scientifiques, ne serait-ce que pour le familiariser avec quelques unes des découvertes scientifiques les plus importantes de tous les temps : la méthode axiomatique, l'"analyse de l'infini" et son application au calcul différentiel, la notion de structure algébrique, ... Par le choix–plus large–des sujets abordés, mais surtout par la façon–plus approfondie–de les traiter, nous essayons de mettre à profit la rupture naturelle entre les études secondaires et les études supérieures.

Le cours est organisé en quatre parties :

I. Introduction à l'aspect formel des mathématiques
II. Nombres et limites
III. Mathématiques et réel
IV. Méthodes de calcul

Après une revue de quelques énoncés de mathématiques classiques et une mise en perspective historique, la première partie se focalise sur la notion de démonstration. L'approche est tout autant basée sur l'étude d'exemples que sur une présentation formelle (sans compromis). La deuxième partie, en partant de l'idée d'approximation d'une mesure (géométrique), développe une théorie des nombres réels, qui sans être complète, permet néanmoins de donner un sens à une expression comme "la racine carrée de deux est un nombre réel" (ceux qui considèrent ceci comme allant de soi devraient s'interroger sur la manière de calculer ne serait-ce que le produit par trois de la racine carrée de deux.) Dans la partie "Mathématiques et réel" on veut d'abord mettre en évidence le fait que ce n'est pas du tout clair *a priori* que le réel se laisse mettre en équation (ou, comme le disait Galilei, que le monde est écrit en langage mathématique). Cette "efficacité déraisonnable" des mathématiques est illustrée à travers la démarche de modélisation et la richesse des moyens pour représenter les phénomènes. La dernière partie aborde rapidement différentes méthodes de calcul, qui servent pour obtenir des solutions pratiques et efficaces à de nombreux problèmes abordés dans les parties précédentes. S'il n'est pas vrai qu'un mathématicien est seulement quelqu'un qui sait faire des calculs compliqués, c'est presque toujours le cas qu'un mathématicien amené à faire un calcul compliqué trouvera une méthode adéquate pour le faire.

Les auteurs de cet ouvrage enseignent à tous les niveaux–du Bac au doctorat–et mènent des recherches actives, dans différents domaines des mathématiques, au sein de l'Université Bordeaux 1.

Septembre 2004

C.-H. Bruneau,
B. Erez,
E. Kowalski,
N. Lince,
J.-M. Sebag,
A. Yger

Mise en garde.

Avec ce texte, nous avons voulu proposer un complément *utile*
au cours et aux travaux dirigés du cours de mathématiques de base
du semestre d'orientation MISMI. Suivant les chapitres,
le texte est soit trop (?) complet, soit trop vague par rapport au programme.
Vous êtes invités à passer rapidement sur les parties que vous trouvez
compliquées ou trop détaillées par rapport au cours d'amphi.

Ceci est la seconde édition du texte : y demeurent forcément encore
quelques erreurs (typographiques, oubli d'hypothèses, ...). Si vous en trouvez,
merci de nous les signaler. Si un exercice vous semble mal formulé,
envisagez l'exercice supplémentaire qui consiste à reformuler l'exercice
en question...
Nous avons complété le texte avec quelques références historiques.
Celles-ci sont très sommaires, et mériteraient d'être étoffées.
En effet, nous n'avons pas la prétention de faire un cours d'histoire des sciences,
bien que nous pensions qu'il soit important de placer les découvertes dans leur contexte.

Table des matières

Première partie

Introduction à l'aspect formel des mathématiques.

Voici comment D. Hilbert, un des grands mathématiciens du début du 20ème siècle, explique le développement des mathématiques de son temps à des lycéens, lors d'un cours pendant les vacances de Pâques de 1896 (notre traduction libre de *"Feriencursus : Über den Begriff des Unendlichen"* [Cours inter-semestre : Sur le concept d'infini.], Chap. 3 de "David Hilbert's Lectures on the Foundations of Geometry, 1891–1902", ed. M. Hallett et U. Majer, Springer-Verlag, Berlin, 2004).

> "La direction moderne des mathématiques tend vers la *précision des concepts* et la *force* des démonstrations, pour faire en sorte que la *proverbiale affidabilité* des vérités mathématiques *soit justifiée*. Les mathématiques modernes ont reconnu comme *erronées* un certain nombre de notions issues des mathématiques plus anciennes (*notions de fonction, de nombre, ...*). Il est important de trouver la *source* des erreurs. La principale est : que l'on a la tendance à transposer des propriétés évidentes pour les *ensembles finis* aux *ensembles infinis*. [Par exemple :] parmi un nombre donné de personnes il y en a toujours une qui est la plus petite, la plus grande, la plus jeune, la plus vieille, ... Mais parmi une infinité de nombres il n'en existe pas toujours un plus petit ou plus grand. Ainsi :
>
> $1, 2, 3, \ldots$ pas de plus grand
>
> $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$ pas de plus petit
>
> $1, \frac{1}{2}, 2, \frac{1}{3}, 3, \frac{1}{4}, 4, \ldots$ ni de plus grand, ni de plus petit.
>
> [Le fait que] l'esprit humain ait toujours eu une tendance à transposer des propriétés du fini vers l'infini a été une source importante d'erreurs par le passé, mais plus récemment a donné lieu à une *critique* constructive et un *approfondissement des concepts*. Par exemple la théorie des séries[1]. Lorsqu'on effectue la somme d'un nombre fini de termes on a le droit d'échanger l'ordre des termes ; pour un nombre infini de termes on n'a plus le droit :
>
> $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} \pm \cdots$ converge et vaut $\log 2$, mais
>
> $1 + \frac{1}{3} + \frac{1}{5} + \cdots$ $= \infty$, et
>
> $-\frac{1}{2} - \frac{1}{4} - \frac{1}{6} \cdots$ $= -\infty$,
>
> ainsi si l'on réordonne différemment, on peut obtenir toutes les valeurs. Par exemple :
>
> $$\begin{aligned} S = \quad & (\tfrac{1}{1} + \tfrac{1}{3} - \tfrac{1}{2}) && + && (\tfrac{1}{5} + \tfrac{1}{7} - \tfrac{1}{4}) && + (\tfrac{1}{9} + \tfrac{1}{11} - \tfrac{1}{6}) + \cdots \\ \log 2 = \quad & (\tfrac{1}{1} - \tfrac{1}{2} + \tfrac{1}{3} - \tfrac{1}{4}) && + && (\tfrac{1}{5} - \tfrac{1}{6} + \tfrac{1}{7} - \tfrac{1}{8}) && + (\tfrac{1}{9} - \tfrac{1}{10} + \tfrac{1}{11} - \tfrac{1}{12}) + \cdots \\ S - \log 2 = \quad & (\tfrac{1}{2} - \tfrac{1}{4}) && + && (\tfrac{1}{6} - \tfrac{1}{8}) && + (\tfrac{1}{10} - \tfrac{1}{12}) + \cdots \\ = \quad & \tfrac{1}{2}(1 - \tfrac{1}{2} && + && \tfrac{1}{3} - \tfrac{1}{4} && + \tfrac{1}{5} - \tfrac{1}{6} + \cdots) = \tfrac{1}{2} \log 2, \end{aligned}$$
>
> c'est-à-dire que $S = \frac{3}{2} \log 2$. Ces considérations mènent à la notion de *convergence absolue* [d'une série, qui garantit que la valeur de la somme infinie ne dépend pas de l'ordre de sommation].

[1] Il s'agit de sommes infinies de nombres ou de fonctions, définies par un processus de passage à la limite, que nous traiterons dans la deuxième partie de ce cours. Il était important pour Hilbert de donner des exemples de notions, qui venaient juste d'être éclaircies, notamment par Weierstraß. Les sommes infinies de fonctions circulaires/trigonométriques interviennent de manière essentielle dans la description de phénomènes physiques tels que la propagation de la chaleur (analyse de Fourier).

De même une somme finie de fonctions continues est encore continue. Ce n'est pas le cas pour une somme infinie. Par exemple : la série

$$f(x) = x + (x^2 - x) + (x^3 - x^2) + \cdots$$

converge pour $0 \leq x \leq 1$. [On montre que,]

$f(x) = 0 \quad$ si $0 \leq x < 1$ et

$f(x) = 1 \quad$ si $x = 1$,

ainsi la fonction f *n'est pas continue* en $x = 1$. Et on est amené à la notion de *convergence uniforme* [qui garantit–entre autre–la continuité d'une somme infinie de fonctions continues.][...] D'où la notion fondamentale de série *absolument et uniformément convergente*, qui nous dispense dans toute l'Analyse de considérer d'autres types de séries.

Mais encore plus intéressant est ce que l'on a gagné avec le traitement *moderne* du concept d'infini. Considérons les exemples d'ensembles infinis :

$1, 2, 3, 4, 5, \ldots$	
$2, 3, 4, 5, 6 \ldots$	
$2, 4, 6, 8, 10, \ldots$	
$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots$	(tous les nombres rationnels)
$1,\ 1,011 \ldots, 1,1011 \ldots$	(tous les nombres réels)
(x, y)	(tous les couples de nombres, par ex. rationnels)
(x, y, z)	avec x, y et z entre 0 et 1 (les points à l'intérieur d'un cube)

toutes les fonctions continues.

Peut-on comparer [la taille de] ces différents ensembles ? Pour les *ensembles finis* la *définition de l'égalité* est : il y a égalité [du nombre d'éléments], lorsque il est possible d'établir une correspondance qui fait correspondre les éléments un à un. *Dames et Messieurs* sont en nombre égal, s'ils peuvent se donner les mains par couples, sans oublier personne.

[C'est cette notion d'égalité, qu'il est bien de généraliser aux ensembles infinis.] Notons que la notion de dimension ne nous avance pas : il est même possible d'obtenir un carré comme l'image d'un segment par une fonction continue. [Il y a beaucoup plus de nombres réels que de nombres rationnels.] Les nombres réels ne sont pas dénombrables. [Ceci se démontre sans exhiber des nombres particuliers et montre donc qu'] il existe des nombres réels, qui ne sont pas rationnels. On connaît évidemment des exemples de tels nombres ; par exemple $\sqrt{2}$ ou e, [mais au moment où je vous parle] on ne sait pas encore si $\sqrt{2}^{\sqrt{2}}$ est rationnel[2].”

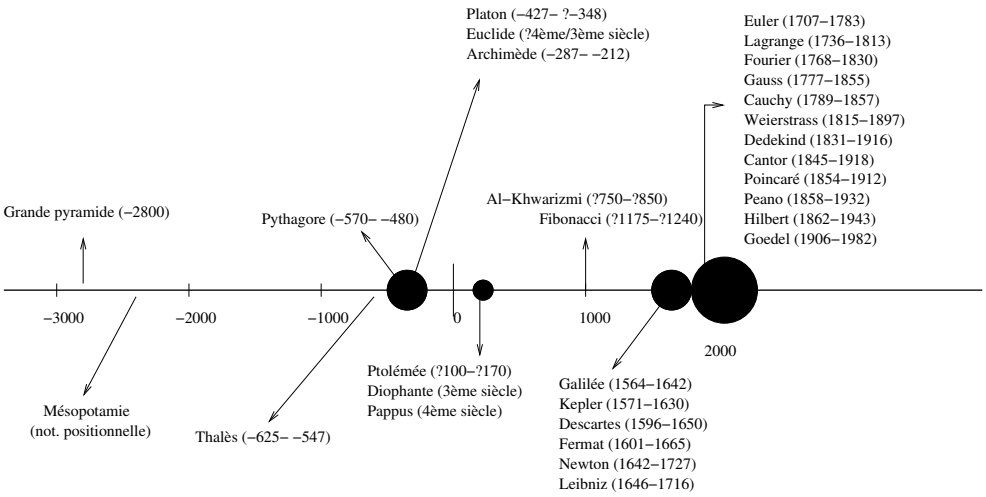
[2]Cette question précise a été résolue en 1920 par Siegel, et est un cas particulier d'un résultat plus général de Gel'fond et Schneider, qui ont montré indépendamment l'un de l'autre en 1934, que si α et β sont deux nombres, chacun solution d'une équation polynomiale à coefficients rationnels, avec α différent de 0 et de 1 et avec β irrationnel, alors α^β ne peut satisfaire aucune équation polynomiale à coefficients rationnels. Par conséquent $\sqrt{2}^{\sqrt{2}}$ n'est pas rationnel.

Chapitre 1

Les mathématiques ont une histoire.

1.1 Une présence variable.

Le domaine d'études recouvert par les mathématiques a changé avec les temps. L'astronomie au temps des Grecs était essentiellement une application de considérations mathématiques à la solution du problème posé par la description du mouvement des astres. De nos jours l'astronomie est considérée comme une branche de la physique, qui à son tour utilise beaucoup plus qu'au temps des Grecs les méthodes mathématiques. En plus des branches classiques des mathématiques (analyse, algèbre, géométrie, probabilités et statistiques, ...), que l'on cultive suivant leur logique propre ou en vue d'applications, on a aussi vu se développer avec vigueur des disciplines nouvelles telles que les mathématiques financières ou la biologie mathématique. De plus, les classifications des anciens ont perdu un peu de leur intérêt, après que l'on ait réussi à unifier les différentes branches en leur donnant un fondement formel commun. Cela dit on peut parler de l'histoire des mathématiques (ou de *la* mathématique), et–peut-être mieux–de la place des mathématiques dans l'Histoire. Celle-ci a eu des hauts et des bas. Toutes les civilisations n'ont pas accordé la même importance à cette science. Le nombre de ceux qui en cultivaient l'étude a grandement varié à travers les âges.



Sur la figure on représente la densité des travaux en mathématiques à travers le temps. On voit que, par exemple, les Romains n'ont pas donné de grande importance aux mathématiques. Ils ont plutôt développé le droit et la dialectique, et ils ont été d'excellents bâtisseurs/ingénieurs.

Une des raisons pour s'intéresser à l'histoire des mathématiques est que *tous les résultats mathématiques, même les plus simples, sont des conquêtes de l'esprit, le fruit du travail d'un ou de plusieurs hommes ou femmes, qui nous ont précédés.* La même chose est évidemment vraie d'autres disciplines, mais on a souvent tendance à penser que les mathématiques sont "éternelles", ou alors que l'on ne peut les comprendre, que si l'on dispose d'antennes spéciales, qui permettent d'entrer en contact avec un monde abstrait, figé et qui a une existence propre, indépendante du "monde réel".

Or, des propriétés de base des nombres, comme la notation positionnelle et l'utilisation du zéro, ou la notion même de nombre, ont mis des siècles a être précisées par des générations de penseurs, qui ont ainsi forgé des outils conceptuels très puissants pour résoudre les problèmes qui se posaient à eux : mesure de grandeurs variées (distances, superficies,...), conversion entre différentes unités de mesure, précision comptable, etc.

Les raisons qui ont fait que l'on a étudié les mathématiques ont aussi évolué avec le temps. Elles présentent plusieurs facettes, dont voici un choix important. Les mathématiques permettent :

- de donner une *description* de régularités observées ; ainsi Kepler a pu faire usage de la théorie des coniques d'Apollonius pour interpréter les mesures astronomiques de Brahé ; on dit souvent que les mathématiques sont un langage particulier ;
- de faire des *prédictions*, par exemple avec l'établissement de modèles mathématiques construits sur la base de lois physiques formulées en langage mathématique ;
- d'améliorer la *compréhension* de la structure des choses et de concepts ; comment les choses s'agencent, quelles lois sont fondamentales, qu'est-ce que l'infini ?
- de satisfaire un certain goût *esthétique* ; la beauté d'un raisonnement, d'un enchaînement d'idées.

Les mathématiques ont souvent été utilisées pour séparer les sciences "dures", celles qui en font un grand usage (physique, chimie), des sciences "molles" (ou, mieux, "souples"), qui n'auraient pas encore suffisamment clarifié leurs concepts pour arriver à être mathématisées et ne sont pas encore en mesure de fournir des prévisions quantitatives fiables (sociologie). Nous n'attachons pas une grande importance à ces distinctions.

1.2 Ce que l'on apprenait autrefois à l'université (arithmétique digitale).

Le lecteur se sera certainement déjà posé la question de savoir pourquoi les programmes scolaires comportent autant de mathématiques. Puis encore maintenant, pourquoi même si son choix de parcours de formation aurait dû l'éloigner de mathématiques de plus en plus en abstraites, il se voit contraint de se familiariser avec des concepts en apparence si éloignés de ses intérêts. Le fait est que peu de choses dans un cours comme celui-ci s'avèrent inutiles dans la constitution d'un bagage scientifique de base au niveau universitaire.

Au Moyen âge on enseignait à l'Université comment effectuer sur les doigts des calculs longs et parfois complexes. Beaucoup d'étudiants se destinaient alors à exercer des activités commerciales. Ils payaient directement les enseignants, qui n'avaient pas vraiment d'autre salaire que ces "frais de scolarité". L'institution universitaire a beaucoup changé, mais les programmes d'étude sont toujours construits autour de ce qui est utile pour les étudiants (et non pour les enseignants !). *Si il y a autant de mathématiques dans les programmes, c'est que les mathématiques ont été développées à tel point, que l'on ne peut obtenir une bonne compréhension de ce qui nous entoure, sans une culture mathématique de base.* Or, ce qui nous entoure est de plus en plus le produit de l'Homme et de sa technologie, et celle-ci est largement basée sur des concepts mathématiques.

Pour marquer une pause après ce discours grandiloquent et pour vraiment entrer en matière, passons en revue quelques règles de calcul digital, c'est-à-dire de calcul sur les doigts, comme elles auraient pu être enseignées dans une université du Moyen âge.

Multiplication par 9. On pose les deux mains à plat devant soi, avec les pouces au centre. On numérote les doigts de gauche à droite. Le résultat de la multiplication $a \times 9$ (pour a au plus égal à 10) s'obtient en pliant le doigt numéro a : le nombre de doigts à gauche du doigt plié donne les dizaines et le nombre de doigts à droite de ce doigt donne le nombre d'unités. On peut vérifier que cette règle est correcte en vérifiant tous les cas. Ceci donne une *démonstration par épuisement (de tous les cas)*.

Exercice. Expliquer pourquoi l'identité algébrique

$$10(x-1) + (10-x) = 9x$$

donne une autre démonstration du fait que la règle est correcte.

Multiplier deux nombres plus grands que 5. On pose encore les mains devant soi, comme ci-dessus, mais cette fois fermées avec les doigts pliés. Si on veut multiplier $5+a$ par $5+b$ (pour a et b au plus égaux à 5), on procède comme suit. On déplie a doigts de la main gauche et b doigts de la main droite (en commençant par les pouces). Puis, on somme le nombre de doigts dépliés : cela donne les dizaines. Le produit du nombre des doigts pliés de la main gauche avec ceux de la main droite donne le nombre d'unités. A nouveau on peut vérifier la règle par épuisement.

Exercice. Trouver une identité algébrique qui explique la règle ci-dessus.

Exercice. Trouver une façon "digitale" pour multiplier deux nombres supérieurs à 10.

Chapitre 2

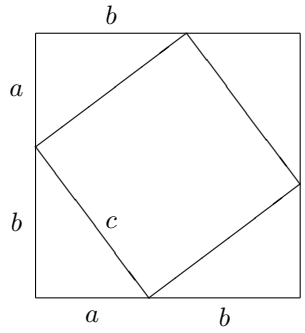
Quelques énoncés et démonstrations de mathématiques classiques.

Avant de procéder à une introduction formelle de notre sujet, nous passons en revue des résultats fondamentaux et des idées de leurs démonstrations. L'objectif est multiple : de nous rappeler ces énoncés, de mieux comprendre leur origine, de préparer la suite du cours et de nous familiariser avec certaines techniques de démonstration, qui vont jouer un rôle important plus tard.

2.1 Le Théorème de Pythagore.

Le Théorème de Pythagore (déjà connu sous des formes analogues par les Égyptiens, les Indiens et les Chinois) énonce, dans sa version géométrique, une relation entre les aires des carrés construits sur les côtés d'un triangle rectangle.

Exercice. Donner un énoncé du Théorème et expliquer comment utiliser la figure pour en donner une démonstration. Quelles sont les notions qui interviennent dans une telle démonstration ?



Une démonstration d'une généralisation du Théorème de Pythagore aux triangles quelconques, basée sur l'utilisation des coordonnées cartésiennes sera donnée dans la Partie III.

2.2 Longueurs commensurables, l'algorithme d'Euclide et l'identité de Bézout.

On dit que deux segments sont *commensurables* s'il existe un segment dont les deux segments sont des multiples *entiers*. Deux segments commensurables ont donc une (unité de) mesure commune. Si deux segments ne sont pas commensurables, on dit qu'ils sont *incommensurables*.

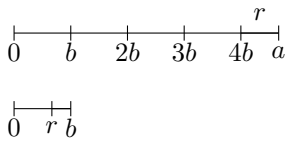
Pour essayer de trouver une (unité de) mesure commune à deux segments on peut procéder comme suit. On prend le segment le plus court[1] et on voit combien de fois il rentre dans le segment le plus grand. Il y a deux cas. Si un multiple (entier) du segment le plus petit donne exactement le segment le plus grand, les deux segments sont commensurables et ont pour mesure commune le segment le plus petit.

Sinon, on considère la différence entre le plus grand multiple du segment le plus petit, qui rentre dans le grand segment. Cette différence est par définition un segment plus petit, que le petit segment. On recommence ensuite en voyant combien de fois la différence trouvée rentre dans le petit segment. A nouveau il y a deux cas.

Si un multiple (entier) de la différence donne exactement le petit segment, c'est que la différence et le petit segment sont commensurables. Mais, c'est beaucoup mieux ! En effet dans ce cas la différence est aussi une mesure commune entre les deux segments de départ : le grand segment est par définition égal à la somme du petit et de la différence ; vu que le petit est multiple de la différence, le grand est aussi multiple de la différence.

Sinon on recommence avec la différence et ce que l'on pourrait appeler la deuxième différence. On peut itérer cette construction. Il y a deux cas : si la procédure s'arrête, c'est que les deux segments *de départ* sont commensurables et ont pour mesure commune la dernière différence ; si la procédure ne s'arrête pas c'est que les segments ne sont pas commensurables. Évidemment ce n'est pas parce que après dix millions d'itérations la procédure ne s'est pas arrêtée, qu'elle ne va pas s'arrêter à l'itération suivante. La procédure n'est donc pas utile pour démontrer l'incommensurabilité de deux segments.

Cette procédure est connue sous le nom d'*algorithme d'Euclide*. [2] L'algorithme est basé sur l'itération de soustractions, mais d'habitude, dans sa forme arithmétique, il est présenté comme un algorithme permettant de trouver le plus grand commun diviseur de deux entiers. Avant de nous restreindre aux entiers reprenons la description ci-dessus sous forme numérique.



Soient donc a et b les longueurs des segments considérés et supposons, que b est inférieur à a. Soit q le plus grand *entier* tel que $qb \leq a$ (q vaut 4 dans l'exemple de la figure). On a donc

$$a = qb + r \ , \text{ avec } 0 \leq r < b \ .$$

On appelle q le *quotient* et r le *reste*. Ici r représente la longueur de la différence. Si $r = 0$, on a que a est un multiple entier de b, autrement dit b divise a. Si par contre $r > 0$, alors on recommence et on cherche un entier q_1 tel que

$$b = q_1 r + r_1 \ , \text{ avec } 0 \leq r_1 < r \ .$$

[1] Si les deux segments sont égaux ils sont évidemment commensurables.

[2] Nous parlerons plus loin des "Éléments" d'Euclide. L'algorithme dont il est question figure dans les Propositions 1 et 2 du Livre VII de cet ouvrage.

A nouveau, si $r_1 = 0$, alors r divise b, mais aussi a : il suffit d'écrire $a = q(q_1 r) + r = (qq_1 + 1)r$. Si $r_1 > 0$, alors on cherche un entier q_2 tel que

$$r = q_2 r_1 + r_2 \ , \text{ avec } 0 \leq r_2 < r_1 \ .$$

Montrons que si $r_2 = 0$, alors r_1 divise a et b. En effet, alors r_1 divise r et donc r_1 divise b, car il divise les deux termes de la somme $b = q_1 r + r_1$. De même r_1 divise a car on vient de voir que r_1 divise r et b. On voit donc que si on a une suite (finie) d'égalités

$$\begin{aligned} a &= qb + r \\ b &= q_1 r + r_1 \\ r &= q_2 r_1 + r_2 \\ &\cdot \\ &\cdot \\ r_k &= q_{k+2} r_{k+1} + r_{k+2} \\ &\cdot \\ &\cdot \\ r_{n-1} &= q_{n+1} r_n + r_{n+1} \\ r_n &= q_{n+2} r_{n+1} \ , \end{aligned}$$

avec les q_k entiers et $r_{n+2} = 0 < r_{n+1} < r_n < r_{n-1} < \cdots < r_1 < r$, alors *$a$ et b sont multiples du dernier reste non-nul r_{n+1}*.

Ce qui précède est valable pour tous a et b, mais comme on l'a noté il n'est pas du tout clair que l'algorithme s'arrête. Notons que si les segments de départ sont multiples d'un même segment S, alors dans la version numérique on peut supposer que a et b sont eux-mêmes entiers. En effet, si le premier segment est k fois le segment S et le second ℓ fois S, et si s dénote la longueur de S, alors $a = ks$ et $b = \ell s$. En divisant par s, on peut remplacer a par k et b par ℓ.

Si a et b sont entiers, alors l'algorithme donne le plus grand commun diviseur $pgcd(a, b)$ de a et b. Que le dernier reste non-nul soit un diviseur commun résulte de ce qui précède. Pour voir que celui-ci est le plus grand parmi tous les diviseurs communs de a et de b, soit d un tel diviseur. Alors, en utilisant les notations ci-dessus, d divise a et b, donc d *divise la différence* $r = a - qb$. De même, d divise donc la différence $r_1 = b - q_1 r$, et ainsi de suite on voit, que d divise forcément le dernier reste non-nul r_{n+1}. Donc celui-ci est bien le plus grand des diviseurs communs.

En remontant la chaîne d'égalités ci-dessus on voit aussi qu'il existe deux entiers (relatifs) u et v, tels que

$$\text{pgcd}(a, b) = ua + vb \ .$$

C'est *l'identité de Bézout*[3], qui joue un rôle important en arithmétique. Elle permet en particulier de trouver toutes les solutions en entiers x et y d'une équation de la forme $ax - by = c$, avec a, b et c entiers. On résout d'abord l'équation $ax - by = d$, où $d = \text{pgcd}(a, b)$ par la méthode ci-dessus. Puis on montre que la première équation n'a de solution, que si c est multiple de d.

Exercice. Trouver le plus grand commun diviseur de 71755875 et 61735500.

Exercice. Résoudre l'équation $17x - 11y = 542$ en entiers.

Les démonstrations dans ce paragraphe ont été plutôt informelles. En fait, nous nous sommes convaincus du bien fondé des affirmations *par inspection*, en allant voir le processus de près. Même si on peut formaliser les démonstrations de ce type, il est souvent plus convaincant que de voir pourquoi et comment elles marchent.

[3] E. Bézout, mathématicien français, 1730–1783.

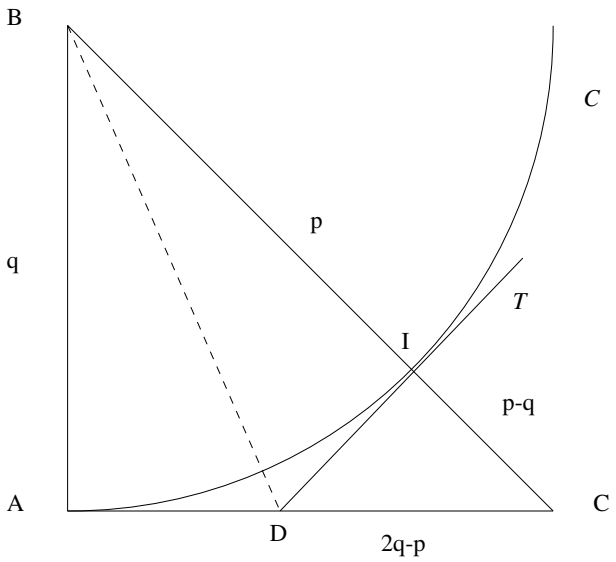
2.3 Incommensurabilité de la diagonale avec le côté d'un carré.

Nous avons souligné que l'algorithme d'Euclide ne permet pas de montrer que deux segments sont incommensurables. D'ailleurs, on a pensé à un certain moment que toutes les longueurs étaient commensurables (du moins dans l'Ecole pythagoricienne), et cela fut une grande surprise d'apprendre que *la diagonale d'un carré n'est pas commensurable au côté*. Cela peut se montrer avec une *démonstration par l'absurde*, en supposant que la diagonale et le côté sont commensurables, c'est-à-dire qu'il existe un segment tel que la diagonale et le côté soient des multiples *entiers* du segment, et de montrer que cette hypothèse mène à une contradiction.

1) On peut se ramener à considérer un carré dont les côtés ont une longueur entière. On a donc un *triangle* ABC qui est isocèle, rectangle et à côtés entiers, disons p et q (avec p le côté le plus long).

2) A partir de ABC on construit géométriquement un triangle ayant les mêmes propriétés que ABC, mais avec des côtés plus petits.

3) Cela suffit. La contradiction consiste dans le fait que d'une part on peut itérer l'étape (2) (encore et encore) et d'autre part le fait que il existe un entier plus petit que tous les autres (parmi tous ceux qui sont longueurs de côtés de triangles envisagés).



Voici une figure qui explique la deuxième étape : on commence par construire le cercle de centre B et rayon q, puis on trace la tangente au cercle par le point d'intersection I du cercle et du segment BC. Cette tangente coupe le segment AC en D. Il est clair que IC mesure $p-q$, qui est entier. Il est un peu plus difficile de montrer que le triangle DIC est rectangle et isocèle et que le côté DC est de longueur $2q-p$ et donc entier. (En fait une démonstration complète est hors de portée pour nous ! Voir ce qu'il faut faire pour démontrer des assertions très simples à partir des axiomes d'Euclide.)

Exercice. Notons que l'on peut résumer l'essentiel de la démonstration de l'incommensurabilité du côté et de la diagonale du carré en disant qu'*en pliant le côté sur la diagonale* on obtient un triangle (rectangle, isocèle) semblable à celui de départ, mais plus petit. Ce triangle plus petit aura des côtés entiers, si celui de départ a des côtés entiers, *etc.* Montrer que la même approche permet de montrer :

a) l'incommensurabilité de la diagonale et du côté d'un pentagone régulier ;

b) l'incommensurabilité de la petite diagonale et du côté d'un hexagone régulier.

2.4 Il n'existe pas de rationnel dont le carré est 2.

Comme l'algorithme d'Euclide, l'énoncé du paragraphe précédent a lui aussi une version numérique. En fait, pour l'énoncer il ne faut même pas faire appel aux nombres rationnels. La voici : *il n'existe pas d'entiers a et b tels que* $2b^2 = a^2$. A nouveau on peut procéder par l'absurde. On suppose que l'on peut trouver a et b entiers avec $2b^2 = a^2$, et on en tire l'existence d'entiers $a' < a$ et $b' < b$ tels que (encore) $2b'^2 = a'^2$. Ceci mène clairement à une contradiction, car à partir de a' et b' on peut alors trouver des entiers $a'' < a' < a$ et $b'' < b' < b$ avec $2b''^2 = a''^2$ etc. La contradiction réside dans le fait que cela donne une suite (infinie) d'entiers sans plus petit élément.

Pour trouver a' et b' on étudie la parité des entiers en jeu. Si $2b^2 = a^2$, alors a^2 est clairement pair, mais ceci implique que a lui-même est pair. En effet on a la proposition :

si n est un entier et n^2 est pair, alors n est pair.

Nous allons admettre cette proposition, que nous démontrerons plus tard (par contraposition, dans le chapitre sur la logique, voir Sect 3.1). Donc a est pair. Mais alors, a^2 est divisible par 4 et donc b^2 est aussi pair, d'où par la proposition b est pair. Écrivons $a = 2a'$ et $b = 2b'$ avec a' et b' entiers. Alors, en remplaçant et en divisant on obtient bien $2b'^2 = a'^2$.

Exercice. Pour x un nombre réel, on note $[x]$ la *partie entière* de x, c'est-à-dire le plus grand entier inférieur ou égal à x. La *partie fractionnaire* de x est $\{x\} = x - [x]$. Soit N un entier naturel, qui n'est pas le carré d'un entier. Supposons (par l'absurde) que $\sqrt{N}$ est rationnel et qu'il s'écrit B/A comme fraction réduite.

a) Montrer que $B/A = NA/B$.
b) Montrer que les parties fractionnaires de B/A et de NA/B ont respectivement la forme a/A et b/B, avec a et b des entiers positifs, plus petits que A et B.
c) En déduire que $b/a = B/A$ et que par conséquent B/A n'est pas réduite...

Notons que si $x^2 = 2$, alors on peut écrire $x^2 - 1 = 1$, ou encore $(x+1)(x-1) = 1$ et finalement $x = 1 + 1/(1+x)$. On peut voir cette écriture comme le résultat de la division de x par 1 dans l'algorithme d'Euclide. Avec un peu de courage on peut alors remplacer le x du membre de droite par $1 + 1/(1+x)$, ce qui donne

$$x = 1 + \frac{1}{2 + \frac{1}{1+x}} \ .$$

Puis encore

$$x = 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2+\cdots}}}$$

Un tel développement est appelé un développement en *fraction continue*. En remplaçant les petits points par 0 on obtient des nombres rationnels, qui approchent la racine de deux x : on trouve les nombres 1, $3/2 = 1,5$, $7/5 = 1,4$, $17/12$, $41/29$, $99/70$, $293/169$, $577/408$, etc.

Exercice. Expliciter le lien entre l'algorithme d'Euclide et le développement en fraction continue. (Indications : par exemple si on applique l'algorithme pour trouver le plus grand commun diviseur de 105 et 24 on trouve, que $105 = 4 \cdot 24 + 9$, que l'on peut réécrire $105/24 = 4 + 9/24$, puis $24 = 2 \cdot 9 + 6$ s'écrit $24/9 = 2 + 6/9$, qui donne une expression pour $9/24$ sous la forme $1/(2 + 6/9)$, etc.)

Exercice. Trouver le développement en fraction continue et des approximations rationnelles du nombre ϕ tel que $\phi = 1/(1+\phi)$.

2.5 Le nombre π et l'impossibilité de la quadrature du cercle.

On connaît les formules pour l'aire $A(r)$ et le périmètre $L(r)$ d'un cercle de rayon r :

$$A(r) = \pi r^2 \quad \text{et} \quad L(r) = 2\pi r \ .$$

Mais qu'est-ce que π ? On peut définir π à l'aide de l'une de ces formules, par exemple ainsi : π *est le rapport entre le périmètre et le diamètre d'un cercle.* Il faut remarquer que dans une telle définition on ne précise pas quel cercle on considère ! On fait comme si l'on savait que le rapport en question est toujours le même. On peut se convaincre (?) que c'est bien le cas en observant que la longueur du périmètre change de manière linéaire avec la longueur du rayon...

On peut aussi utiliser l'autre formule. Alors la définition serait : π *est le rapport entre l'aire du cercle et l'aire d'un carré construit sur le rayon.* Mêmes remarques, sauf que maintenant l'aire varie avec le carré du rayon (c'est clair pour des carrés et on peut donc extrapoler en remplissant le cercle de petits carrés...).

Il faut au moins vérifier que ces deux définitions sont compatibles. On peut rendre la compatibilité plausible de la manière suivante. On considère les polygones réguliers à n côtés P_n, inscrits dans le cercle de rayon r. On note b_n la longueur d'un côté de P_n et h_n la hauteur du triangle ayant pour base l'un des côtés de P_n et de sommet l'origine. Alors l'aire $A(P_n)$ et le périmètre $L(P_n)$ de P_n sont donnés par :

$$A(P_n) = n \cdot \frac{b_n h_n}{2} \quad \text{et} \quad L(P_n) = n \cdot b_n \ .$$

Lorsque n croît $A(P_n)$ (resp. $L(P_n)$) s'approche de $A(r)$ (resp. $L(r)$) et h_n s'approche de r. Ainsi (?) le quotient

$$\frac{A(P_n)}{L(P_n)} = \frac{n \cdot (b_n h_n)}{2(n \cdot b_n)} = \frac{h_n}{2}$$

s'approche de $r/2$.

Les approximations polygonales permettent de calculer des approximations de π défini comme la longueur du demi-cercle de rayon 1. [4] On considère ici les polygones réguliers à $m = 6 \cdot 2^n$ côtés (petit changement de notation). Sur la figure qui suit on représente le cas de $n = 0$, donc $m = 6$. Noter que ce dessin montre déjà que π est plus grand que 3, qui est la longueur du demi-périmètre de l'hexagone régulier.
Plus généralement soit $a_n = |AB|3 \cdot 2^n$ et $b_n = |CD|3 \cdot 2^n$, où les segments AB et CD se rapportent à la figure où serait représenté le polygone régulier à $6 \cdot 2^n$ côtés. Donc a_n (resp. b_n) est le demi-périmètre du polygone régulier à $m = 6 \cdot 2^n$ côtés exinscrit (resp. inscrit) au cercle de rayon 1. Par exemple $a_0 = 2\sqrt{3}$ et $b_0 = 3$. On peut montrer (exercice !) que

$$b_n < b_{n+1} < a_{n+1} < a_n$$

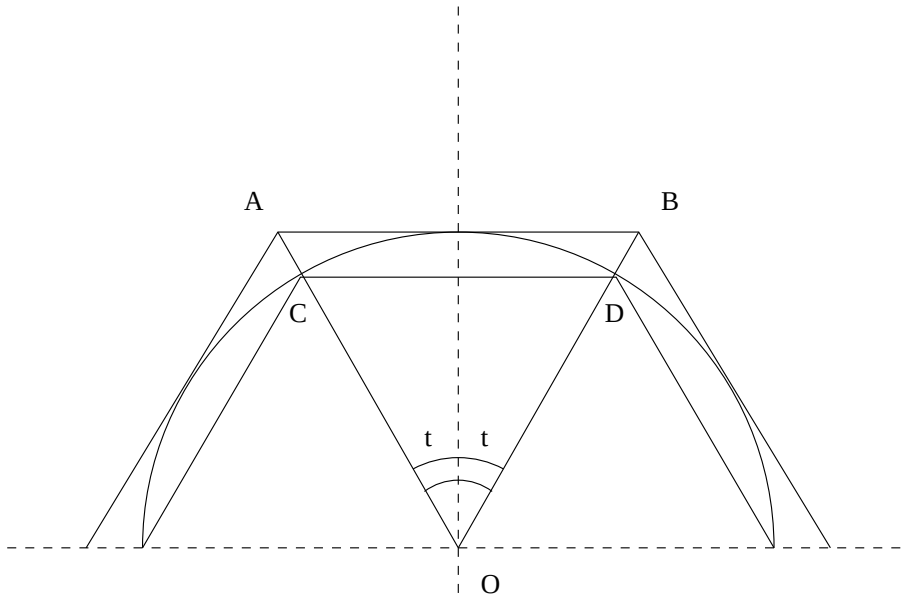
et

$$0 < a_n - b_n < \left(\frac{1}{3}\right)^n (a_0 - b_0) < \left(\frac{1}{3}\right)^n \ .$$

Ceci montre (!) qu'il y a un unique nombre *coincé* entre les b_n et les a_n. Ce nombre est π et les encadrements donnés ici permettent de l'approcher. En fait on a les formules :

$$a_{n+1} = \frac{2a_n b_n}{a_n + b_n} \quad \text{et} \quad b_{n+1} = \sqrt{a_{n+1} b_n} \ .$$

[4] Cette méthode d'approximation peut servir de base au calcul des aires et de longueurs. Voir le chapitre sur les intégrales.



Si $n = 4$, c'est-à-dire $m = 96$, ceci donne pour a_4 et b_4 les valeurs approchées $3,142715$ et $3,141031$. En fait Archimède (287–212 av. J.-Chr.) avait déjà procédé comme nous l'avons fait et avait *de plus* trouvé l'encadrement remarquable[5]

$$3 + \frac{10}{71} < b_4 < \pi < a_4 < 3 + \frac{10}{70} \ .$$

L'*impossibilité de la quadrature du cercle* s'énonce en disant que l'on ne peut pas construire à la règle et au compas un carré dont l'aire serait égale à l'aire d'un cercle de rayon unité. Terminologie équivalente : que *le cercle n'est pas quarrable*.

Par *construction à la règle et au compas* on entend une construction (en plusieurs étapes) effectuée à partir de la seule donnée d'un segment (unité) à l'aide d'intersections de cercles et de lignes droites, où les cercles ont pour rayons et centres des longueurs et des points déjà construits et où les lignes sont pareillement définies par des points déjà construits.

La quadrature du cercle à la règle et au compas est seulement un d'une série de problèmes que déjà les Grecs avaient énoncés et qui demandent de construire certaines grandeurs définies géométriquement, sous certaines contraintes. D'autres problèmes de ce type sont : la construction (à la règle et au compas) d'un cube de volume double d'un cube de côté l'unité ou la division en trois parties égales d'un angle (quelconque) donné. Les démonstrations que ces constructions sont impossibles (à la règle et au compas) n'ont été données que beaucoup de temps plus tard. Ainsi, ce n'est qu'au 19ème siècle que l'on a démontré l'impossibilité de la quadrature du cercle.[6] Ces démonstrations passent par une traduction en termes numériques–on abandonne la géométrie–et reposent sur une analyse de la "complexité" des

[5]Dans l'Ancien Testament, "Livre des Rois", 7.23, on donne à π la valeur 3. Il y est dit : "Hiram, l'architecte engagé par Salomon pour construire son temple, construisit un grand bassin de bronze, qui avait 10 [longueurs] d'un bord à l'autre, était entièrement rond, haut de 5 [longueurs] et une corde de 30 [longueurs] l'aurait entouré".

Une façon pour se souvenir des premiers chiffres dans le développement décimal de π est d'apprendre par coeur la phrase suivante et de considérer la longueur des mots qui la composent : "Que j'aime à faire apprendre un nombre utile aux sages. Immortel Archimède, artiste ingénieur, ...". De telles phrases avaient déjà été concoctées par les Grecs : "'Aει ὁ θεὸς ὁ μεγας γεωμετρεῖ", est une phrase attribuée à Platon et qui signifie à peu près "Toujours le grand Dieu géométrise".

[6]Lindemann, "Über die Zahl π", Math. Ann. **20**(1882), 213–225.

nombres réels. L'impossibilité de la quadrature du cercle est une conséquence du fait que le nombre π ne satisfait aucune équation polynomiale à coefficients rationnels (on dit que π est un *nombre transcendant*; en particulier il n'est pas rationnel).

Les Grecs avaient montré que certaines surfaces étroitement liées au cercle étaient quarrables. C'est le cas des *lunules hippocratiques*, dont la plus simple s'obtient en retranchant à un demi-disque construit sur la diagonale d'un carré de côté a la portion du quartier du disque de rayon a, qui a la diagonale pour corde. On montre (exercice !), que la lunule ainsi obtenue a la même aire que le triangle isocèle de base la diagonale du carré et de coté a.

Les Grecs avaient aussi observé que la quadrature des lunules aurait pu amener à la quadrature du cercle. Considérons en effet la lunule $\mathcal{L}$ obtenue comme suit : on construit hexagone régulier inscrit dans un disque $\mathcal{D}$ de rayon 2 et on construit le demi-disque de diamètre le côté L de l'hexagone ; $\mathcal{L}$ est obtenue en retranchant à ce demi-disque la portion du disque $\mathcal{D}$ coupé par la corde L. Soit $\mathcal{T}$ le trapèze égal à la moitié de l'hexagone considéré, alors (exercice !) on a : $\pi = 2(\mathcal{T} - 3\mathcal{L})$.

Dans une autre voie, Hippias d'Elée (env. 420 av. J.C.) avait réduit la quadrature du cercle à la construction à la règle et au compas d'une courbe définie de manière "mécanique" : la *quadratrice*. Cette courbe est définie par le parcours dans le plan qu'effectue un point P, intersection de deux droites d_1 et d_2 qui se déplacent comme nous allons maintenant l'expliciter. On trace un quart de cercle avec centre l'origine O et rayon OA, et on considère un rayon OB perpendiculaire à OA. La droite d_1 passe par l'origine O et coupe le cercle en Q ; la droite d_2 est perpendiculaire à OB (donc parallèle à OA) et coupe OB en R. Le mouvement des deux droites est déterminé par la condition que Q et R parcourent respectivement BA et BO à vitesse uniforme, de manière à ce que, s'ils partent en même temps de B, alors ils arrivent en même temps sur OA. On se convainc que ceci définit bien une courbe dans le plan et que, malgré le fait que d_1 et d_2 sont confondues à l'arrivée, la "limite des P" existe. La quadratice a une description analytique, en coordonnés cartésiennes. Soient (x, y) les cordonnées de P et soit r la longueur d'un rayon du cercle. Soit θ l'angle formé par d_1 et OA. L'hypothèse dit que y/θ est constant égal à $2r/\pi$. On en déduit que la quadratrice est le lieu des points (x, y), qui satisfont $x = y \cdot \cot(\pi y/2r)$ et l'on voit que pour y tendant vers zéro, x tend vers $2r/\pi$. Ceci permet d'obtenir un segment de longueur (un multiple) de π.

(Pour d'autres informations sur ce sujet, voir E.W. Hobson, "*Squaring the circle*", Chelsea Pub. Co.)

Ce qui nous intéresse dans les développements reproduits ci-dessus est de montrer que pour attaquer le problème qui se posait à eux, les mathématiciens du passé n'ont pas hésité à l'inscrire dans une famille de problèmes plus généraux (quadrature des lunules), ou à commencer par résoudre le problème par des moyens "illicites" (emploi de la quadratrice). Ces approches sont encore parmi les plus fructueuses : il est rare qu'un problème cède après une (première) tentative de solution directe ! Souvenez-vous en.

Exercice. Lesquelles des affirmations suivantes sont fausses et lesquelles sont vraies ? Pourquoi ?

a) On ne peut pas construire de carré d'aire égale à π.
b) Peut-être qu'avec d'autres moyens que la règle et le compas on peut construire un carré d'aire égale à l'aire d'un cercle de rayon l'unité.
c) Il est possible de construire à la règle et au compas un carré dont l'aire serait égale à l'aire d'un cercle de rayon deux fois l'unité.
d) Il n'y a aucun carré construit à la règle et au compas d'aire plus petite que celle d'un cercle de rayon l'unité.
e) On peut construire un carré dont l'aire est égale à l'aire d'un cercle de rayon l'unité si on dispose d'un nombre suffisant de compas et de règles.
f) On ne peut construire à la règle et au compas un carré dont l'aire serait égale au périmètre d'un cercle de rayon l'unité.

2.6 La formule du binôme.

Pour l'instant nous avons fait des démonstrations : par épuisement, par inspection et par l'absurde. Une autre technique de démonstration très importante est celle des *démonstrations par récurrence*. Un des premiers exemples de démonstration par récurrence est celle qu'a donné Pascal de la formule du binôme. Il s'agit de l'énoncé suivant.

Théorème. Soient x et y des nombres [7]. Alors pour tout entier naturel n on a

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

où pour $k=0$ on pose $\binom{n}{0} = 1$ pour $k \geq 1$ les *coefficients binomiaux* [8] sont donnés par

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{1\cdot 2\cdots \quad k} \quad . \quad (*)$$

Démonstration. L'affirmation est vraie pour $n=0$: l'égalité se réduit à $1=1$. Par définition

$$(x+y)^n = (x+y)^{n-1}(x+y) \ .$$

Si on suppose la formule vraie pour $n-1$, le membre de gauche devient

$$\left(\sum_{k=0}^{n-1} \binom{n-1}{k} x^k y^{n-1-k}\right)(x+y)$$

ce qui s'écrit encore

$$\sum_{k=0}^{n-1} \binom{n-1}{k} (x^{k+1} y^{n-1-k} + x^k y^{n-k}) \ .$$

En décomposant la somme et en renumérotant on voit que pour vérifier la formule pour n, il suffit de montrer que pour $k \geq 1$ on a la relation

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k} \quad . \quad (**)$$

On vérifie facilement que avec (*), cette relation est satisfaite. Ceci termine la démonstration.

Exercice. Écrire la formule du binôme pour $n=-1$ et pour $n=1/2$. (Noter que les coefficients binomiaux sont bien définis par (*), même si n n'est pas entier.)

Il faut noter que si on admet la formule (*) la démonstration que nous venons de donner est essentiellement formelle. Le vrai travail a consisté à deviner cette formule. Voici comment Pascal a procédé pour la trouver. Ce qui suit donnera une nouvelle démonstration du théorème.

[7] Tout ce que l'on doit savoir faire dans le "système de nombres" auquel appartiennent x et y est multiplier, additionner et ce de manière à ce que l'on ait les identités $ab = ba$, $a+b = b+a$, $a(b+c) = ab+ac$ et $1a = a$.

[8] Les coefficients binomiaux sont souvent notés C_n^k.

La manipulation algébrique qui mène à (**) montre que les coefficients binomiaux sont déterminés par (**), c'est-à-dire que l'on peut les retrouver à l'aide de ce qu'on appelle le *triangle de Pascal* [9]

$$\begin{array}{ccccccccccccc}
 & & & & & 1 & & & & & & & \\
 & & & & 1 & & 1 & & & & & & \\
 & & & 1 & & 2 & & 1 & & & & & \\
 & & 1 & & 3 & & 3 & & 1 & & & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & & & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 & &
\end{array}$$

que l'on prolonge à l'infini. La ligne suivante, qui correspond à $n = 6$, est obtenue d'après (**) en sommant deux à deux les termes de la dernière ligne : elle commence par 1 et continue avec $6 = 1 + 5$, $15 = 5 + 10$, $20 = 10 + 10$, etc. L'idée de Pascal est la suivante : à partir de ce triangle *on construit un nouveau triangle en divisant chaque terme par le terme qui se trouve à sa gauche.* Si on n'écrit pas la première anti-diagonale ceci donne le triangle

$$\begin{array}{ccccccccccccc}
 & & & & & \frac{1}{1} & & & & & & & \\
 & & & & \frac{2}{1} & & \frac{1}{2} & & & & & & \\
 & & & \frac{3}{1} & & \frac{2}{2} & & \frac{1}{3} & & & & & \\
 & & \frac{4}{1} & & \frac{3}{2} & & \frac{2}{3} & & \frac{1}{4} & & & & \\
 & \frac{5}{1} & & \frac{4}{2} & & \frac{3}{3} & & \frac{2}{4} & & \frac{1}{5} & & &
\end{array}$$

Qui représente $\binom{n}{k} / \binom{n}{k-1}$ sur l'intersection de la n-ième ligne et de la k-ième diagonale. Pour les besoins de la suite on a simplifié quelques fractions : $\frac{3}{3} = \frac{2}{2}$, $\frac{6}{4} = \frac{3}{2}$, etc. Sur le triangle ainsi présenté on voit apparaître la loi : *sur une ligne donnée, de gauche à droite, les numérateurs des fractions décroissent et les dénominateurs croissent*! Ainsi ce qu'on est amené à montrer est l'égalité

$$\frac{\binom{n}{k}}{\binom{n}{k-1}} = \frac{n-k+1}{k} \ . \quad (***)$$

La démonstration de (***) se fait par récurrence sur le numéro de la ligne. D'abord il est clair que l'égalité est vérifiée pour les termes sur la première ligne du triangle. Or, si on représente une portion du triangle de Pascal, par

$$\begin{array}{ccccc}
A & & B & & C \\
 & D & & E &
\end{array}$$

où l'on suppose donc que la deuxième ligne est obtenue de la première par $D = A + B$, $E = B + C$, et si l'on suppose que les quotients de la première ligne satisfont à (***), c'est-à-dire

$$\frac{B}{A} = \frac{m}{l-1} \quad \text{et} \quad \frac{C}{B} = \frac{m-1}{l}$$

pour certains m et l, alors un bref calcul montre que les quotients de la deuxième ligne satisfont aussi (***), à savoir

$$\frac{E}{D} = \frac{m}{l} \ .$$

[9] Ce triangle remonte en fait beaucoup plus loin et il a été redécouvert un certain nombre de fois : Omar Alkhaijâma (Monde arabe, 1080), Tsu shi Kih (Chine 1303), Stifel (Allemagne, 1544), Cardano (Italie, 1545), Pascal (France, 1654).

Il est amusant de voir comment Pascal décrit sa démonstration (voir B. Pascal, "Traité du triangle arithmétique...", Paris, 1654) :

> *"Quoy que cette proposition ait une infinité de cas, j'en doneray une démonstration bien courte, en supposant* 2 *lemmes.*
>
> *Le 1. qui est évident de soy-mesme, que cette proportion se rencontre dans la seconde base ; car il est bien visible que* ϕ *est à* σ *comme* 1 *à* 1*.*
>
> *Le 2. que si cette proportion se trouve dans une base quelconque, elle se trouvera nécessairement dans la base suivante."*

Ici, la "base" est la ligne du triangle de Pascal, et la "proportion" est le quotient $\binom{n}{k} / \binom{n}{k-1}$ [10].

Corollaire. Le coefficient binomial $\binom{n}{k}$ égale le nombre de façons de choisir k objets parmi n.

Démonstration. Si on écrit $(x+y)^n$ comme le produit de n termes $(x+y)$ on voit que le coefficient de $x^{n-k}y^k$ est bien égal au nombre de possibilités de choisir simultanément k fois y dans les n termes du produit.

2.7 Les Éléments d'Euclide d'Alexandrie.

Pendant des siècles un unique livre de mathématiques a servi de modèle pour tous les autres : "Les Éléments", rédigés par Euclide d'Alexandrie. Galilée n'utilisait essentiellement que des mathématiques issues des Éléments, et même Newton a rédigé les "Principes mathématiques de philosophie naturelle" en 1687, en se basant sur les Éléments.

Il s'agit d'un texte dont la lecture est encore intéressante de nos jours, qui a inspiré et inspire encore de nombreux mathématiciens, ne serait ce que parce qu'il peut servir pour mesurer les progrès effectués depuis sa rédaction. Les Éléments traitent de géométrie plane en treize Livres, qui occupent (au moins) trois volumes de format moderne. Ce qui est remarquable, et qui a fait que pendant très longtemps ils n'ont pas été égalés, est le style de présentation. En effet les éléments sont l'exemple par excellence d'un *traitement axiomatique* d'une théorie. Ce type de traitement permet de dégager les principes sur lesquels se fonde la théorie et permet d'en mettre en évidence les éventuelles lacunes. Vu le degré de sophistication des mathématiques Grecques, et la quantité de résultats déjà obtenus, cette présentation axiomatique est un vrai tour de force. Nous allons voir la complexité logique du traité sur un exemple. Il faut voir, que les Éléments n'épuisent pas du tout les connaissances mathématiques des Grecs. Il semblerait que cet ouvrage servait à l'époque comme un ouvrage destinés aux étudiants de l'Académie d'Alexandrie. [11] Pour avoir une image plus complète des mathématiques Grecques il faudrait aussi passer en revue les travaux en astronomie, statique et mécanique, ainsi que les œuvres d'Archimède, Apollonius, etc.

Néanmoins, les Éléments abordent des sujets primordiaux pour la compréhension de ces autres travaux et aussi pour la compréhension du coeur de ce cours : que ce soit l'aspect formel des mathématiques ou les notions de nombre et de limite ! En effet, il faut comprendre pourquoi Euclide ne fonde pas son

[10] La démonstration de Pascal, ainsi que beaucoup d'autres démonstrations "d'origine", sont reprises dans l'ouvrage très intéressant de E. Hairer et G. Wanner *Analysis by its history.* Undergraduate Texts in Mathematics. Readings in Mathematics. Springer-Verlag, New York, 1996, ISBN 0-387-94551-2. Existe aussi en Français.

[11] On pense d'ailleurs que les Éléments puissent être un ouvrage collectif, rédigé par plusieurs auteurs. En tout cas on a très peu d'informations sur Euclide d'Alexandrie et il est sûr que nombreux résultats du traité étaient bien connus avant sa rédaction.

étude sur la notion de nombre, telle que nous l'entendons aujourd'hui. [12] Un nombre est pour lui la mesure d'une grandeur (donc positif). "Un nombre est une multitude composée d'unités", dit-il ; donc l'unité n'est pas un nombre...[13] Malgré ça, Euclide développe un puissant calcul avec les segments, qui donne des bases suffisamment larges et solides pour ses besoins. Son traité contient aussi la méthode d'exhaustion, qui est un précurseur du concept moderne de limite, et qu'il utilise par exemple pour calculer le volume d'une pyramide.

Les Éléments présentent une théorie des proportions, qui permet de faire des calculs très complexes, et qui a commencé à se révéler insuffisante seulement lorsque Galilei et d'autres ont eu besoin de définir des notions comme la vitesse instantanée. On peut dire, que c'est seulement avec les définitions des nombres réels de Cauchy, Cantor et Dedekind vers la fin du 19ème siècle, que l'on a pu combler ces insuffisances.

Par ailleurs, le fait que les Éléments ne s'appuient pas sur une théorie des nombres réels, mais plutôt sur l'étude systématique des conséquences d'un ensemble restreint d'axiomes, leur donne une généralité plus grande et a permis d'arriver à découvrir d'autres types de géométries, qui se sont révélées très utiles en physique (voir le texte de Poincaré dans le paragraphe qui suit).

Nous reproduisons ci-après le début du Livre I, des Éléments, ainsi que quelques propositions à titre d'exemple. L'ouvrage commence par une liste de Définitions, de Demandes et de Notions communes. Quelques autres définitions sont données plus loin dans le texte, mais essentiellement tous les résultats sont des conséquences des quelques postulats énoncés ci-dessous !

Début du Livre I. Traduction de Bernard Vitrac, dans Bibliothèque d'Histoire des Sciences, Presses Universitaires de France.

Définitions.

1) Un *point* est ce dont il n'y a aucune partie.
2) Une *ligne* est une longueur sans largeur.
3) Les *limites d'une ligne* sont des points.
4) Une *ligne droite* est celle qui est placée de manière égale par rapport aux points qui sont sur elle.
5) Une *surface* est ce qui a seulement longueur et largeur.
6) Les *limites d'une surface* sont des lignes.
7) Une *surface plane* est celle qui est placée de manière égale par rapport aux droites qui sont sur elle.
8) Un *angle plan* est l'inclinaison, l'une sur l'autre, dans un plan, de deux lignes qui se touchent l'une sur l'autre et ne sont pas placées en ligne droite.
9) Et quand les lignes contenant l'angle sont droites, l'angle est appelé *rectiligne*.
10) Et quand une droite, ayant été élevée sur une droite, fait les angles adjacents égaux entre eux, chacun de ces angles égaux est *droit*, et la droite qui a été élevée est appelée *perpendiculaire* à celle sur laquelle elle a été élevée.
11) Un angle *obtus* est celui qui est plus grand qu'un droit.
12) Un angle *aigu* est plus petit qu'un droit.
13) Une *frontière* est ce qui est limite de quelque chose.
14) Une *figure* est ce qui est contenu par quelque ou quelques frontière(s).

[12] On a dit qu'à cause du fait qu'ils avaient découvert des segments incommensurables, les Grecs pensaient ne plus pouvoir fonder les mathématiques et la description du monde sur les nombres (entiers), comme voulaient le faire les pythagoriciens.

[13] Noter que si on pense aux nombres naturels comme alignés sur une demi-droite, alors tous ces nombres, sauf l'unité, apparaissent comme la moyenne (arithmétique) des deux nombres à leur gauche et à leur droite : par exemple 4 est la moyenne de 3 et de 5 ; à gauche de l'unité il devrait y avoir zéro, qui n'était pas considéré un nombre.

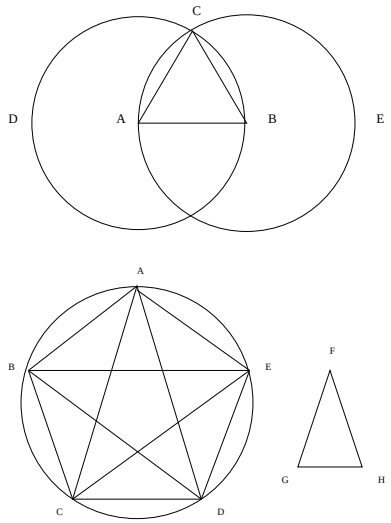
15) Un *cercle* est une figure plane contenue par une ligne unique { celle appelée circonférence } par rapport à laquelle toutes les droites menées à sa rencontre à partir d'un unique point parmi ceux qui sont placés à l'intérieur de la figure, sont { jusqu'à la circonférence du cercle } égales entre elles.
16) Et le point est appelé *centre* du cercle.
17) Et un *diamètre* du cercle est n'importe quelle droite menée par le centre, limitée de chaque côté par la circonférence du cercle, laquelle coupe le cercle en deux parties égales.
18) Un *demi-cercle* est la figure contenue par le diamètre et la circonférence découpée par lui ; le centre du demi-cercle est le même que celui du cercle.
19) Les *figures rectilignes* sont les figures contenues par les droites ; *trilatères* : celles qui sont contenues par trois droites, *quadrilatères* par quatre ; *multilatères* par plus de quatre.
20) Parmi les figures trilatères est un *triangle équilatéral* celle qui a les trois côtés égaux ; *isocèle* celle qui a deux côtés égaux seulement ; *scalène* celle qui a les trois côtés inégaux.
21) De plus, parmi les figures trilatères est un triangle *rectangle* celle qui a un angle droit ; *obtusangle*, celle qui a un angle obtus ; *acutangle*, celle qui a les trois angles aigus.
22) Parmi les figures quadrilatères est un *carré* celle qui est à la fois équilatérale et rectangle ; est *oblongue* celle qui est rectangle mais non équilatérale ; un *losange*, celle qui est équilatérale mais non rectangle ; un *rhomboïde*, celle qui a les côtés et les angles opposés égaux les uns aux autres mais qui n'est ni équilatérale ni rectangle ; et que l'on appelle *trapèzes* les quadrilatères autres que ceux-là.
23) Des droites *parallèles* sont celles qui étant dans le même plan et indéfiniment prolongées de part et d'autre, ne se rencontrent pas, ni d'un côté ni de l'autre.

Demandes.
1) Qu'il soit demandé de mener une ligne droite de tout point à tout point.
2) Et de prolonger continûment en ligne droite une ligne droite limitée.
3) Et de décrire un cercle à partir de tout centre et au moyen de tout intervalle.
4) Et que tous les angles droits soient égaux entre eux.
5) Et que, si une droite tombant sur deux droites faits les angles intérieurs et du même côté plus petits que deux droits, les deux droites, indéfiniment prolongées, se rencontrent du côté où sont les angles plus petits que deux droits.

Notions communes.
1) Les choses égales à une même chose sont aussi égales entre elles.
2) Et si, à partir de choses égales, des choses égales sont ajoutées, les touts sont égaux.
3) Et si, à partir de choses égales, des choses égales sont retranchées, les restes sont égaux.
4) { Et si, à des choses inégales, des choses égales sont ajoutées, les touts sont inégaux.
5) Et les doubles du même sont égaux entre eux.
6) Et les moitiés du même sont égales entre elles. }
7) Et les choses qui s'ajustent les unes sur les autres sont égales entre elles.

Proposition 1.
Sur une droite limitée donnée, construire un triangle équilatéral.
Livre IV. Proposition 11.
Dans un cercle donné, inscrire un pentagone équilatéral et équiangle.

Livre VII. Définitions.
1) Est *unité* ce selon quoi chacune des choses existantes est dite une.
2) Et un *nombre* est la multitude composée d'unités.
3) ...



Les démonstrations sont souvent accompagnées par des figures, comme celles ci-dessus. Le texte énonce clairement comment déduire les énoncés des postulats ou des propositions précédentes. Ainsi la Prop. I.1 repose sur l'utilisation successive de la Déf. 20, de la Dem. 3, de la Dem. 1, des Déf. 15/16, et de la NC 1. La démonstration de la Prop. IV.11 est déjà beaucoup plus compliquée. Elle nécessite l'emploi de la Dem. 1, des NC 1, 2 et 6, ainsi que des Prop. I.9, III.26, 27, 29, IV.2, 10. Évidemment ces propositions ont aussi des démonstrations qui dépendent de plein d'autres résultats antérieurs. Par exemple la Prop. IV.2 dépend de la Dem. 1, des NC 1, 3 et des Prop. I.23, 32, III.16 (Por.), 32. On voit bien la complexité de la construction logique.

Exercice. On se propose de démontrer comment faire coïncider la partie d'une demi-droite AB à un segment C donné, sans utiliser un instrument transporteur de segments.

Pour cela on va d'abord démontrer la Proposition I.2 des Éléments, qui dit : *donné un segment BC et un point A, on se propose de tracer un segment égal à BC dont une extrémité est A.*

Indications pour la démonstration de la Proposition I.2 :

- utiliser la Proposition I.1 pour construire un triangle équilatéral ABD sur AB ;
- prolonger DB et DA en deux droites ;
- construire le cercle de centre B par C, et appeler G l'intersection de ce cercle avec le prolongement de DB ;
- construire le cercle de centre D par G, et appeler L l'intersection de ce cercle avec le prolongement de DA ;
- conclure.

Appliquer la Proposition 2 pour avoir AD égal à C avec extrémité A, puis construire le cercle de centre A par D, qui coupera AB en E ; la partie AE de la demi-droite répond à la demande. (Faire des dessins !)

2.8 Les géométries non euclidiennes.

Vers la fin du 19ème siècle il est apparu que l'on pouvait développer sur le modèle euclidien des géométries, qui gardaient certaines des caractéristiques de la géométrie euclidienne, mais pas toutes.

De plus, on pouvait même démontrer que ces nouvelles géométries étaient tout aussi consistantes que la géométrie euclidienne, c'est-à-dire que si l'une d'entre elles se révélait contenir des énoncés contradictoires, on aurait alors pu en trouver aussi dans la géométrie euclidienne. La possibilité de ces géométries non euclidiennes a amené les mathématiciens et les philosophes à se poser la question de savoir laquelle des géométries connues était la "vraie" géométrie, celle qui décrit le monde dans lequel nous vivons. Il était difficile d'admettre que l'espace était lui aussi redevable de plusieurs représentations non équivalentes. On avait l'habitude de penser l'espace immuable support des phénomènes et donc unique. Les développements de la physique du début du 20ème siècle ont montré tout le profit que l'on pouvait tirer de ces nouvelles théories géométriques : la relativité se base sur une représentation homogène de l'espace et du temps dans un espace (!) à quatre dimensions, où les distances sont mesurées suivant des règles différentes de celles que l'on déduirait simplement en généralisant celles de l'espace euclidien à trois dimension.

Comme cela arrive souvent, on n'a pas abouti à ces résultats révolutionnaires parce qu'on "voulait faire la révolution". Au contraire, c'est en voulant mieux comprendre leur héritage mathématique, et plus précisément le choix d'axiomes dans les Éléments d'Euclide, que divers mathématiciens parmi lesquels d'abord Saccheri, puis au 19ème siècle Bolyai, Gauss, Hilbert, Lobatchevsky et Riemann en sont arrivés à se voir forcés d'admettre que d'autres géométries étaient possibles.

Pour présenter succinctement ces géométries nous reproduisons ci-après un extrait de "La science et l'hypothèse", de H. Poincaré (1854-1912). Éminent mathématicien et physicien, avec ce premier livre philosophique publié en 1902 Poincaré a contribué de manière significative au débat passionné et passionnant de la philosophie des sciences du début du 20ème siècle autour des fondements des sciences. (Nous utilisons l'édition en livre de poche dans la collection Champs, Flammarion de 1968.)

Chapitre III : "Toute conclusion suppose des prémisses ; ces prémisses elles-mêmes ou bien sont évidentes par elles-mêmes et n'ont pas besoin de démonstration, ou bien ne peuvent être établies qu'en s'appuyant sur d'autres propositions, et comme on ne saurait remonter ainsi à l'infini, toute science déductive, et en particulier la géométrie, doit reposer sur un certain nombre d'axiomes indémontrables. Tous les traités de géométrie débutent donc par l'énoncé de ces axiomes. Mais il y a entre eux une distinction à faire : quelques-uns, comme celui-ci par exemple : "deux quantités égales à une même troisième sont égales entre elles", ne sont pas des propositions d'analyse. Je les regarde comme des jugements analytiques *a priori*, je ne m'en occuperai pas.

Mais je dois insister sur d'autres axiomes qui sont spéciaux à la géométrie. La plupart des traités en énoncent trois explicitement[14] :

1° Par deux points ne peut passer qu'une droite ;

2° La ligne droite est le plus court chemin d'un point à un autre ;

3° Par un point on ne peut faire passer qu'une parallèle à une droite donnée.

Bien que l'on se dispense généralement de démontrer le second de ces axiomes, il serait possible de le déduire des deux autres et de ceux, beaucoup plus nombreux, que l'on admet implicitement sans les énoncer. [...]

On a longtemps cherché en vain à démontrer également le troisième axiome, connu sous le nom de *postulat d'Euclide*. Ce qu'on a dépensé d'efforts dans cet espoir chimérique est vraiment inimaginable. Enfin au commencement du siècle et à peu près en même temps, deux savants, un Russe et un Hongrois, Lobatchevsky et Bolyai établirent d'une façon irréfutable que cette démonstration est impossible ; ils nous ont à peu près débarrassés des inventeurs de géométries sans postulat ; depuis lors l'Académie des Sciences ne reçoit guère qu'une ou deux démonstrations nouvelles par an[15]. [...]

[14]Parmi les axiomes qui suivent, le deuxième est la Prop. I.20 des Éléments et le troisième est étroitement lié au cinquième postulat d'Euclide.

[15]Il existe encore bon nombre des problèmes ouverts fameux auxquels s'attaquent des amateurs passionnés, qui n'hésitent pas à soumettre leurs solutions aux instances "officielles". A titre d'exemple, il arrive assez souvent, que l'on demande à des membres de l'Institut de mathématiques de Bordeaux de vérifier des écrits tendant à prouver une

La Géométrie de Lobatchevsky.– S'il était possible de déduire le postulat d'Euclide des autres axiomes, il arriverait évidemment qu'en niant le postulat, et en admettant les autres axiomes, on serait conduit à des conséquences contradictoires ; il serait donc impossible d'appuyer sur de telles prémisses une géométrie cohérente.

Or c'est précisément ce qu'a fait Lobatchevsky. Il suppose au début que :

L'on peut par un point mener plusieurs parallèles à une droite donnée.

Et il conserve d'ailleurs tous les autres axiomes d'Euclide. De ces hypothèses, il déduit une suite de théorèmes entre lesquels il est impossible de relever aucune contradiction et il construit une géométrie dont l'impeccable logique ne le cède en rien à celle de la géométrie euclidienne.

Les théorèmes sont bien entendu, très différents de ceux auxquels nous sommes accoutumés et ils ne laissent pas de déconcerter un peu d'abord.

Ainsi la somme des angles d'un triangle est toujours plus petite que deux droits et la différence entre cette somme et deux droits est proportionnelle à la surface du triangle.

Il est impossible de construire une figure semblable à une figure donnée mais de dimensions différentes. [...]

Il est inutile de multiplier ces exemples ; les propositions de Lobatchevsky n'ont plus aucun rapport avec celles d'Euclide, mais elles ne sont pas moins logiquement reliées les unes aux autres.

La Géométrie de Riemann. – Imaginons un monde uniquement peuplé d'êtres dénués d'épaisseur ; et supposons que ces animaux "infiniment plats" soient tous dans un même plan et n'en puissent sortir. Admettons de plus que ce monde soit assez éloigné des autres pour être soustrait à leur influence. Pendant que nous sommes en train de faire des hypothèses, il ne nous coûte pas plus de douer ces êtres de raisonnement et de les croire capables de faire de la géométrie. Dans ce cas, ils n'attribueront certainement à l'espace que deux dimensions.

Mais supposons maintenant que ces animaux imaginaires, tout en restant dénués d'épaisseur, aient la forme d'une figure sphérique, et non d'une figure plane et soient tous sur une même sphère sans pouvoir s'en écarter. Quelle géométrie pourront-ils construire ? Il est clair d'abord qu'ils n'attribueront à l'espace que deux dimensions ; ce qui jouera pour eux le rôle de la ligne droite, ce sera le plus court chemin d'un point à un autre sur la sphère, c'est-à-dire un arc de grand cercle, en un mot leur géométrie sera la géométrie sphérique.

Ce qu'ils appelleront l'espace, ce sera cette sphère d'où ils ne peuvent sortir et sur laquelle se passent tous les phénomènes dont ils peuvent avoir connaissance. Leur espace sera donc *sans limites* puisqu'on peut sur la sphère aller toujours devant soi sans jamais être arrêté, et cependant il sera *fini* ; on n'en trouvera jamais le bout, mais on pourra en faire le tour.

Eh bien la géométrie de Riemann, c'est la géométrie sphérique étendue à trois dimensions. Pour la construire, le mathématicien allemand a dû jeter par-dessus bord, non seulement le postulat d'Euclide, mais encore le premier axiome : *Par deux points on ne peut faire passer qu'une droite.*

Sur une sphère, par deux points donnés on ne peut *en général* passer qu'un grand cercle (qui, comme nous venons de le voir, jouerait le rôle de la droite pour nos êtres imaginaires), mais il y a une exception : si les deux points donnés sont diamétralement opposés, on pourra faire passer par ces deux points une infinité de grands cercles.

De même dans la géométrie de Riemann (au moins sous une de ces formes), par deux points ne passera en général qu'une seule droite ; mais il y a des cas exceptionnels où par deux points pourront passer une infinité de droites.

Il y a une sorte d'opposition entre la géométrie de Riemann et celle de Lobatchevsky.

Ainsi la somme des angles d'un triangle est :

- Égale à deux droits dans la géométrie d'Euclide.
- Plus petite que deux droits dans celle de Lobatchevsky.

---

des conjectures de théorie des nombres, comme le "problème des premiers jumeaux", l'Hypothèse de Riemann ou encore–malgré sa démonstration récente–le Grand théorème de Fermat.

- Plus grande que deux droits dans celle de Riemann.

Le nombre de parallèles qu'on peut mener à une droite donnée par un point donné est égal :

- À un dans la géométrie d'Euclide.
- À zéro dans celle de Riemann.
- À l'infini dans celle de Lobatchevsky.

Ajoutons que l'espace de Riemann est fini, quoique sans limite, au sens donné plus haut à ces deux mots."

Puis, Poincaré explique rapidement comment l'on peut démontrer que les théorèmes de Lobatchevsky et de Riemann ne présentent aucune contradiction. En gros il s'agit de réaliser des modèles euclidiens des nouvelles géométries (cela se comprend assez aisément pour la géométrie sphérique, où la sphère est la sphère usuelle, euclidienne). Ensuite, il pose la question de savoir si "les axiomes explicitement énoncés dans les traité sont les seuls fondements de la géométrie ?", pour répondre immédiatement que l'"on peut être assuré du contraire en voyant qu'après les avoir successivement abandonnés on laisse encore debout quelques propositions communes aux théories d'Euclide, de Lobatchevsky et de Riemann". A titre d'exemple, Poincaré met en évidence le fait que "la possibilité du mouvement d'une figure invariable n'est pas une vérité évidente en elle-même", et il indique comment, en explicitant un axiome qui rendrait un tel mouvement possible, on pourrait donner une définition non défectueuse d'une droite : "Il peut arriver que le mouvement d'une figure invariable soit tel que tous les points d'une ligne appartenant à cette figure restent immobiles pendant que tous les points situés en dehors de cette ligne se meuvent. Une pareille ligne s'appellera une ligne droite." Poincaré souligne encore que, d'après un théorème de Lie, le nombre de géométries sur un espace de dimension donnée n, dans lesquelles le mouvement d'une figure invariable est possible, pour lesquelles il faut un nombre fini p de conditions pour déterminer la position de cette figure dans l'espace, est limité par un nombre qui ne dépend que de n et de p.

Pour terminer le chapitre, Poincaré se tourne vers un ultérieur type de géométrie, avant de conclure avec des considérations philosophiques. Il écrit :

"Les Géométries de Hilbert. – Enfin M. Veronese et M. Hilbert ont imaginé de nouvelles géométries plus étranges encore, qu'ils appellent *non-archimédiennes*. Ils les construisent en rejetant l'*axiome d'Archimède* en vertu duquel toute longueur donnée, multipliée par un entier suffisamment grand, finira par surpasser toute autre longueur donnée si grande qu'elle soit. Sur une droite non archimédienne, les points de notre géométrie ordinaire existent tous, mais il y en a une infinité d'autres qui viennent s'intercaler entre eux, de telle sorte qu'entre deux segments, que les géomètres de la vieille école auraient regardés comme contigus, on puisse caser une infinité de points nouveaux. [...]

De la nature des axiomes. – La plupart des mathématiciens ne regardent la géométrie de Lobatchevsky que comme une simple curiosité logique ; quelques-uns d'entre eux sont allés plus loin cependant. Puisque plusieurs géométries sont possibles, est-il certain que ce soit la nôtre qui soit vraie ? L'expérience nous apprend sans doute que la somme des angles d'un triangle est égale à deux droits ; mais c'est seulement parce que nous n'opérons que sur des triangles trop petits ; la différence, d'après Lobatchevsky, est proportionnelle à la surface du triangle : ne pourra-t-elle devenir sensible quand nous opérons sur des triangles plus grands ou quand nos mesures deviendront plus précises[16] ? La géométrie euclidienne ne serait ainsi qu'une géométrie provisoire.

Pour discuter cette opinion, nous devons d'abord nous demander quelle est la nature des axiomes géométriques.

Sont-ce des jugements synthétiques *a priori*, comme disait Kant ?

[16]Il est intéressant d'observer que Gauss a effectivement essayé de vérifier par la mesure directe sur un grand triangle formé par les cimes de trois montagnes, si la somme de ses angles était vraiment égale à deux droits. Les erreurs implicites dans les mesures ont de fait rendu la vérification impossible.

Ils s'imposeraient alors à nous avec une telle force, que nous ne pourrions concevoir la proposition contraire, ni bâtir sur elle un édifice théorique. Il n'y aurait pas de de géométrie non euclidienne.

Pour s'en convaincre, qu'on prenne un véritable jugement synthétique *a priori*, par exemple celui-ci [qui joue un] rôle prépondérant :

Si un théorème est vrai pour le nombre 1*, si on a démontré qu'il est vrai de* $n+1$*, pourvu qu'il le soit de* n*, il sera vrai de tous les nombres entiers positifs.*

Qu'on essaie ensuite de s'y soustraire et de fonder, en niant cette proposition, une fausse arithmétique analogue à la géométrie non euclidienne, – on n'y pourra pas parvenir ; on serait même tenté au premier abord de regarder ces jugements comme analytiques.

D'ailleurs, reprenons notre fiction des animaux sans épaisseur ; nous ne pouvons guère admettre que ces êtres, s'ils ont l'esprit fait comme nous, adopteraient la géométrie euclidienne qui serait contredite par toute leur expérience ?

Devons-nous donc conclure que les axiomes de la géométrie sont des vérités expérimentales ? Mais on n'expérimente pas sur des droites ou des circonférences idéales ; on ne peut le faire que sur des objets matériels. Sur quoi porteraient donc les expériences qui serviraient de fondement à la géométrie ? La réponse est facile.

Nous avons vu plus haut que l'on raisonne constamment comme si les figures géométriques se comportaient à la manière des solides. Ce que la géométrie emprunterait à l'expérience, ce seraient donc les propriétés de ces corps.

Les propriétés de la lumière et sa propagation rectiligne ont été aussi l'occasion d'où sont sorties quelques-unes des propositions de la géométrie, et en particulier celles de la géométrie projective, de sorte qu'à ce point de vue on serait tenté de dire que la géométrie métrique est l'étude des solides et que la géométrie projective est celle de la lumière.

Mais une difficulté subsiste, et elle est insurmontable. Si la géométrie était une science expérimentale, elle ne serait pas une science exacte, elle serait soumise à une continuelle révision. Que dis-je ? elle serait dès aujourd'hui convaincue d'erreur puisque nous savons qu'il n'existe pas de solide rigoureusement invariable.

Les axiomes géométriques ne sont donc ni des jugements synthétiques a priori *ni des faits expérimentaux.*

Ce sont des *conventions ;* notre choix, parmi toutes les conventions possibles, est *guidé* par des faits expérimentaux ; mais il reste *libre* et n'est limité que par la nécessité d'éviter toute contradiction. C'est ainsi que les postulats peuvent rester *rigoureusement* vrais quand même les lois expérimentales qui ont déterminé leur adoption ne sont qu'approximatives.

En d'autre termes, *les axiomes de la géométrie* (je ne parle pas de ceux de l'arithmétique) *ne sont que des définitions déguisées.*

Dès lors, que doit-on penser de cette question : La géométrie euclidienne est-elle vraie ?

Elle n'a aucun sens.

Autant demander si le système métrique est vrai et les anciennes mesures fausses ; si les coordonnées cartésiennes sont vraies et les coordonnées polaires fausses. Une géométrie ne peut pas être plus vraie qu'une autre ; elle peut seulement être *plus commode.*

Or la géométrie euclidienne est et restera la plus commode :

1° Parce qu'elle est la plus simple ; et elle n'est pas telle seulement par suite de nos habitudes d'esprit ou de je ne sais quelle intuition directe que nous aurions de l'espace euclidien ; elle est la plus simple en soi de même qu'un polynôme du premier degré est plus simple qu'un polynôme du second degré ; les formules de la trigonométrie sphérique sont plus compliquées que celles de la trigonométrie rectiligne, et elles paraîtraient encore telles à un analyste qui en ignorerait la signification géométrique.

2° Parce qu'elle s'accorde assez bien avec les propriétés des solides naturels, ces corps dont se rapprochent nos membres et notre œil et avec lesquels nous faisons nos instruments de mesure."

En reproduisant cette partie philosophique nous ne disons pas que nous adhérons aux conclusions auxquelles arrive Poincaré, mais nous pensons que celles-ci sont suffisamment intéressantes pour que tous ceux qui s'intéresseraient à la question de la nature des axiomes de la géométrie en aient connaissance.

Plus loin dans son ouvrage, au chapitre suivant, Poincaré essaie d'imaginer un monde non euclidien, qui aurait amené ses habitants à adopter des conventions différentes de celles que nous avons adoptées et qui auraient conduit à une géométrie non euclidienne. Cette invention montre bien qu'une telle géométrie n'est pas si exotique qu'elle apparaît au premier abord.

"Le monde non euclidien. Si l'espace géométrique était un cadre imposé à *chacune* de nos représentations, considérées individuellement, il serait impossible de se représenter une image dépouillée de ce cadre, et nous ne pourrions rien changer à notre géométrie.

Mais il n'en est pas ainsi, la géométrie n'est que le résumé des lois suivant lesquelles *se succèdent* ces images. Rien n'empêche alors d'imaginer une série de représentations, de tout point semblables à nos représentations ordinaires, mais se succédant d'après des lois différentes de celles auxquelles nous sommes accoutumés.

On conçoit alors que des êtres dont l'éducation se ferait dans un milieu où ces lois seraient ainsi bouleversées pourraient avoir une géométrie très différente de la nôtre.

Supposons, par exemple, un monde renfermé dans une grande sphère et soumis aux lois suivantes :

La température n'y est pas uniforme ; elle est maxima au centre, et elle diminue à mesure qu'on s'en éloigne, pour se réduire au zéro absolu quand on atteint la sphère où ce monde est renfermé.

Je précise davantage la loi suivant laquelle varie cette température. Soit R le rayon de la sphère limite ; soit r la distance du point considéré au centre de cette sphère. La température absolue sera proportionnelle à $R^2 - r^2$.

Je supposerai de plus que, dans ce monde, tous les corps aient même coefficient de dilatation, de telle façon que la longueur d'une règle quelconque soit proportionnelle à la température absolue.

Je supposerai enfin qu'un objet transporté d'un point à un autre, dont la température est différente, se met immédiatement en équilibre calorifique avec son nouveau milieu.

Rien dans ces hypothèses n'est contradictoire ou inimaginable.

Un objet mobile deviendra alors de plus en plus petit à mesure qu'on se rapproche de la sphère limite.

Observons d'abord que, si ce monde est limité au point de vue de notre géométrie habituelle, il paraîtra infini à ses habitants.

Quand ceux-ci, en effet, veulent se rapprocher de la sphère limite, ils se refroidissent et deviennent de plus en plus petits. Les pas qu'ils font sont donc aussi de plus en plus petits, de sorte qu'ils ne peuvent jamais atteindre la sphère limite.

Si, pour nous, la géométrie n'est que l'étude des lois suivant lesquelles se meuvent les solides invariables, pour ces êtres imaginaires, ce sera l'étude des lois suivant lesquelles se meuvent les solides *déformés par ces différences de température* dont je viens de parler.

Sans doute, dans notre monde, les solides naturels éprouvent également des variations de forme et de volume dues à l'échauffement ou au refroidissement. Mais nous négligeons ces variations en jetant les fondements de la géométrie ; car, outre qu'elles sont très faibles, elles sont irrégulières et nous paraissent par conséquent accidentelles.

Dans ce monde hypothétique, il n'en serait plus de même, et ces variations suivraient des lois régulières et très simples.

D'autre part, les diverses pièces solides dont se composerait le corps des habitants, subiraient les mêmes variations de forme et de volume.

Je ferai encore une autre hypothèse ; je supposerai que la lumière traverse des milieux diversement réfringents et de telle sorte que l'indice de réfraction soit inversement proportionnel à $R^2 - r^2$. Il est aisé de voir, que, dans ces conditions, les rayons lumineux ne seraient pas rectilignes, mais circulaires.

Pour justifier ce qui précède, il me reste à montrer que certains changements survenus dans la position des objets extérieurs peuvent être *corrigés* par des mouvements corrélatifs des êtres sentant qui habitent ce monde imaginaire ; et cela de façon à restaurer l'ensemble primitif des impressions subies par ces êtres sentant.

Supposons qu'un objet se déplace, en se déformant, non comme un solide invariable, mais comme un solide éprouvant des dilatations inégales exactement conformes à la loi de température que j'ai supposée plus haut. Qu'on me permette pour abréger le langage, d'appeler un pareil mouvement *déplacement non euclidien.*

Si un être sentant se trouve dans le voisinage, ses impressions seront modifiées par le déplacement de l'objet, mais il pourra les rétablir en se mouvant lui-même d'une manière convenable. Il suffit que l'ensemble de l'objet et de l'être sentant, considéré comme formant un seul corps, ait éprouvé un de ces déplacements particuliers que je viens d'appeler non euclidiens. Cela est possible si l'on suppose que les membres de ces êtres se dilatent d'après la même loi que les autres corps du monde qu'ils habitent.

Bien qu'au point de vue de notre géométrie habituelle les corps se soient déformés dans ce déplacement et que leur diverses parties ne se retrouvent plus dans la même situation relative, cependant nous allons voir que les impressions de l'être sentant sont redevenues les mêmes.

En effet, si les distances mutuelles des diverses parties ont pu varier, néanmoins les parties primitivement en contact sont revenues en contact. Les impressions tactiles n'ont donc pas changé.

D'autre part, en tenant compte de l'hypothèse faite plus haut au sujet de la réfraction et de la courbure des rayons lumineux, les impressions visuelles seront aussi restées les mêmes.

Ces êtres imaginaires seront donc comme nous conduits à classer les phénomènes dont ils seront témoins et à distinguer parmi eux, "les changements de position" susceptibles d'être corrigés par un changement volontaire corrélatif.

S'ils fondent une géométrie, ce ne sera pas comme la nôtre, l'étude des mouvements de nos solides invariables ; ce sera celle des changements de position qu'ils auront ainsi distingués, et qui ne sont autres que les "déplacements non euclidiens", *ce sera la géométrie non euclidienne.*

Ainsi des êtres comme nous, dont l'éducation se ferait dans un pareil monde, n'auraient pas la même géométrie que nous."

2.9 On peut construire une courbe continue qui passe par tous les points d'un carré.

Les définitions modernes des nombres réels sont toutes basées sur la notion d'approximation rationnelle. C'est-à-dire que l'on part de la donnée des nombres rationnels et on définit un nombre réel comme étant une suite de rationnels particulière. Dans ce cours, nous allons définir un nombre réel comme étant un développement décimal illimité

$$a_0, a_1 a_2 a_3 \dots \ ,$$

avec a_0 entier et $0 \leq a_i \leq 9$. Ce développement doit être compris comme donnant une suite de rationnels

$$a_0,\ a_0 + a_1/10,\ a_0 + a_1/10 + a_2/10^2,\ \dots \ ,$$

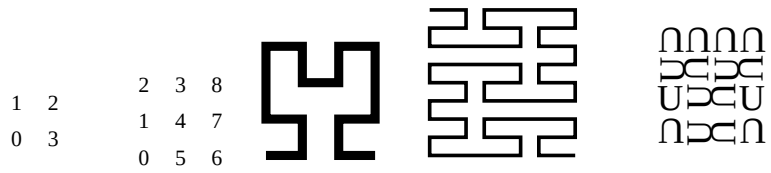
qui l'approximent aussi bien que l'on veut. Une fois que nous aurons défini les nombres réels, que nous aurons appris à les additionner, à les multiplier et que nous aurons mis en évidence leurs propriétés les plus importantes, nous allons passer à l'étude des fonctions de la variable réelle. Dans cette étude la notion de limite joue un rôle central. En particulier nous allons dire, qu'une fonction f est *continue* en un point a, où elle est définie, si pour tout nombre k' de décimales de l'image $b = f(a)$ de a par f il existe un entier k tel que si x est un réel ayant (au moins) k décimales en commun avec a, alors $f(x)$ a (au moins) k' décimales en commun avec b. On se convainc assez facilement que cette définition traduit bien la notion intuitive de continuité.

Mais, la notion de continuité d'une fonction est assez subtile. En 1890, G. Peano, mathématicien à Turin, publie dans un des plus prestigieux journaux de l'époque un court article dans lequel il montre comment définir une courbe continue qui remplit un carré ! On savait, d'après les travaux de G. Cantor, que d'un point de vue ensembliste un segment et un carré ne pouvaient pas être distingués : les deux ensembles ont le même nombre d'éléments, ou, comme on dit de manière plus précise, il existe une application bijective entre le segment et le carré. La courbe de Peano n'est pas une application bijective, mais elle fait l'essentiel : elle passe par tous les points du carré et même on nombre (fini) de fois par certains de ces points ! Le fait que la courbe soit continue est tout à fait remarquable et a obligé les mathématiciens à revoir la notion de dimension. Nous reproduisons ci-dessous le texte original de Peano, qui a été rédigé en Français. Puis nous présentons une construction géométrique due à D. Hilbert permettant de retrouver les résultats de Peano. Le travail de Hilbert, est paru un an après celui de Peano dans la même revue.

Peano est un personnage attachant, qui a contribué de manière significative aux mathématiques de son temps : en théorie des ensembles (Axiomatique de Peano pour les entiers naturels), avec des travaux sur les équations différentielles (conditions générales d'existence de solutions), en algèbre (avec la première définition axiomatique de la notion d'espace vectoriel), ... De plus il s'est intéressé à l'enseignement primaire et a développé une langue, qui aurait dû devenir une langue universelle (comme l'espéranto) : le "latino sine flexione". En fait, à cause de sa croyance exagérée (?) dans les vertus de sa nouvelle langue il a été moins lu, qu'il n'aurait dû l'être, car beaucoup de ses ouvrages ont été rédigés dans cette langue.

Nous commençons par un extrait du "Formulario Matematico", datant de 1908, où Peano décrit sa découverte.

> *Existe complexo de ordine* n*, vel puncto in spatio ad* n *dimensiones, functio continuo de variabile reale, vel tempore, tale que trajectoria de puncto mobile ple toto spatio.*
>
> *Nos pone quadratos partiale, ut illo fi adjacente. In basi* 2 *de numeratione, nos sume* 4 *quadratos partiale in ordine ut in figura* (a)*, et in basi* 3 *ut in figura* (b)*.*
>
> *Tunc me divide omni quadrato partiale un alios quadrato, et ita ad infinito. Fig.* (c) *repraesenta successione de* 16 *quadratos in basi* 2 *; fig.* (d) *successione de* 81 *quadratos in basi* 3*.*
>
> *Si nos repraesenta per signo* $\cap$ *successione* $\begin{matrix} 1 & 2 \\ 0 & 3 \end{matrix}$ *, vel figura* (a)*, tunc figura* (e) *repraesenta successione de* 64 *quadratos in basi* 2*.*



(a) (b) (c) (d) (e)

> *In scripto* Sur une courbe qui remplit toute une aire plane*, MA. a.1890 t.36 p.157, me da expressione analytico de corrispondentia continuo inter numero reale* t*, et numero complexo* $(x; y)$*.*

L'article auquel Peano fait référence est l'article que nous reproduisons ci-après. Il a été publié dans les Mathematischen Annalen : c'est ce que signifie MA. Les dessins proviennent plutôt du travail de Hilbert.

Exercice. Pour préparer la lecture de l'article il est bien de se familiariser un peu avec les développements illimités.

a) Montrer que 1/3 a un développement décimal illimité infini, mais périodique. (On a : $1/3 = 0,3333\ldots$, que l'on peut aussi noter $0,(3)$ avec des parenthèses ou $0,\overline{3}$.)
b) Calculer le développement décimal illimité de 2/3 et de 5/9.
c) Montrer que le développement décimal illimité d'un nombre rationnel est fini si et seulement si le dénominateur du nombre n'est divisible que par les nombres premiers 2 et/ou 5.
d) Définir les développements illimités en base 3 et donner un critère pour la finitude du développement d'un nombre rationnel.

Reproduction de *Sur une courbe, qui remplit toute une aire plane*, par G. Peano à Turin.[17]

"Dans cette Note on détermine deux fonctions x et y, uniformes et continues d'une variable (réelle) t, qui, lorsque t varie dans l'intervalle $(0, 1)$, prennent toutes les couples de valeurs telles que $0 \leq y \leq 1$. Si l'on appelle, suivant l'usage, *courbe continue* le lieu des points dont les coordonnées sont des fonctions continues d'une variable, on a ainsi un arc de courbe qui passe par tous les points d'un carré. Donc, étant donné un arc de courbe continue, sans faire d'autres hypothèses, il n'est pas toujours possible de le renfermer dans une aire arbitrairement petite.

Adoptons pour base de numération le nombre 3 ; appelons *chiffre* chacun des nombres 0, 1, 2 ; et considérons une suite illimitée de chiffres a_1, a_2, a_3, ... que nous écrirons

$$T = 0, a_1 a_2 a_3 \ldots \ .$$

(Pour ce moment, T est seulement une suite de chiffres).

Si a est un chiffre, désignons par $\mathbf{k}a$ le chiffre $2 - a$, *complémentaire* de a ; c'est-à-dire, posons

$$\mathbf{k}0 = 2, \mathbf{k}1 = 1, \mathbf{k}2 = 0 \ .$$

Si $b = \mathbf{k}a$, on déduit $a = \mathbf{k}b$; on a aussi $\mathbf{k}a \equiv a(\mathrm{mod}.2)$. [18]

Désignons par $\mathbf{k}^n a$ le résultat de l'opération $\mathbf{k}$ répétée n fois sur a. Si n est pair, on a $\mathbf{k}^n a = a$; si n est impair, $\mathbf{k}^n a = \mathbf{k}a$. Si $m \equiv n(\mathrm{mod}.2)$, on a $\mathbf{k}^m a = \mathbf{k}^n a$.

Faisons correspondre à la suite T les deux suites

$$X = 0, b_1 b_2 b_3 \ldots, \quad Y = 0, c_1 c_2 c_3 \ldots,$$

où les chiffres b et c sont donnés par les relations

$$\begin{array}{lllllllll} b_1 = a_1, & c_1 & = & \mathbf{k}^{a_1} a_2, & b_2 = \mathbf{k}^{a_2} a_3, & & c_2 & = & \mathbf{k}^{a_1+a_3} a_4, \quad b_3 = \mathbf{k}^{a_2+a_4} a_5, \ldots \\ & b_n & = & \mathbf{k}^{a_2+a_4+\cdots+a_{2n-2}} a_{2n-1}, & & & c_n & = & \mathbf{k}^{a_1+a_3+\cdots+a_{2n-1}} a_{2n} \ . \end{array}$$

Donc b_n, $n^{\text{ième}}$ chiffre de X, est égal à a_{2n-1}, $n^{\text{ième}}$ chiffre de rang impaire [sic]dans T, ou à son complémentaire, selon que la somme $a_1 + \cdots + a_{2n-2}$ des chiffres de rang pair, qui le précédent, est paire ou impaire. Analoguement pour Y. On peut aussi écrire ces relations sous la forme :

$$a_1 = b_1, a_2 = \mathbf{k}^{b_1} c_1, a_3 = \mathbf{k}^{c_1} b_2, a_4 = \mathbf{k}^{b_1+b_2} c_2, \ldots,$$

$$a_{2n-1} = \mathbf{k}^{c_1+c_2+\cdots+c_{n-1}} b_n, a_{2n} = \mathbf{k}^{b_1+b_2+\cdots+b_n} c_n \ .$$

Si l'on donne la suite T, alors X et Y résultent déterminées, et si l'on donne X et Y, la T est déterminée.

[17] Dans l'original il manque des accents et certaines expressions sonnent bizarrement à l'oreille contemporaine, mais nous restons fidèles à l'original.

[18] L'écriture $x \equiv y(\mathrm{mod}.2)$ signifie simplement que x et y ont la même parité, ils donnent le même reste lorsqu'on les divise par 2.

Appelons *valeur* de la suite T la quantité (analogue à un nombre décimal ayant même notation)

$$t = \text{val.}T = \frac{a_1}{3} + \frac{a_2}{3^2} + \cdots + \frac{a_n}{3^n} + \cdots \ .$$

A chaque suite T correspond un nombre t, et l'on a $0 \leq t \leq 1$. Réciproquement les nombres t, dans l'intervalle $(0, 1)$ se divisent en deux classes :

α) Les nombres, différents de 0 et de 1, qui multipliés par une puissance de 3 donnent un entier ; ils sont représentés par deux suites, l'une

$$T = 0, a_1 a_2 \ldots a_{n-1} a_n 222 \ldots$$

où a_n est égal à 0 ou à 1 ; l'autre

$$T' = 0, a_1 a_2 \ldots a_{n-1} a_n^{'} 000 \ldots$$

où $a_n^{'} = a_n + 1$.

β) Les autres nombres ; ils sont représentés par une seule suite T.

Or la correspondance établie entre T et (X, Y) est telle que si T et T' sont deux suites de forme différente, mais $\text{val.}T = \text{val.}T'$, et si X, Y sont les suites correspondantes à T, et X', Y' celles correspondantes à T', on a

$$\text{val.}X = \text{val.}X' \ , \text{val.}Y = \text{val.}Y' \ .$$

En effet considérons la suite

$$T = 0, a_1 a_2 \ldots a_{2n-3} a_{2n-2} a_{2n-1} a_{2n} 222 \ldots$$

où a_{2n-1} et a_{2n} ne sont pas toutes deux égales à 2. Cette suite peut représenter tout nombre de la classe α. Soit

$$X = 0, b_1 b_2 \ldots b_{n-1} b_n b_{n+1} \ldots$$

on a :

$$b_n = \mathbf{k}^{a_2 + \cdots + a_{2n-2}} a_{2n-1}, \ b_{n+1} = b_{n+2} = \cdots = \mathbf{k}^{a_2 + \cdots + a_{2n-2} + a_{2n}} 2 \ .$$

Soit T' l'autre suite dont la valeur coïncide avec $\text{val.}T$,

$$T' = 0, a_1 a_2 \ldots a_{2n-3} a_{2n-2} a_{2n-1}^{'} a_{2n}^{'} 000 \ldots$$

et

$$X' = 0, b_1 \ldots b_{n-1} b_n^{'} b_{n+1}^{'} \ldots \ .$$

Les premiers $2n-2$ chiffres de T' coïncident avec ceux de T ; donc les premiers $n-1$ chiffres de X' coïncident aussi avec ceux de X ; les autres sont déterminés par les relations

$$b_n^{'} = \mathbf{k}^{a_2 + \cdots + a_{2n-2}} a_{2n-1}^{'}, \ b_{n+1}^{'} = b_{n+2}^{'} = \cdots = \mathbf{k}^{a_2 + \cdots + a_{2n-2} + a_{2n}^{'}} 0 \ .$$

Nous distinguerons maintenant deux cas, suivant que $a_{2n} < 2$, ou $a_{2n} = 2$.

Si a_{2n} a la valeur 0 ou 1, on a $a_{2n}^{'} = a_{2n} + 1$, $a_{2n-1}^{'} = a_{2n-1}$, $b_n^{'} = b_n$,

$$a_2 + a_4 + \cdots + a_{2n-2} + a_{2n}^{'} = a_2 + \cdots + a_{2n-2} + a_{2n} + 1 \ ,$$

d'où

$$b_{n+1}^{'} = b_{n+2}^{'} = \cdots = b_{n+1} = b_{n+2} = \cdots = \mathbf{k}^{a_2 + \cdots + a_{2n}} 2 \ .$$

Dans ce cas les deux séries X et X' coïncident en forme et en valeur.

Si $a_{2n} = 2$, on a $a_{2n-1} = 0$ ou 1, $a'_{2n} = 0$, $a'_{2n-1} = a_{2n-1} + 1$, et en posant

$$s = a_2 + a_4 + \cdots + a_{2n-2}$$

on a

$$b_n = \mathbf{k}^s a_{2n-1}, \quad b_{n+1} = b_{n+2} = \cdots = \mathbf{k}^s 2 \ ,$$
$$b'_n = \mathbf{k}^s a'_{2n-1}, \quad b'_{n+1} = b'_{n+2} = \cdots = \mathbf{k}^s 0 \ .$$

Or puisque $a'_{2n-1} = a_{2n-1} + 1$, les deux fractions $0, a_{2n-2}222\ldots$ et $0, a'_{2n-1}000\ldots$ ont la même valeur ; en faisant sur les chiffres la même opération $\mathbf{k}^s$ on obtient les deux fractions $0, b_n b_{n+1} b_{n+2}\ldots$ et $0, b'_n b'_{n+1} b'_{n+2}\ldots$, qui ont aussi, comme l'on voit facilement, la même valeur ; donc les fractions X et X', bien que de forme différente, ont la même valeur.

Analoguement on prouve que val.Y = val.Y'.

Donc si l'on pose x = val.X, et y = val.Y, on déduit que x et y sont deux fonctions uniformes de la variable t dans l'intervalle $(0, 1)$. Elles sont continues ; en effet si t tend à t_0, les $2n$ premiers chiffres du développement de t finiront par coïncider avec ceux du développement de t_0, si t_0 est un β, ou avec ceux de l'un des deux développements de t_0, si t_0 est un α ; et alors les n premiers chiffres de x et y correspondantes à t coïncideront avec ceux des x, y correspondantes à t_0.

Enfin à tout couple (x, y) tel que $0 \leq x \leq 1$, $0 \leq y \leq 1$ correspond au moins un couple de suites (X, Y), qui en expriment la valeur ; à (X, Y) correspond une T, et à celle-ci t ; donc on peut toujours déterminer t de manière que les deux fonctions x et y prennent des valeurs arbitrairement données dans l'intervalle $(0, 1)$.

On arrive aux même conséquences si l'on prend pour base de numération un nombre impaire quelconque, au lieu de 3. On peut prendre aussi pour base un nombre pair, mais alors il faut établir entre T et (X, Y) une correspondance moins simple.

On peut former un arc de courbe continue qui remplit entièrement un cube. Faisons correspondre à la fraction (en base 3)

$$T = 0, a_1 a_2 a_3 a_4 \ldots$$

les fractions

$$X = 0, b_1 b_2 \ldots \ , Y = 0, c_1 c_2 \ldots \ , Z = 0, d_1 d_2 \ldots$$

où

$$b_1 = a_1, \ c_1 = \mathbf{k}^{b_1} a_2, \ d_1 = \mathbf{k}^{b_1 + c_1} a_3, \ b_2 = \mathbf{k}^{c_1 + d_1} a_4, \ \ldots$$
$$\begin{aligned} b_n &= \mathbf{k}^{c_1 + \cdots + c_{n-1} + d_1 + \cdots + d_{n-1}} a_{3n-2}, \\ c_n &= \mathbf{k}^{d_1 + \cdots + d_{n-1} + b_1 + \cdots + b_n} a_{3n-1}, \\ d_n &= \mathbf{k}^{b_1 + \cdots + b_n + c_1 + \cdots + c_n} a_{3n} \ . \end{aligned}$$

On prouve que x = val.X, y = val.Y, z = val.Z sont des fonctions uniformes et continues de la variable t = val.T ; et si t varie entre 0 et 1, x, y, z prennent tous les termes de valeurs qui satisfont aux conditions $0 \leq x \leq 1$, $0 \leq y \leq 1$, $0 \leq z \leq 1$.

M. Cantor, (Journal de Crelle, t. 84, p. 242) a démontré qu'on peut établir une correspondance univoque et réciproque (unter gegenseitiger Eindeutigkeit) entre les points d'une ligne et ceux d'une surface. M. Netto (Journal de Crelle, t. 86, p. 263), et d'autres ont démontré qu'un telle correspondance est nécessairement discontinue. (Voir aussi G. Loria, *La definizione dello spazio ad n dimensioni ... secondo le ricerche di G. Cantor*, Giornale di Matematiche, 1877). Dans ma Note on démontre qu'on peut établir d'un côté l'uniformité et la continuité, c'est-à-dire, aux points d'une ligne on peut faire correspondre les points d'une surface, de façon que l'image de la ligne soit l'entière surface, et que le point sur la surface soit fonction continue du point de la ligne. Mais cette correspondance n'est point univoquement réciproque, car aux points (x, y) du carré, si x et y sont des β, correspond bien une seule valeur de t,

mais si x, ou y, ou toutes les deux sont des α, les valeurs correspondantes de t sont en nombre de 2 ou de 4.

On a démontré qu'on peut enfermer un arc de courbe plane continue dans une aire arbitrairement petite :

1) Si l'une des fonctions, p. ex. la x coïncide avec la variable indépendante t ; on a alors le théorème d'intégrabilité des fonctions continues.
2) Si les deux fonctions x et y sont à variation limitée (Jordan, Cours d'analyse, III, p. 599). Mais, comme démontre l'exemple précédent, cela n'est pas vrai si l'on suppose seulement la continuité des fonctions x et y.

Ces x et y, fonctions continues de la variable t, manquent toujours de dérivée.

Turin, Janvier 1890."

Une autre façon d'obtenir une courbe continue qui remplit un carré est par un processus limite, qui construit une suite de fonctions en se basant sur la construction géométrique esquissée dans le premier extrait de Peano.

On commence par une fonction continue $\varphi(t) = (x(t), y(t))$ *quelconque* de l'intervalle $I = [0, 1]$ dans les carré $I \times I$, qui relie les points $(0, 0)$ et $(1, 0)$, avec $\varphi(0) = (0, 0)$ et $\varphi(1) = (1, 0)$. On définit une nouvelle fonction $F(\varphi)$ ayant les mêmes propriétés que φ en posant

$$F(\varphi)(t) = \begin{cases} \frac{1}{2}(y(4t), x(4t)) & \text{si } 0 \le t \le \frac{1}{4} \\ \frac{1}{2}(x(4t-1), 1 + y(4t-1)) & \text{si } \frac{1}{4} \le t \le \frac{2}{4} \\ \frac{1}{2}(1 + x(4t-2), 1 + y(4t-2)) & \text{si } \frac{2}{4} \le t \le \frac{3}{4} \\ \frac{1}{2}(2 - y(4t-3), 1 - x(4t-3)) & \text{si } \frac{3}{4} \le t \le 1 \ . \end{cases}$$

Exercice. Vérifier les affirmations suivantes, qui précisent cette définition.

Le fait que l'on multiplie t par 4 signifie que l'on parcourt l'intervalle I quatre fois plus rapidement. Sur chaque quart de I on fait subir une transformation simple à la courbe définie par φ. D'abord, le facteur $1/2$ diminue la taille de l'image d'un facteur 2. Puis en inversant les coordonnées x et y on fait subir à la courbe une réflexion le long de la diagonale $x = y$. Puis en ajoutant 1 à la coordonnée y on la translate vers le haut (de combien ? ; pourquoi retranche-t-on 1 à $4t$?). De même pour la transformation sur les deux derniers intervalles.

Vu que la fonction $F(\varphi)$ a les mêmes propriétés que φ, nous pouvons l'utiliser pour construire une nouvelle fonction, et ainsi de suite. Nous obtenons donc une suite de fonctions

$$\varphi_0 = \varphi,\ \varphi_1 = F(\varphi),\ \varphi_2 = F(\varphi_1),\ \ldots,\ \varphi_n = F(\varphi_{n-1}),\ \ldots$$

Il est clair que si ψ est une autre fonction ayant les mêmes propriétés que φ, et telle que la distance maximale pour tous les choix de t dans I, entre les valeurs $\varphi(t)$ et $\psi(t)$ est inférieure ou égale à une constante K, alors la distance maximale entre les valeurs $F(\varphi)(t)$ et $F(\psi)(t)$ est inférieure ou égale à $K/2$ (c'est le facteur $1/2$ dans la définition de F, qui le garantit). On en déduit que si $\psi = \varphi_m$ pour un certain m et si on majore K par 1, alors pour tout t dans I et tout entier n la distance entre $\varphi_k(t)$ et $\varphi_{k+m}(t)$ satisfait

$$d(\varphi_n(t), \varphi_{n+m}(t)) \le 2^{-k} \ .$$

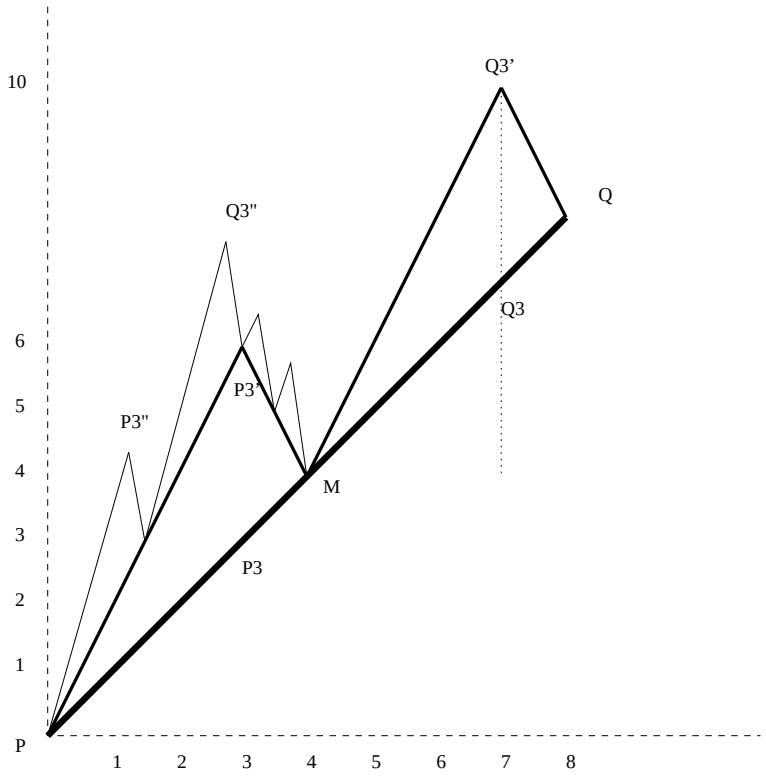
On peut montrer que ceci suffit à garantir que la suite $\varphi_n(t)$ converge vers une fonction $\varphi_\infty(t)$, qui est continue : non seulement pour chaque valeur de t la suite des points $\varphi_n(t)$ converge vers un point

bien déterminé, mais la variation du point limite comme fonction de t est continue. (L'inégalité mise en évidence garantit la convergence *uniforme* de la suite $\varphi_n(t)$.)

Exercice. Dessiner le graphe de la fonction $x \mapsto x^n$, sur l'intervalle $[0, 1]$, pour des valeurs croissantes de l'entier n. Quelle semble être la limite de cette suite de fonctions pour n devenant de plus en plus grand ? La limite est-elle continue ? Pourquoi ?

2.10 Il existe des fonctions partout continues et nulle part dérivables.

Comme Peano le remarque à la fin de son article, les fonctions coordonnées de sa courbe (comme de celle de Hilbert) sont des fonctions partout continues et nulle part dérivables. Un autre exemple d'une telle fonction s'obtient par un autre processus limite–dû à Bolzano–, représenté sur la figure ci-dessous. Ici le segment MQ_3' a pente double que le segment MQ.



Chapitre 3

Logique

Résumé : quand on fait des démonstrations en mathématiques on s'appuie sur des règles de déduction bien précises. Ces règles sont un des objets de la logique. Dans ce chapitre on présente rapidement le calcul propositionnel (1) par les tableaux de vérité et (2) comme système déductif. Puis on introduit le calcul des prédicats. On insiste sur la différence entre la propriété pour une proposition d'être vraie (on dira plutôt valide) et d'être démontrée. En passant on discute ce qui fonde les méthodes de démonstration par contraposition et par l'absurde.

Les mathématiques sont une science déductive. Elles sont exposées comme une suite d'énoncés que l'on déduit les uns des autres dans un ordre précis. Bien que souvent, en vue d'applications diverses, on ne considère comme intéressants que ces énoncés, l'intérêt des mathématiques résidant dans le fait que ces énoncés sont démontrés. De manière vague, la démonstration d'un énoncé est une suite finie d'énoncés, qui commence par des énoncés admis et qui se termine par l'énoncé en question. On passe d'un terme à l'autre d'une telle suite en appliquant des règles, qui–il faut l'admettre–sont rarement complètement explicitées. [1]

Les mathématiques sont présentées sous une forme, qui est un mélange complexe de langage courant et de symboles particuliers. Les concepts et symboles particuliers aux mathématiques sont introduits au moyen de définitions. Après les définitions on passe aux énoncés qui à leur tour seront suivis des démonstrations. [2]

Les concepts spécifiques aux mathématiques tels que "nombre", "fonction", etc. seront définis à l'aide de la théorie des ensembles, que nous allons considérer au chapitre suivant. On commence ici par s'intéresser aux principes logiques qui sont à la base de cette théorie et aussi des méthodes de démonstration. On présente un calcul, qui s'applique de manière beaucoup plus générale et qui peut servir à formaliser tout type de raisonnement. En particulier, ce calcul permet de traduire en symboles le genre de phrase couramment utilisée en mathématiques.

[1] En quelque sorte les énoncés *démontrés* des mathématiques sont l'analogue des faits *observés* en physique. Par ailleurs, lorsqu'il/elle fait de la recherche le/la mathématicien(ne) comme le physicien expérimente, conjecture, tâtonne, ... C'est seulement pour présenter ses résultats que le/la mathématicien(ne) se met en "mode déductif".

[2] Les énoncés portent des noms variés : "théorème" s'il s'agit d'un résultat important, "proposition" s'il l'est moins (?), "corollaire" si c'est un énoncé qui est une conséquence (directe) d'un autre énoncé, "lemme" si c'est un énoncé "qui sert à démontrer".

3.1 Le calcul propositionnel.

Ce calcul ainsi que la théorie de la quantification exposée plus loin remontent à G. Frege (1848-1925), ses lois affirment quelque chose sur les propriétés des propositions *quelconques.* Il s'agit en quelque sorte de l'analogue des notions communes mises en évidence dans les Éléments d'Euclide, sauf que l'on s'attache ici à décrire les lois du raisonnement (formel). (Cependant la relation d'égalité n'est généralement pas considérée comme faisant partie du calcul propositionnel.)

On commence par la syntaxe. On se donne des *variables propositionnelles* p, q, ... à partir desquelles, à l'aide de certains symboles, on construit des *fonctions propositionnelles* (expressions bien formées). Les symboles que l'on utilise le plus couramment sont les suivants (*connecteurs logiques*) :

$$\neg \ , \ \vee \ , \ \wedge \ , \ \Rightarrow \ , \ \Leftrightarrow , \ldots$$

On combine ces symboles avec des parenthèses et des signes de ponctuation. Les connecteurs ci-dessus sont appelés respectivement *négation, disjonction, conjonction, implication (philonienne ou logique), équivalence.* Noter que l'on pourrait par exemple se restreindre à ne considérer que des expressions construites à l'aide de $\neg$ et $\vee$ (voir plus bas).

Voici des fonctions propositionnelles

$$\neg p \ \ , \ \ p \wedge q \ \ , \ \ (p \vee q) \Rightarrow (p \vee r) \ .$$

Voici des exemples d'agrégats de symboles qui n'en sont pas [3] :

$$pq \Rightarrow \ \ , \ \ p\neg \ \ , \ \ ()p \wedge \vee \ .$$

Les variables propositionnelles peuvent prendre deux *valeurs de vérité* : V et F (ou 1 et 0) ; elles ne prennent qu'une valeur à la fois. Les fonctions propositionnelles deviennent alors des *fonctions de vérité* à l'aide des définitions suivantes (tableaux de vérité) [4] :

p	$\neg p$
V	F
F	V

p	q	$p \wedge q$	$p \vee q$	$p \Rightarrow q$	$p \Leftrightarrow q$	$p \oplus q$	$p \mid q$	$p \downarrow q$
V	V	V	V	V	V	F	F	F
F	V	F	V	V	F	V	V	F
V	F	F	V	F	F	V	V	F
F	F	F	F	V	V	F	V	V

On peut lire ces fonctions propositionnelles élémentaires comme suit :

$\neg p$	"non p"
$p \wedge q$	"p et q"
$p \vee q$	"p ou q
$p \Rightarrow q$	"si p alors q"
$p \Leftrightarrow q$	"p si et seulement si q".

[3] On peut être plus précis. Sont des fonctions propositionnelles (fp) les expressions suivantes et aucune autre : (a) les variables propositionnelles, (b) toute fp précédée de $\neg$, (c) toute fp suivie de l'un des connecteurs logiques suivi par une fp (le tout entre parenthèses).

[4] En tout il y aurait ici 16 possibilités, on n'explicite que les 7 connecteurs les plus courants ; noter qu'il y a d'autres notations en usage : $-p$ ou $\overline{p}$ pour $\neg p$, $\rightarrow$ pour $\Rightarrow$, etc.

Ainsi, on peut lire le tableau en disant, par exemple, que la conjonction de deux propositions n'est vraie que si les deux propositions le sont.

Exemple. La phrase "Si ce n'est pas le cas que $x > y$ et si ce n'est pas le cas que $y > z$, alors ce n'est pas le cas que $x > z$", se traduit par $(\neg p \wedge \neg q) \Rightarrow \neg r$, où p (resp. q, r) représente la "phrase" $x > y$ (resp. $y > z$, $x > z$). La traduction peut servir à déterminer si la phrase d'origine est correcte (noter que par exemple $\neg p$ signifie en fait $x \leq y$).

Noter que l'implication philonienne $p \Rightarrow q$ a la particularité que si l'antécédent p prend la valeur F ("faux") alors, *indépendamment de la valeur du conséquent* q, elle prend la valeur V ("vrai"). Aussi, $p \Rightarrow q$ est faux seulement si l'antécédent p est vrai et le conséquent q est faux. Il existe d'autres systèmes logiques, utiles à d'autres fins, où le "si ... alors ..." est traduit différemment.

Comme indiqué plus haut, on n'a pas besoin d'utiliser tous les connecteurs logiques que nous venons de définir. On dira que G et H sont *équivalentes* si les fonctions propositionnelles G et H prennent la même valeur de vérité pour toute distribution de valeurs de vérité des variables apparaissant dans G et H. On écrira $G \Leftrightarrow H$. Par exemple

$$(p \wedge q) \Leftrightarrow \neg(\neg p \vee \neg q) \quad , \quad (p \Rightarrow q) \Leftrightarrow \neg(p \wedge \neg q) \quad \text{et aussi} \quad (p \Rightarrow q) \Leftrightarrow (\neg p \vee q) \ .$$

Pour vérifier ces équivalences il suffit de calculer à l'aide des tableaux de vérité ! Vérifions la première :

p	q	$(p$	$\wedge$	$q)$	$\Longleftrightarrow$	$\neg$	$(\neg p$	$\vee$	$\neg q)$
V	V		V		**V**	V	F	F	F
V	F		F		**V**	F	F	V	V
F	V		F		**V**	F	V	V	F
F	F		F		**V**	F	V	V	V
			1		6	5	2	4	3

On lit la table définissant la conjonction et on remplit la colonne 1. Puis la définition de la négation donne les colonnes 2 et 3. Ensuite on remplit 4 et 5. La colonne 6 est obtenue à l'aide de la 1 et de la 5, vu qu'elle ne contient que des V, on a vérifié l'équivalence.

Exercices : comment lire les fonctions $p \oplus q$, $p \mid q$ et $p \downarrow q$, d'après la définition du tableau ? Montrer que avec $\neg$ et $\vee$ on peut exprimer tous les autres connecteurs considérés. Montrer que la barre de Scheffer $\mid$ permet d'exprimer la négation et la disjonction (et donc tout).

Un peu de vocabulaire : si $p \Rightarrow q$ est une implication alors

$q \Rightarrow p$ est l'implication *réciproque*
$\neg p \Rightarrow \neg q$ est l'implication *inverse*
$\neg q \Rightarrow \neg p$ l'implication *contraposée*.

On vérifie par calcul que *une implication et sa contraposée sont équivalentes*. Ceci fonde la technique de démonstration par contraposition.

Exemple : "s'il pleut, je prends mon parapluie" est équivalent à "si je ne prends pas mon parapluie, alors il ne pleut pas".

Exemple : d'après ce qui précède, pour montrer la parité d'un entier n, dont le carré n^2 est pair, il suffit de montrer que si n est impair (la négation d'être pair), alors n^2 est impair.

Exercice : montrer par contraposition que si un nombre rationnel x est positif, alors $2x$ est positif.

3.2 Validité I.

Une *proposition valide* (ou *tautologie* ou *loi logique*) est une fonction propositionnelle ne prenant que la valeur V. En particulier une équivalence est une proposition valide. Voici des exemples de propositions valides :

$\neg(\neg p) \Leftrightarrow p$	"double négation"
$\neg(p \wedge \neg p)$	"loi de contradiction"
$p \vee \neg p$	"tiers exclu".

Les propositions valides sont les énoncés qui nous intéressent : ce sont les "résultats" de la théorie. D'après ce qui précède il est facile de vérifier si une proposition est valide : il suffit de dresser un tableau de vérité, qui va avoir 2^n lignes, si la fonction fait intervenir n variables (ça peut donc être très fastidieux à faire en pratique, mais une machine peut le faire !)

Au lieu de dresser un tableau on peut aussi procéder par *réduction* : par exemple pour vérifier si l'implication

$$((p \wedge q) \vee (\neg p \wedge \neg r)) \Rightarrow (q \Leftrightarrow r)$$

est valide il suffit de voir s'il est possible que le conséquent $C : (q \Leftrightarrow r)$ est faux en même temps que l'antécédent $A : ((p \wedge q) \vee (\neg p \wedge \neg r))$ est vrai. Or, C est faux précisément quand q et r n'ont pas la même valeur et A est vrai dès que l'un de $(p \wedge q)$ ou $(\neg p \wedge \neg r)$ est vrai. Supposons q vrai et r faux, alors il suffit de prendre p vrai pour que A soit vrai. Ainsi l'implication n'est pas valide.

Exercice. Montrer la validité de la proposition :

$$((p \Rightarrow q) \wedge (q \Rightarrow r)) \Rightarrow (p \Rightarrow r) \ .$$

3.3 Méthode déductive I.

Une autre notion est celle de *proposition démontrée*. Cette notion aura un sens lorsqu'on aura donné au calcul propositionnel la forme d'un système déductif. Pour ça, on doit choisir un ensemble de propositions (les *axiomes*) et on doit spécifier des *règles d'inférence*. Ces règles décrètent quelles sont les suites de propositions qui vont former des *démonstrations* (synonyme de *preuves*). Une démonstration de la proposition B sera alors une suite de propositions A_1, ... A_n, B, qui se termine par B et qui est formée de propositions déjà démontrées. L'on passe de l'une des propositions à la suivante en appliquant une règle d'inférence et on considère les axiomes comme étant démontrés.

Citons un maître :

> *"Une preuve complète peut [...] se caractériser comme suit : elle consiste dans la construction d'une chaîne de propositions jouissant des propriétés que voici : les membres initiaux sont des propositions déjà tenues [...] pour [démontrées] ; chaque membre subséquent s'obtient des précédents en appliquant une* règle d'inférence *; et enfin le dernier membre est la proposition à prouver."*

(A. Tarski, "Introduction à la logique", p. 44, Gauthier-Villars, 1971, Paris)

Les quatre propositions suivantes ont été retenues comme axiomes par J.H. Whitehead et B. Russel dans leur travail monumental sur les fondements des mathématiques "Principia mathematica", Cambridge, 1910-1913 ; ce sont les *axiomes logiques* :

AL1 $(p \vee p) \Rightarrow p$

AL2 $p \Rightarrow (p \vee q)$
AL3 $(p \vee q) \Rightarrow (q \vee p)$
AL4 $(r \Rightarrow s) \Rightarrow ((r \vee t) \Rightarrow (s \vee t))$

Ici $p \Rightarrow q$ est considéré comme une abréviation de $q \vee \neg p$.

D'habitude on retient deux règles d'inférence : la *règle de substitution* et la *règle de détachement.*

> *La règle de substitution.* "Si une proposition de caractère universel, qui a déjà été acceptée comme [démontrée] contient des variables propositionnelles, et si ces variables sont remplacées par d'autres variables propositionnelles ou par des fonctions propositionnelles ou par des propositions–en substituant partout autant d'expressions à autant de variables–, alors la proposition obtenue de cette façon peut aussi être tenue pour [démontrée]." (*loc. cit.* p. 42)

> *La règle de détachement.* "Si deux propositions acceptées comme [démontrées], l'une ayant la forme d'une implication tandis que l'autre est l'antécédent de cette implication, alors la proposition qui constitue le conséquent de l'implication peut être reconnue comme [démontrée]." (*loc. cit.* p. 43)

On peut généraliser cette règle et déduire une proposition q d'une famille de propositions p_1, ..., p_n, si la *conjonction* $p_1 \wedge \ldots \wedge p_n$ des propositions de la famille apparaît comme l'antécédent d'une implication dont la proposition q est le conséquent. De manière schématique on représente la règle de détachement comme suit :

$$\frac{\begin{array}{c} A \\ A \Rightarrow B \end{array}}{B}$$

Ceci résume le fait que A et $A \Rightarrow B$ sont démontrées et que par conséquent on peut en tirer B.

Exemples. 1) Voici deux exemples d'utilisation des règles d'inférence. Donnons une démonstration de $p \Rightarrow (q \Rightarrow p)$. En substituant $\neg q$ à q dans (AL2) on déduit $p \Rightarrow (p \vee \neg q)$, qui équivaut à la proposition voulue par la définition (utilisée ici) de $\Rightarrow$.

Supposons montrées les propositions :

$$(I) \;\; p \Rightarrow (q \Rightarrow p) \;\; \text{et} \;\; (II) \;\; (p \Rightarrow (p \Rightarrow q)) \Rightarrow (p \Rightarrow q) \; .$$

On veut en déduire $p \Rightarrow p$. D'abord on opère sur (I) la substitution de q avec p, ce qui donne

$$p \Rightarrow (p \Rightarrow p) \;\; (*) \; .$$

Ensuite on opère dans (II) la substitution de q par p, pour obtenir :

$$(p \Rightarrow (p \Rightarrow p)) \Rightarrow (p \Rightarrow p) \; .$$

Noter que l'antécédent de cette dernière implication est la proposition (*), ainsi par détachement on obtient la proposition voulue, à savoir on utilise le schéma

$$\frac{\begin{array}{c} p \Rightarrow (p \Rightarrow p) \\ (p \Rightarrow (p \Rightarrow p)) \Rightarrow (p \Rightarrow p) \end{array}}{p \Rightarrow p}$$

2) Il n'est peut-être pas inutile de donner un *analogue d'un système déductif*, comme nous venons de le définir. Dans cet analogue on appellera *proposition* tout mot anglais contenu dans un dictionnaire fixé, disons le dictionnaire "The concise Oxford French dictionary", Claredon Press, Oxford, 1980. L'unique *règle d'inférence* consiste à changer une lettre dans un mot. Voici un exemple de démonstration dans

ce "mini-système déductif". Nous allons déduire la proposition CASH de la proposition SLOT. Il faut exhiber une suite de mots contenus dans le dictionnaire cité. Chaque mot dans la suite ne doit différer du précédent que par une lettre. Voici une démonstration (avec entre parenthèses la traduction en Français) :

SLOT (fente)–SOOT (suie)–MOOT (débattre v.t.)–MOST (le plus de)–COST (prix)–CAST (jet, moule)–CASH (argent).

Vous pouvez essayer de montrer, que dans une démonstration de CASH à partir de SLOT il faut nécessairement "passer" par un mot/proposition contenant deux voyelles.

Ainsi l'impossibilité de la quadrature du cercle *à la règle et au compas* est analogue à la non-dénombrabilité de CASH à partir de SLOT *sans utiliser les mots à deux voyelles.*

3.4 Cohérence et complétude I.

Appelons **CP** le système déductif avec pour axiomes (AL1-4) et pour règles d'inférence la substitution et le détachement. Notons que les axiomes logiques (AL1-4) sont des propositions valides et que les règles d'inférence mènent de propositions valides à propositions valides, ainsi : une proposition démontrée (dans **CP**) est valide (c'est la *cohérence*). Réciproquement, et c'est remarquable, dans **CP** on a l'équivalence

$$\boxed{\text{proposition démontrée} \leftrightarrow \text{proposition valide}}$$

C'est-à-dire que l'on peut montrer l'énoncé suivant :

> *"Toute proposition valide du calcul propositionnel s'obtient par une démonstration à partir des axiomes logiques (AL1-4)."*

(c'est la *complétude*; voir par exemple P. Bernays : "Axiomatische Untersuchung des Aussagenkalküls der Principia Mathematica," Math. Zeit. **25**(1926) ; voir aussi les travaux de E.L. Post).[5] Une conséquence de ce résultat est qu'il y a une procédure automatique (les tableaux) pour vérifier si une proposition admet une démonstration ou pas.

Remarque. Il faut noter que le système **CP** n'est pas le seul à rendre compte du calcul propositionnel. Nous avons mis celui de Whitehead et Russel en avant à cause de son importance historique. Ce système avait été précédé par le système de Frege (voir son œuvre majeure "Begriffsschrifft [Idéographie]", Halle, 1879) et il a été suivi par d'autres tout aussi puissants et–peut-être–plus élégants. Par exemple Nicod en 1917 a proposé un système avec le seul axiome

$$[p|(q|r)]|\left([t|(t|t)]|\{(s|q)|[(p|s)|(p|s)]\}\right) \ ,$$

et avec unique règle d'inférence, en plus de la règle de substitution, la règle

$$\frac{\begin{array}{c} P \\ P|(Q|R) \end{array}}{R}$$

Une alternative avec laquelle il est plus facile de travailler est celle proposée par Hilbert et Bernays dans "Grundlagen der Mathematik", vol. 1, Berlin, 1934 (p. 66). Ce choix est présenté dans les exercices du livre de Tarski déjà cité (voir *loc. cit.* Ex. 13, p. 136).

[5] On dit que la notion sémantique de proposition valide équivaut à la notion syntaxique de proposition démontrée.

Il est clair qu'expliciter toutes les étapes d'une démonstration d'une expression donnée est presque toujours une tâche impossible, d'autant plus que l'on a l'impression d'avancer à reculons, avec les yeux rivés sur les axiomes ! On sait d'où on vient mais on ne sait pas où on va.

En pratique, lorsque l'on fait une démonstration, on se base sur une réserve de propositions dont on sait (par ailleurs) qu'elles sont démontrées (par quelqu'un, quelque part, ...). Ainsi, par exemple, on utilise souvent des propositions valides, dont on montre la validité par tableau de vérité, sans chercher à les déduire des axiomes. [6] Pour des exemples voir la discussion des démonstrations indirectes donnée à la fin de ce chapitre.

3.5 Le calcul des prédicats ; quantificateurs.

En fait, en mathématiques on utilise un calcul propositionnel enrichi pour avoir une plus grande expressivité : on travaille avec des *prédicats* et avec des *quantificateurs* [7] . Par exemple on a besoin d'énoncer des propositions telles que "pour tout couple d'entiers x et y, il existe un entier z tel que $x + z = y$" ou "toute fonction continue est dérivable".

Exemples. Donnons des exemples de *prédicats*, des expressions susceptibles d'être quantifiées.
1) Gx = "x est chauve" (prédicat à un terme ou monadique)
2) Gxy = "x est le père de y" (prédicat relatif ou dyadique)
3) Pxy = "x est perpendiculaire à y"
4) $Gxyz$ = "x donne y à z" (prédicat triadique)
5) $Gxyz$ = "x se trouve entre y et z sur un cercle"
6) $Sxyz$ = "z est la somme de x et de y"
7) $Rxyzw$ = "x paye y à z pour w"
8) $Pxyzw$ = "x est à y comme z est à w" (proportion)

Ici on note x une variable et on note Gx une expression dont on met en évidence qu'elle contient la variable x. La variable x joue le rôle d'un pronom (comme "premier", "un", ...). On tâchera d'utiliser des lettres majuscules pour des expressions contenant des variables susceptibles d'être quantifiées et si Gx contient aussi d'autres variables on a intérêt à en faire mention explicite, on écrira alors $Gxyz$ (par exemple) au lieu de Gx.

On introduit le *quantificateur universel* $\forall$ et le *quantificateur existentiel* $\exists$. Ce sont des préfixes qui ne portent que sur les variables, à l'exclusion des fonctions. Les expressions

$$\forall x\ Gx \quad et \quad \exists x\ Gx \quad (*)$$

se lisent respectivement "pour tout x on a Gx" et "il existe x tel que Gx". A nouveau on aurait pu se borner à ne considérer que l'un des quantificateurs vu que l'on impose l'équivalence

$$\exists x Px \Leftrightarrow \neg(\forall x(\neg Px)) \ ,$$

qui se lit : l'existence d'un x tel que Px équivaut au fait qu'il est faux que pour tout x on a la négation de Px. De même

$$\neg(\exists x Rx) \Leftrightarrow \forall x(\neg Rx) \ .$$

[6] Il existe aussi des méthodes de déduction dites *naturelles*, qui sont "moins formelles" dans les sens qu'elles font en sorte de construire les démonstrations à partir des énoncés à montrer, de manière plus directe. Dans ces approches on n'explicite essentiellement que des règles d'inférence. La première de ces méthodes a été proposée par Gentzen en 1934.

[7] Ce que nous appelons ici calcul des prédicats est aussi appelé logique du premier ordre ou théorie de la quantification.

3.6 Validité II.

Il faut préciser ce qu'est une proposition admettant une valeur de vérité en présence de quantificateurs.

Exemple. L'expression "être impair" n'est pas vraie en tant que telle, elle est *vraie de* tous les nombres entiers qui ne sont pas divisibles par 2. Si Gx dénote le prédicat relatif "x est impair" on obtient une proposition en quantifiant et, par exemple, la proposition $\forall x\ Gx$ est fausse (ici on sous-entend que x ne parcourt que l'ensemble des nombres entiers).

Plus généralement auront une valeur de vérité les expressions fermées au sens des définitions qui suivent. On dit que la variable x est *liée* dans une expression, si elle tombe sous le coup d'un quantificateur. A l'opposé, si dans une expression une variable n'est pas quantifiée, alors on dit qu'elle est *libre*. Une expression est dite *ouverte* (resp. *fermée/close*) si elle contient (resp. ne contient pas) de variable libre. Les variables propositionnelles sont fermées. Les expressions ouvertes ne sont ni vraies, ni fausses.

> *"Une variable libre est ce qui correspond dans le langage courant à un pronom pour lequel on n'exprime ou ne sous-entend pas d'antécédent grammatical ; et l'analogue d'une proposition ouverte est une phrase qui contiendrait un tel pronom errant."*

(W.V.O. Quine : "Methods of logic", §17, Holt, Rinehart and Winston, New York, 1959). Une expression fermée est dite *valide* si elle est valide pour toute interprétation dans un univers de discours ; il faut s'imaginer que l'on laisse les variables prendre leurs valeurs dans tous les mondes possibles : celui des nombres, celui des animaux, etc.

Exemple : voici deux expressions fermées valides

$$\forall x F(x) \Rightarrow \exists x F(x) \quad , \quad \exists x (F(x) \wedge G(x)) \Rightarrow \exists x F(x)\ .$$

La *clôture universelle* d'une proposition ouverte est la proposition fermée obtenue en liant toute variable libre qui y apparaît par un quantificateur universel. On décrète qu'une proposition ouverte est valide si sa clôture universelle est valide.

Exemple : voici deux propositions ouvertes valides

$$\forall x F(x) \Rightarrow F(y) \quad , \quad F(y) \Rightarrow \exists x F(x)\ .$$

Notons quelques équivalences utiles (où l'on écrit R et S pour Rx et Sx) :

$$\begin{aligned} \forall x(R \wedge S) &\Leftrightarrow (\forall x R) \wedge (\forall x S) \\ \exists x(R \vee S) &\Leftrightarrow (\exists x R) \vee (\exists x S) \\ \forall x \forall y R &\Leftrightarrow \forall y \forall x R \\ \exists x \exists y R &\Leftrightarrow \exists y \exists x R \end{aligned}$$

Exercices : traduire en symboles "Il n'est pas vrai que tous les habitants ont été torturés" et "Il n'existe pas d'habitants qui étaient torturés". Ces propositions sont-elles équivalentes ? Montrer l'équivalence :

$$\neg(\forall x \exists y R x y) \Leftrightarrow \exists x \forall y (\neg R x y)\ .$$

Étudier l'implication $(\exists x R) \wedge (\exists x S) \Rightarrow \exists x(R \wedge S)$ et sa réciproque. S'agit-il de propositions valides ? (Indication : penser à l'exemple où Rx est "x est un homme chauve" et où Sx est "x est un homme barbu".)

En général, *on ne peut pas inverser les quantificateurs* existentiel et universel. Si Bxy signifie "l'étudiant y boit la bière x" il n'est pas du tout équivalent de dire "$\forall x \exists y Bxy$" ou "$\exists y \forall x Bxy$". En effet la première proposition se traduit par "toute bière est bue par un étudiant (chaque b. a son é.)" et la deuxième par "toutes les bières sont bues par un étudiant (le même é. pour toutes les b.)".

Un autre exemple est le suivant. Soit Nx l'énoncé "x est un nombre (rationnel)" et soit Gxy l'énoncé "x est inférieur à y". Alors

$$\forall x \exists y (Nx \Rightarrow (Ny \wedge Gxy))$$

signifie que pour tout nombre x il existe un nombre y supérieur à x (ce qui est vrai). Par contre

$$\exists y \forall x (Nx \Rightarrow (Ny \wedge Gxy))$$

signifie qu'il existe un nombre qui est supérieur à tout autre nombre (ce qui est faux).

3.7 Méthode déductive II.

Malgré les apparences il est possible de trouver des méthodes mécaniques pour contrôler la validité dans le calcul des prédicats [8] et on sait que même dans ce contexte plus général toute expression valide est démontrable ! Précisons quelque peu la notion de démonstration pour le calcul avec quantificateurs. On obtient un système d'axiomes en ajoutant aux axiomes logiques (AL1-4) deux axiomes, qui portent sur les quantificateurs. Ils s'énoncent comme suit :

AQ1 $(\forall x\ Fx) \Rightarrow Fy$

AQ2 $(\forall x\ (p \Rightarrow Fx)) \Rightarrow (p \Rightarrow \forall x\ Fx)$

Exemple. D'autres implications valides sont

$$\exists y\ (Fy \Rightarrow \forall x\ Fx) \quad \text{et} \quad Fy \Rightarrow \exists x\ Fx\ .$$

Pour compléter la description de ce qu'est une preuve du calcul des prédicats, il faut énoncer les *règles d'inférence* qui régissent les quantificateurs et qui se rajoutent aux règles de substitution et de détachement, que nous avons énoncées plus haut. On a

> la *règle de généralisation universelle*, qui permet de déduire d'une proposition n'importe quelle quantification universelle de celle-ci.

Ensuite on précise ce qu'est une *substitution légitime* dans une proposition avec quantificateurs. Faisons un exemple :

Exemple : le résultat de la substitution de $Gx \vee \exists z Hzx$ "pour F" dans $\forall x Fx \Rightarrow Fy$ est

$$\forall x (Gx \vee \exists z Hzx) \Rightarrow Gy \vee \exists z Hzy$$

(bien suivre x). Dans l'antécédent F porte sur x et dans le conséquent F porte sur y : dans la proposition que l'on substitue on garde x pour le F de l'antécédent et on change x en y pour le F dans le conséquent. Pourquoi ? En fait la substitution n'a de sens que si l'on spécifie une variable, qui sera celle sur laquelle porte F.

La substitution est soumise à deux restrictions. (1) Les quantificateurs de la proposition introduite ne doivent pas porter sur les variables de la proposition dans laquelle on l'introduit et (2) les variables de la proposition introduite ne doivent pas "tomber sous le coup" des quantificateurs de la proposition dans

[8] L. Löwenheim a présenté une telle méthode en 1915. Par contre il n'y a pas de méthode automatique pour vérifier la non-validité.

laquelle on l'introduit. En termes imagés, dans une expression $Gxyz$ les variables "tiennent une place (avec un nom), qui est entourée d'une barrière étanche à la quantification". Si l'on devait substituer une expression à x on la remplacerait à toute occurrence de x et l'on veillerait à ce que les quantificateurs "ne traversent pas" les parenthèses imaginaires qui marquent la place de x.

3.8 Cohérence et complétude II.

Avec ces définitions on a l'énoncé :

> *"Toute proposition valide du calcul des prédicats est démontrable (à partir des axiomes AL1-4 et AQ1-2)."*

(K. Gödel : "Die Vollständigkeit der Axiome des logischen Funktionenkalküls," Monatshefte für Math. und Phy. **37**(1930) (sa thèse de doctorat)).

Nous allons voir que, même si on le voulait, en mathématiques on ne pourrait pas faire en sorte que les démonstrations soient "mécaniques". En fait, en mathématiques, aucune procédure de démonstration ne permet d'atteindre tous les énoncés valides. Ceci est analogue au fait que dans les sciences expérimentales "vérité" ne coïncide pas avec "vérifiabilité".

Les deux résultats cités de P. Bernays et de K. Gödel sont des résultats de *complétude* : on peut démontrer toute proposition valide. Le fameux Théorème d'incomplétude de Gödel affirme que, au contraire,

> "*tout système formel (consistant* [9]*) assez riche est incomplet.*"

(voir K. Gödel, "Ueber formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme," Monatshefte für Math. und Phys., **38**(1931), 173–198.) Ici "assez riche" signifie qu'il contient l'arithmétique élémentaire, c'est-à-dire un système dans lequel on puisse "compter" [10]. Gödel construit dans un tel système une expression qui est vraie si et seulement si elle est... indémontrable (pour n'importe quelle procédure de démonstration) [11] ! Le système sur lequel se basent les mathématiques est bien "assez riche", il contient donc au moins une proposition indémontrable (si il est consistant).

Références : des résultats de complétude semblables à ceux énoncés, ainsi qu'un premier traitement systématique du calcul propositionnel par tableaux de vérité, sont contenus dans l'article de E.L. Post : "Introduction to a general theory of elementary propositions," Am. J. Math. **43**(1921). Cet article et celui de Gödel de 1930, cité plus haut, sont traduits dans J. Largeault "Logique mathématique—textes." A. Colin Ed., Paris, 1972. Voir aussi l'appendice au livre de Quine déjà cité.

---

[9] Il s'agit d'un système dans lequel on ne peut pas montrer à la fois une proposition et sa négation. **CP** est consistant, mais on ne sait pas montrer que le système (plus riche) utilisé en mathématique est consistant.

[10] En plus de l'égalité $=$ on ajoute à **CP** deux signes nouveaux 0 et σ (la fonction successeur) et on impose les axiomes $\neg(\sigma x = 0)$, $(\sigma x = \sigma y) \Rightarrow (x = y)$ et $(P(0) \wedge \forall\ x\ (P(x) \Rightarrow P(\sigma x))) \Rightarrow (\forall x\ P(x))$. Ces axiomes correspondent à trois des cinq axiomes que Peano a introduits pour caractériser les entiers naturels. Le dernier est l'axiome d'induction, il permet par exemple de donner une définition (récursive) de la somme de deux entiers, etc. On en parlera dans le chapitre sur la théorie des ensembles.

[11] L'idée vague est la suivante : toute expression du calcul propositionnel étendu peut se mettre sous une forme canonique ne comportant que les signes $\neg$, $\vee$, $($, $)$, $\forall$, des signes de variable $x, x', x'', \ldots$ et $p, p', p'', \ldots$ et des signes de prédicat $F, F', F'', \ldots$. Si donc, par exemple, on attribue les valeurs entières de 1 à 5 aux premiers signes et les valeurs de 6 à 9 à x, F, p et $'$, tout énoncé se verra associer un entier naturel : son *nombre de Gödel*. De même une suite d'énoncés aura un nombre de Gödel si on attribue la valeur 10 au "passage à la ligne" entre un énoncé et le suivant. Or, Gödel montre comment lire les propriétés des énoncés à partir des nombres qui leurs sont associés, en particulier il montre comment construire un énoncé (ouvert) $E(x, x', \ldots)$, dans la notation du système, dépendant de deux variables x et x', qui n'est vrai de deux entiers x et x', que si x est le nombre de Gödel d'une suite d'énoncés fournissant une preuve de x' (disons au sens de **CP**). Alors $\exists x\ E(x, x', \ldots)$ ne sera vrai que de ces entiers x', qui sont des nombres de Gödel d'énoncés démontrables. Ensuite Gödel montre, par un procédé semblable à celui de "la diagonale de Cantor" (voir plus bas), que l'on peut trouver un entier n, tel que n soit le nombre de Gödel de $\neg\exists x E(x, n, \ldots)$ (le même n !). Cet énoncé dit bien de lui-même qu'il n'est pas démontrable.

Un exemple de théorie déductive qui n'est pas complète est donné dans le livre de Tarski à la page 192.

Il existe plusieurs livres de divulgation sur le théorème d'incomplétude de Gödel. En voici un qui est bien connu : Nagel, E., Newman, J.R., Gödel, K., Girard, J.Y. "Le théorème de Gödel." Édifions du Seuil.

Sur le problème de la décision voir : A. Church, "A note on the *Entscheidungsproblem*", J. of symbolic Logic, **1**(1936), 40–41 et 101–102 et S.C. Kleene, "Introduction to metamathematics", North-Holland, 1952.

3.9 Démonstrations indirectes.

On peut maintenant expliciter les bases sur lesquelles se fondent quelques types de démonstration couramment utilisés en mathématiques.

Démonstrations par contraposition. Nous avons déjà rencontré ce type de démonstration, qui consiste à montrer une implication $p \Rightarrow q$ en montrant sa contraposée $\neg q \Rightarrow \neg p$, que l'on sait lui être équivalente.

Démonstrations par l'absurde. De manière générale, pour démontrer une proposition p par l'absurde on procède comme suit. On suppose que $\neg p$ est vraie/démontrée et on en déduit une contradiction... [12] Des livres entiers ont été consacrés à cette méthode de démonstration, ce qui laisse penser qu'elle ne va pas de soi (voir par exemple J.-L. Gardies, "Le raisonnement par l'absurde", P.U.F., Paris, 1991).

Un *premier type* de démonstration par l'absurde de p consiste à montrer $\neg p \Rightarrow q$ et, par ailleurs, $\neg q$. Par contraposition on alors $\neg q \Rightarrow p$, ce qui permet de déduire p par détachement

$$\frac{\neg q \qquad \neg q \Rightarrow p}{p}$$

Un *deuxième type* de démonstration par l'absurde de p est basé sur la proposition valide

$$(q \Rightarrow \neg q) \Rightarrow \neg q \ . \quad (*)$$

Pour la démonstration on suppose $\neg p$ vraie/démontrée et on montre $\neg p \Rightarrow p$. On termine alors par détachement en utilisant (*).

Un *troisième type* de démonstration par l'absurde concerne le cas particulier de propositions p qui sont des implications $q \Rightarrow r$. Pour montrer $q \Rightarrow r$ par l'absurde on suppose $q \wedge \neg r$ et on en tire une contradiction. En fait ceci n'est qu'un cas particulier des précédents, vu que $q \Rightarrow r$ équivaut à $\neg q \vee r$, dont la négation est bien $q \wedge \neg r$.

Pour terminer notons encore une autre (?) manière de comprendre les démonstrations par l'absurde. En effet on peut montrer qu'une proposition p est démontrable dans un système déductif **S** si et seulement si le système **S**+$\neg p$ obtenu en ajoutant $\neg p$ aux axiomes du système **S** est un système non-consistant, c'est-à-dire un sytème dans lequel on peut montrer une paire de propositions contradictoires. Or dans un système non-consistant on peut montrer *toute* formule et en particulier p. Ainsi, vu que de manière générale, si dans **S**+q on peut montrer r, alors dans **S** on peut montrer $q \Rightarrow r$, par ce qui précède, on obtient $\neg p \Rightarrow p$ *dans* **S**. Vu que $(\neg p \Rightarrow p) \Rightarrow p$ est valide on en déduit p par détachement (voir Chap. 4, 1.6 de R. Cori et D. Lascar, "Logique mathématique", vol. 1, Masson, Paris 1993).

[12] On dit que deux propositions forment une paire de *propositions contradictoires* si l'une est la négation de l'autre.

Exemples.

1) Voici deux propositions qu'il est commode de démontrer par contraposition : "si un entier n est tel que n^2 est pair, alors c'est que n est pair" et "si a est un entier, et si $4^a 7$ ne peut pas s'écrire comme une somme de trois carrés d'entiers, alors $4^{a+1}7$ n'est pas non plus la somme de trois carrés d'entiers."

2) Une des démonstrations usuelles de la proposition p : "il n'existe pas d'entiers naturels a et b tels que $2b^2 = a^2$" est une démonstration par l'absurde du premier type. On suppose donné $\neg p$, c'est-à-dire l'égalité

$$2b^2 = a^2 \quad (*)$$

(pour certains entiers naturels a et b). On construit alors une suite décroissante *infinie* d'entiers naturels (pairs). L'existence d'une telle suite est q. La négation de q est la loi fondamentale de l'arithmétique, qui dit que tout ensemble d'entiers naturels possède un plus petit élément. (Rappelons de la Sect. 2.4 comment l'on montre $\neg p \Rightarrow q$: on se base sur la première proposition de l'exemple précédent ; de (*) on déduit que a est pair, c'est-à-dire qu'il existe un entier naturel a_1 avec $a = 2a_1$, puis de même que b est pair, et donc $b = 2b_1$ pour un certain entier naturel b_1. Alors $2b_1^2 = a_1^2$ et on recommence... La suite que l'on considère est celle des a_n.)

L'alternative qui consiste à supposer (sans perte de généralité), que dans (*) soit a soit b est impair, mène à la proposition qui affirme l'existence d'un entier qui est à la fois pair et impair.

3) Un autre exemple de démonstration par l'absurde du premier type est donné par la démonstration de la loi de simplification

$$p : (x + y = x + z) \Rightarrow (y = z)$$

à partir de la loi de trichotomie

$$t : \text{"ou bien } (x = y) \text{ ou bien } (x < y) \text{ ou bien } (y < x)''$$

et des deux lois

$$I_1 : (y < z) \Rightarrow (x + y) < (x + z) \quad \text{et} \quad I_2 : (y > z) \Rightarrow (x + y) > (x + z)$$

(ici x, y et z sont des nombres (par exemple rationnels)). Voici comment on procède : on suppose $\neg p$; il existe donc x, y et z avec :

$$(x + y = x + z) \wedge \neg(y = z) \ .$$

Par t on en déduit

$$r_1 : \neg(x + y < x + z) \wedge \neg(x + y > x + z)$$

et $r_2 : (y < z) \vee (y > z)$ qui par I_1 et I_2 et détachement donne

$$r_3 : (x + y < x + z) \vee (x + y > x + z) \ .$$

Donc $\neg p \Rightarrow q$ avec $q = r_1 \wedge r_3$. Par ailleurs $\neg q$ est vrai (ici $r_1 = \neg r_3$).

4) Un exemple de démonstration du deuxième type. On admet

$$s : (x < y) \Rightarrow \neg(y < x)$$

et on veut montrer

$$p : \neg(x < x) \ .$$

La négation de p est

$$q : (x < x) \ .$$

Par substitution de y par x dans s il vient

$$(x < x) \Rightarrow \neg(x < x)$$

et par détachement on obtient $\neg(x < x)$. C'est-à-dire $q \Rightarrow \neg q$ dont on a vu que l'on déduit $\neg q$ qui est p.

5) Un exemple classique de démonstration par l'absurde est donné par la démonstration *de la diagonale* de Cantor, qui montre que l'on ne peut pas numéroter les nombres réels. (Cet énoncé sera plus clair après notre discussion des nombres réels, mais si l'on remplace les nombres réels par l'ensemble des suites de 0 et 1 on peut présenter cette démonstration sans savoir ce que sont les réels.)

On suppose (par l'absurde) avoir numéroté les réels entre 0 et 1, que l'on représente par leur développement décimal illimité :

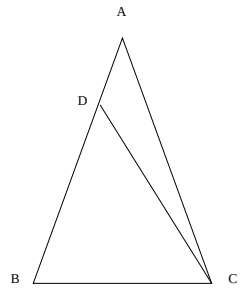
$$\begin{array}{lcl} 1 & : & 0, a_{11}a_{12}a_{13}... \\ 2 & : & 0, a_{21}a_{22}a_{23}... \\ 3 & : & 0, a_{31}a_{32}a_{33}... \\ \text{etc.} & & \end{array}$$

Puis on considère le nombre b défini par le développement décimal

$$0, b_1b_2b_3...$$

où l'on définit b_i comme étant égal à 0 ou 1 suivant que a_{ii} est égal à 1 ou pas. Ceci définit bien un réel b : noter par exemple que b_i ne vaut pas identiquement 9 à partir d'un certain rang. Ce nombre ne peut pas être dans la liste, car sinon il existerait i avec $b_i = a_{ii}$; à b ne correspond donc aucun entier dans la numérotation, qui n'en est donc pas une. On retrouve l'argument *via* $p \Rightarrow \neg p$.

6) Voici une démonstration tirée des Éléments d'Euclide. Il s'agit de la Proposition I.6, qui affirme que "si deux angles d'un triangle sont égaux l'un à l'autre, les côtés opposés à ces angles égaux seront aussi égaux l'un à l'autre". On regarde la figure.



On suppose, par l'absurde, qu'il existe un triangle ABC avec les angles $\widehat{ABC}$ et $\widehat{ACB}$ égaux mais avec AB différent de AC. Supposons que des deux côtés inégaux AB soit le plus grand. On part donc de

$$(\widehat{ABC} = \widehat{ACB}) \wedge (AB > AC) \quad (*)$$

En utilisant la Proposition I.2 des Éléments on montre qu'il existe un point D, compris entre A et B, tel que DB est égal à AC. On en tire que

$$\text{triangle } ABC = \text{triangle } DBC \quad (**)$$

(ici on utilise la Proposition I.4 des Éléments : "si deux triangles ont chacun deux côtés respectivement égaux à deux côtés de l'autre et si les angles compris entre les côtés égaux sont égaux, alors ces triangles

auront aussi leurs bases égales et seront égaux l'un à l'autre" ; on utilise aussi la loi $p \wedge q \Rightarrow p$ pour détacher $\widehat{ABC} = \widehat{ACB}$ de (*)). On conclut la démonstration en observant que (**) contredit la "notion commune", qui dit que "le tout est plus grand que la partie".

7) On vérifie par l'absurde la validité de

$$\forall x\ (Fx \Rightarrow Gx) \Rightarrow (\exists x\ Fx \Rightarrow \exists x\ Gx)\ .$$

Si elle est fausse c'est que son antécédent est vrai et que son conséquent est faux. De ceci on tire d'une part que, $Fy \Rightarrow Gy$ est vrai (par AQ1) et d'autre part, que à la fois $\exists x\ Fx$ est vrai et $\exists x\ Gx$ est faux. Du fait que $\exists x\ Fx$ est vrai on tire que Fy est aussi vrai. Ainsi par détachement on obtient Gy. Mais du fait que $\exists x\ Gx$ est faux on tire aussi que Gy est faux (par la négation de (AQ1)), ce qui est une contradiction.

8) Montrons par l'absurde qu'"il existe deux nombres irrationnels a et b tels que a^b soit rationnel" : supposons la négation de cette affirmation ; alors, puisque $\sqrt{2}$ n'est pas rationnel le nombre $\sqrt{2}^{\sqrt{2}}$ n'est pas rationnel. Par ailleurs $\sqrt{2}^{\sqrt{2}}$ élevé à la puissance $\sqrt{2}$ donne le rationnel 2, par conséquent $\sqrt{2}^{\sqrt{2}}$ ne peut pas être irrationnel non plus... (On notera que cette démonstration ne fournit pas les nombres a et b dont on a montré l'existence.)

3.10 Autres exemples d'utilisation de la méthode déductive.

Signalons rapidement que l'on peut exhiber une axiomatique qui permet de "démontrer les programmes informatiques" (!) (axiomatique de Hoare). Cette axiomatique est à la base de systèmes, que l'on peut implémenter sur machine, pour vérifier les programmes.

De même il existe une présentation axiomatique des théories physiques de la relativité (restreinte) et de la mécanique quantique, qui montre comment déduire l'ensemble des énoncés de la théorie d'un nombre limité de principes.

Dans un autre registre, le langage mathématique "théorème", "proposition", "corollaire", etc. est utilisé dans l'éthique du philosophe Spinoza, et Rameau procède par énoncés et démonstrations dans son "Traité d'harmonie".

3.11 Identité.

On peut rajouter le signe identité $=$ au calcul propositionnel avec quantificateurs en imposant les règles suivantes pour son usage :

$$\forall x\ (x = x) \quad \text{et} \quad \forall x \forall y\ ((Fx \wedge (x = y)) \Rightarrow Fy)\ .$$

Exemple. Avec l'identité on peut définir ce que l'on entend par "existence numériquement définie". Par exemple :

$$\neg(\exists x\ Fx) \ \text{ et } \ \exists x\ (Fx \wedge (\neg(\exists y\ (Fy \wedge \neg(y = x)))$$

signifient respectivement "il n'y a pas de x tel que Fx" et "il existe exactement un x tel que Fx" (on note souvent $\exists! x\ Fx$). Plus généralement on voit comment définir pour tout entier naturel n "il existe exactement n x tels que Fx", qui serait symbolisé par $\exists_n x\ Fx$. Cependant cette procédure ne permet

pas de définir le nombre naturel n, auquel on pense comme étant "l'ensemble de tous les ensembles à n éléments". Par exemple, si on note $\in$ la relation d'appartenance on définirait :

$$0 \stackrel{\text{déf.}}{=} \text{l'ensemble des ensembles } z \text{ tels que } \neg(\exists x\ x \in z)$$

et

$$1 \stackrel{\text{déf.}}{=} \text{l'ensemble des ensembles } z \text{ tels que } \exists y \forall x (x \in z \Leftrightarrow x = y)\ .$$

Mais ceci est déjà de la théorie des ensembles, non plus de la logique...

Chapitre 4

Théorie des ensembles.

Il s'agit d'une théorie belle, profonde et importante. En effet on peut dire que tout objet mathématique est un ensemble. Dans ces quelques pages nous passons en revue les axiomes de la théorie des ensembles, nous nous arrêterons sur le principe de démonstration par récurrence et nous donnerons un sens à ce que l'on entend par un ensemble plus grand qu'un autre. Une des applications majeures des concepts discutés ici sera pour nous la construction de l'ensemble des nombres réels, qui sera présentée dans la deuxième partie.

4.1 Tout objet mathématique est un ensemble.

Au départ, avec les travaux de G. Cantor (1845-1918), la théorie des ensembles a été considérée comme un

> *"outil pour étudier la pathologie des fonctions"*

(Schoenfiels, *Die Entwicklung der Lehre von der Punktmannigfaltigkeiten* [Le développement de la théorie des ensembles.], Jahresbericht der Deutschen Math.-Ver., **8**(1900)).

Plus tard on considérait, qu'il s'agissait là d'une théorie qui est

> *"la branche des mathématiques dont la tâche est d'étudier mathématiquement les notions fondementales de "nombre", ordre", et "fonction", en les considérant dans leur forme élémentaire, simple, et en développant de là les fondements logiques de toute l'arithmétique et l'analyse"*

(E. Zermelo, *Untersuchungen über die Grundlagen der Mengenlehre I* [Investigations sur les fondements de la théorie des ensembles.], Math. Ann. **65**(1908), 261–81). G. Peano (1858-1932) avait donné une description axiomatique de l'arithmétique élémentaire et les travaux de R. Dedekind (1831-1916) et d'autres avaient montré que l'on pouvait reconstruire toute l'analyse à partir de l'arithmétique. Signalons en passant qu'à l'époque et depuis un certain temps, existait un courant constructiviste dont l'un des mots d'ordre était une phrase célèbre, attribuée à L. Kronecker (1823-1891) :

> *"Dieu a crée les nombres entiers, le reste est l'œuvre de l'homme."*

Finalement, la théorie des ensembles donne un des fondements possibles aux mathématiques et est vue de nos jours comme une "théorie de l'infini". On en est arrivé à la situation où *tout objet mathématique peut être considéré comme un ensemble.* Ceci signifie, par exemple, que pour asseoir l'existence d'un objet mathématique on montre comment construire un certain ensemble à partir d'ensembles déjà connus. Il est clair que personne prétend qu'il n'y avait pas d'objets mathématiques avant la création de la théorie des ensembles : on savait déjà numéroter les pages d'un livre ! Il se trouve que la théorie que nous allons exposer rapidement fournit un *langage* tellement puissant qu'il peut rendre compte de

tous les concepts mathématiques intuitifs (considérés jusqu'à présent). De plus, en donnant une façon de manier l'infini elle facilite le maniement de concepts qui se sont révélés très utiles dans la physique moderne.

Exemple. Du Lycée on connaît la fonction "racine carrée", qui à un nombre réel positif x associe sa racine carrée y positive, c'est-à-dire l'unique nombre réel positif y tel que $y^2 = x$. Voici comment on arrive à voir cette fonction comme un ensemble : notons E l'ensemble des réels positifs (ou nuls). Étant donnés deux ensembles A et B, la théorie des ensembles nous permet de construire un ensemble, noté $A \times B$ formé de tous les couples ordonnés (x, y) avec x dans A et y dans B. La fonction en question sera définie comme étant le sous-ensemble de $E \times E$ des couples (x, y) avec $y^2 = x$ (on identifie donc la fonction à son graphe).

Mais on n'a pas fini ! Il faut encore décrire E et montrer que pour tout x positif il existe y dont le carré est x. Il est facile de décrire E une fois que l'on dispose de l'ensemble des nombres réels $\mathbf{R}$ muni de l'ordre. L'existence de la racine carrée est une des propriétés essentielles des réels. Or, il serait décevant de simplement postuler l'existence d'un ensemble aussi complexe que celui des nombres réels. En fait, comme on le verra, un des axiomes de la théorie des ensembles postule l'existence d'un ensemble infini (qui permet de construire un ensemble) $\mathbf{N}$, qui a l'essentiel des propriétés de l'ensemble des entiers naturels. Ceci est moins décevant : il est en effet difficile de s'imaginer un ensemble plus primitif que $\mathbf{N}$. Une fois donné $\mathbf{N}$ on construit assez facilement, les ensembles des entiers relatifs et des nombres rationnels (voir plus loin). Un nombre réel peut alors être identifié à une suite de nombres rationnels (ce que l'on appelle communément son développement décimal illimité). Mais qu'est-ce que une suite d'éléments d'un ensemble S ? C'est tout simplement une fonction définie sur $\mathbf{N}$ à valeurs dans S, donc un sous-ensemble de $\mathbf{N} \times E$. Et c'est tout !

4.2 Le système ZFC.

Le point de départ communément admis est l'axiomatique de la théorie des ensembles de Zermelo-Fraenkel avec axiome du choix (ZFC) [1].

On suppose donné le calcul des prédicats avec égalité ($=$). De plus on va utiliser deux symboles particuliers

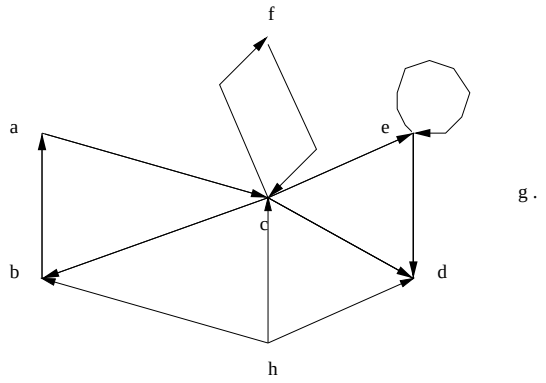
$$\in \quad \text{et} \quad \emptyset \, .$$

Le premier dénote une relation binaire, le deuxième dénote un ensemble particulier. Les axiomes que nous allons énoncer vont cerner le sens à donner à ces symboles : $x \in y$ se lira "l'ensemble x appartient à l'ensemble y" et $\emptyset$ sera l'ensemble vide. Noter que *a priori* la relation $\in$ pourrait être décrite par un graphe comme le suivant, où une flèche $a \to b$ représente $a \in b$.

On verra par exemple, que contrairement à ce qui se passe (en e) pour la relation décrite par le graphe, pour aucun ensemble x on n'aura $x \in x$. De même pour les ensembles on n'aura pas de cycle comme pour a, c et b. Par contre les ensembles x, y et z pour lesquels $x \in y$, $y \in z$ et $x \in z$ (comme pour c, e et d) jouent un rôle très important [2].

[1] "Z" pour E. Zermelo (1871-1953), "F" pour Fraenkel (1891-1965) et "C" pour Axiome du choix. Pour un autre traitement rapide de la théorie des ensembles voir l'appendice de : G. Godefroy "L'aventure des nombres" Ed. Odile Jacob, Sciences, Paris, 1997, ainsi que R. Godement : "Analyse mathématique. I. Convergence, fonctions élémentaires." Springer-Verlag, Berlin, 1998. xx+432 pp. ISBN : 3-540-63212-3 ; une présentation plus détaillée se trouve dans J.-L. Krivine, "Théorie axiomatique des ensembles", P.U.F., Paris 1969, qui a été repris dans le livre du même auteur "Théorie des ensembles", Cassini, Paris, 1999. Voir aussi : Y. Gauthier, "Logique et fondements des mathématiques", Diderot Multimedia, Paris, 1997.

[2] Dans la suite on pensera aux variables comme parcourant un univers U que les axiomes décrivent. Une fois énoncés les axiomes, on pourra préciser U.



Dire qu'un ensemble est une collection d'objets ne mène pas très loin (qu'est-ce qu'un objet, ou une collection ?). Le fait est que l'on ne va pas dire ce qu'est un ensemble, mais *on dit quand deux ensembles sont égaux* :

Axiome 1 : (a. d'extensionnalité)

$$\forall x \forall y \ \ (\forall z \ (z \in x \Leftrightarrow z \in y) \Rightarrow x = y) \ .$$

C'est-à-dire qu'un ensemble est déterminé par ses éléments (et par rien d'autre).

On écrit

$$x \subset y$$

pour abréger l'énoncé $\forall z \ (z \in x \Rightarrow z \in y)$, que l'on lit "l'ensemble x est un sous-ensemble de l'ensemble y". Ainsi un ensemble x est sous-ensemble d'un ensemble y, si tous les éléments de x sont éléments de l'ensemble y et on voit que

$$x = y \Leftrightarrow (x \subset y) \wedge (y \subset x) \ .$$

Si $x \subset y$ on dit aussi que x est une *partie de* y.

A retenir : pour vérifier si deux ensembles sont égaux, vérifier si ils se contiennent mutuellement.

Le prochain axiome définit le sens du symbole $\emptyset$:

Axiome 2 : (a. du vide)

$$\forall x \ \neg(x \in \emptyset) \ .$$

On postule donc l'existence d'un ensemble qui ne contient aucun élément. Ceci correspond à la situation des points g et h de la figure. En fait l'axiome d'extensionnalité montre que l'ensemble vide est l'unique ensemble à ne contenir aucun élément.

Comme déjà dit, les axiomes peuvent être considérés comme spécifiant les règles de construction des ensembles : la difficulté majeure dans le choix des axiomes est qu'ils doivent permettre la construction d'un nombre suffisant d'ensembles sans en construire de trop gros. Par exemple on ne saurait quoi penser de l'ensemble des ensembles qui ne se contiennent pas. S'il se contenait, alors... il ne le ferait pas et réciproquement [3] ! On doit donc procéder avec attention. Cependant, il est clair que d'une manière ou d'une autre on aimerait, par exemple, retrouver l'ensemble des entiers naturels.

[3] C'est là le *paradoxe de Russel*, qui met en garde devant la tentation de dire que la collection des objets ayant une même propriété forment un ensemble (voir l'axiome de séparation plus bas). Notons aussi que, comme nous le verrons, pour aucun ensemble x on a $x \in x$, donc si la collection de tous les ensembles qui ne se contiennent pas était un ensemble elle contiendrait tous les ensembles : on ne peut donc pas considérer l'ensemble de tous les ensembles.

On commence plus modestement en définissant les ensembles à un et à deux éléments. En fait il suffit de demander l'existence des paires. Étant donné des ensembles x et y on a le droit de considérer l'ensemble qui ne contient que ces ensembles :

Axiome 3 : (formation des paires)

$$\forall x \forall y \exists t \ (\forall z \ (z \in t \Leftrightarrow (z = x) \vee (z = y))) \ .$$

L'ensemble t, la *paire*, dont on postule l'existence est noté

$$\{x, y\} \ .$$

On définit le *singleton* $\{x\}$ comme étant l'ensemble $\{x, x\}$, c'est l'ensemble qui ne contient que l'ensemble x. On peut aussi définir le *couple ordonné* (x, y) par

$$(x, y) = \{\{x\}, \{x, y\}\} \ .$$

A retenir : le couple ordonné (x, y) a la propriété d'être déterminé par ces *composantes* x et y, plus précisément on a l'énoncé suivant.

Proposition. Soient x, y, u et v ensembles. Alors :

$$(x, y) = (u, v) \Leftrightarrow ((x = u) \wedge (y = v)) \ .$$

La démonstration de la proposition est laissée en exercice ; elle se fait par épuisement (on analyse toutes les possibilités).

Notons que l'on peut déjà construire beaucoup d'ensembles :

$$\emptyset \ , \ \{\emptyset\} \ , \ \{\emptyset, \{\emptyset, \{\emptyset\}\}\} \ , \ \{\{\emptyset\}\} \ , \ \{\{\emptyset\}, \{\emptyset\}\} \ \ldots$$

Intuitivement on voit que l'on en a déjà une infinité ! Essayons de définir une suite infinie d'ensembles. Voici une tentative :

$$\emptyset \ , \ \{\emptyset\} \ , \ \{\{\emptyset\}\} \ , \ \{\{\{\emptyset\}\}\} \ , \ \text{etc.}$$

Mis à part le premier ce sont là des ensembles à un élément. Ils sont tous différents deux à deux. Mais pour l'instant on ne peut pas les réunir dans un même ensemble : on n'a pas encore d'ensemble infini.

Axiome 4 : (a. de l'union)

$$\forall x \exists t \ (\forall z \ (z \in t \Leftrightarrow \exists y \ (y \in x \wedge z \in y))) \ .$$

Cet axiome permet de construire à partir d'un ensemble donné x un ensemble t, dont les éléments sont les éléments des éléments de x. On note

$$\bigcup x$$

cet ensemble, que l'on appelle la *réunion sur* x. En termes imagés si on pense à x comme étant une collection de sachets (de café, de riz, ...), alors $\bigcup x$ est le résultat de l'opération qui consiste à découper tous les sachets et tout collecter dans un unique récipient.

On définit

$$a \cup b := \bigcup \{a, b\} \ .$$

A retenir : $z \in a \cup b$ si et seulement si $z \in a$ ou $z \in b$.

Ainsi, sait donc faire l'union d'une collection d'ensembles, à condition que ces ensembles soient déjà dans un ensemble. On ne peut toujours pas réunir les ensembles de la suite précédente en un ensemble.

Voici une intuition liée à la suite des nombres naturels : on les passe tous en revue en "ajoutant un" (et on ne s'arrête jamais). L'axiome de l'union nous permet de faire une construction très intéressante. L'*opération successeur* est définie par

$$s(x) := x \cup \{x\} \ .$$

Appliquons cette opération à l'ensemble vide (déjà défini), on obtient

$$s(\emptyset) = \emptyset \cup \{\emptyset\} = \{\emptyset\} \quad , \quad s(\{\emptyset\}) = \{\emptyset\} \cup \{\{\emptyset\}\} = \{\emptyset, \{\emptyset\}\} \ , \quad \text{etc.}$$

C'est-à-dire que l'on obtient la suite

$$\emptyset \quad , \quad \{\emptyset\} \quad , \quad \{\emptyset, \{\emptyset\}\} \quad , \quad \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\} \quad , \quad \text{etc.}$$

On part donc de l'ensemble vide pour arriver à un un ensemble à un, puis deux, puis trois éléments, puis à quatre, etc. En quoi est-ce que la deuxième suite est meilleure que la première ? Tout d'abord les ensembles successifs de la deuxième suite contiennent de plus en plus d'éléments (le n-ième en contient n). De plus, on voit que par exemple $\emptyset$ est contenu dans tous les suivants et en fait un ensemble dans cette suite contient tous ses prédécesseurs. C'est-à-dire que la relation d'appartenance $\in$ correspond ici à l'ordre usuel sur les entiers $<$ ("plus petit") [4]. Ainsi on aurait une façon de définir tous les entiers, de manière à ce qu'ils soient reliés par l'opération successeur, ces entiers sont ordonnés (par $\in$), mais... ce que nous ne pouvons toujours pas faire est de considérer ces entiers comme éléments d'un ensemble unique. Pour cela il nous faut en fait imposer un axiome.

Axiome 5 : (a. de l'infini)

$$\exists x \ \ ((\emptyset \in x) \wedge \ \forall y (y \in x \Rightarrow s(y) \in x)) \ .$$

En mots : on décrète qu'il existe un ensemble, qui contient l'ensemble vide et qui est tel que si il contient un ensemble, alors il contient aussi le successeur de cet ensemble (on dit qu'il est *héréditaire* [5]). On voit donc que la deuxième suite ci-dessus appartient toute entière à n'importe quel ensemble héréditaire. Il est clair que l'on s'attend à ce que l'ensemble des entiers naturels soit héréditaire, mais on n'aimerait pas retrouver parmi les entiers naturels autre chose que des objets obtenus par l'opération successeur à partir d'un zéro choisi. Heureusement on peut montrer qu'il existe un ensemble héréditaire qui est contenu dans tout ensemble héréditaire. *Cet ensemble est l'ensemble des entiers naturels.* D'habitude en mathématiques on le note

$$\mathbf{N} \ .$$

A retenir : l'ensemble $\mathbf{N}$ des entiers naturels est l'ensemble héréditaire contenu dans tout ensemble héréditaire.

On aurait pu espérer construire $\mathbf{N}$ à partir de l'ensemble vide, mais l'on ne sait pas faire une telle construction. Cependant il faut bien comprendre que ce qui a été fait ici est très loin de se donner l'ensemble des entiers naturels *avec toutes ces propriétés.* Ce que l'on a mis en avant est ce qui s'est révélé être l'une des propriétés essentielles de cet ensemble. A l'aide de cette unique propriété on obtiendra tout le reste (par exemple les opérations d'addition et de multiplication sur les entiers).

Parmi les autres axiomes de la théorie des ensembles, le suivant est celui qui donne la plus grande liberté dans la définition d'ensembles nouveaux à partir d'ensembles déjà connus. Nous l'énonçons dans un cas particulier [6].

[4] On dit qu'un ensemble x est un *ordinal* si $\in$ est sur x une relation d'ordre total strict, qui est un bon ordre et si $z \in x \Rightarrow z \subset x$. On voit donc que les éléments de la deuxième suite sont des ordinaux.

[5] On peut montrer que tout ensemble héréditaire est infini, et plus précisément qu'il contient un sous-ensemble strict ayant le même nombre d'éléments (peut-être infini) que lui-même. Penser à l'ensemble des entiers naturels et à son sous-ensemble formé des entiers pairs.

[6] L'*axiome de substitution* dont l'axiome suivant est un cas particulier, est plus difficile à formuler. Nous nous en dispensons pour l'instant. Noter que cet axiome permet de définir l'ensemble vide et la paire, les axiomes 2 et 3 sont donc redondants.

Axiome 6 : (a. de séparation [7]) Soit $P(z,...)$ une propriété [8] pouvant être vraie de l'ensemble z, alors

$$\forall x \exists y \ \forall z \ \ ((z \in y) \Leftrightarrow ((z \in x) \wedge P(z,...))) \ .$$

Cet axiome permet, étant donné un ensemble x et une propriété P, de former le *sous-ensemble* y de x formé des éléments z de x qui vérifient la propriété P. Cet ensemble est noté

$$\{z \in x \ : \ P(z,...)\} \ .$$

On dit aussi que y est défini par *compréhension* [9].

Exemples. 1) Soient x et y des ensembles ; l'ensemble *différence* (aussi appelé le *complémentaire* de y dans x) est l'ensemble

$$x \setminus y := \{z \in x \ : \ \neg(z \in y)\} \ .$$

(Ceci a un sens même si y n'est pas un sous-ensemble de x.)

2) Soit x un ensemble. L'*intersection* sur x est l'ensemble

$$\bigcap x := \{z \in \bigcup x \ : \ \forall y \in x \ (z \in y)\} \ .$$

Cet ensemble est aussi noté $\bigcap_{y \in x} y$. Il s'agit donc de l'ensemble dont les éléments sont les éléments de la réunion $\bigcup x$ sur x, qui sont éléments de tous les éléments de x à la fois (du café, du riz,... tout à la fois). Un cas particulier est celui où $x = \{a, b\}$ est la paire formée des ensembles a et b. On note alors

$$a \cap b := \bigcap \{a, b\} = \{z \in a \cup b \ : \ (z \in a) \wedge (z \in b)\} \ .$$

A retenir : $a \cap b$ est donc l'ensemble des éléments de a qui sont aussi éléments de b (et réciproquement).

3) Avec ces définitions on peut par exemple vérifier les identités

$$x \cap (y \cup z) = (x \cap y) \cup (x \cap z) \ \ , \ \ x \cup (y \cap z) = (x \cup y) \cap (x \cup z) \ .$$

La subtilité de l'axiome de séparation réside dans le fait qu'il permet seulement de "séparer/mettre en évidence" une partie d'un ensemble donné à l'aide d'une propriété. On aurait des problèmes si on voulait qu'une propriété seule définisse un ensemble. Il suffit de reprendre l'exemple déjà considéré, où $P(z)$ signifie "$\neg(z \in z)$". On obtiendrait alors l'ensemble des ensembles qui ne se contiennent pas et, comme on l'a vu, une contradiction.

Pour que l'axiome de séparation soit encore plus puissant, on se permet de considérer l'*ensemble des parties* d'un ensemble, que l'on introduit comme suit.

Axiome 7 : (a. de l'ensemble des parties)

$$\forall x \exists y \ \forall z \ (z \in y \Leftrightarrow z \subset x) \ .$$

Étant donné x il existe donc un ensemble y dont les éléments sont les sous-ensembles (ou parties) de x. On note cet ensemble

$$\mathcal{P}(x) \ .$$

[7] En fait, pour l'axiome de séparation il faudrait parler d'un *schéma d'axiomes*, en effet on a un axiome pour chaque P.

[8] Ici et plus loin nous utilisons une terminologie imagée pour dénoter ce que l'on appelle aussi un énoncé à une variable libre.

[9] On affirme souvent qu'il y a deux façons de se donner un ensemble : par *extension* (en faisant la liste de tous ces éléments) ou alors par *compréhension*. A la lumière de la théorie que nous sommes en train d'exposer, cette affirmation est au mieux une approximation de la réalité des faits.

Exemple.

$$\mathcal{P}(\emptyset) = \{\emptyset\} \quad , \quad \mathcal{P}(\{\emptyset\}) = \{\emptyset, \{\emptyset\}\} \quad , \quad \mathcal{P}(\{\emptyset, \{\emptyset\}\}) = \{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}, \{\emptyset, \{\emptyset\}\}\}$$

Noter que l'ensemble vide est sous-ensemble de tout ensemble (vu qu'il ne contient aucun élément il n'y a rien à vérifier !). Aussi, un ensemble à n éléments aura un ensemble de parties à 2^n éléments [10] .

Le *produit cartésien* des ensembles x et y est défini par

$$x \times y := \{z \in \mathcal{P}(\mathcal{P}(x \cup y)) \ : \ \exists u \in x \ \exists v \in y \ (z = (u, v))\} \ .$$

Il s'agit donc de l'ensemble de tous les couples ordonnés dont la première composante est dans x et la deuxième est dans y. Noter comment nous avons précisé l'ensemble où se trouvent les couples ordonnés (u, v) [11].

Les axiomes que nous avons formulés jusqu'ici sont les axiomes de Zermelo et Fraenkel (ZF). On complète ce système d'axiomes avec deux autres :

Axiome 8 : (a. de fondation)

$$\forall x \ \ (\neg(x = \emptyset) \Rightarrow (\exists y \in x \ \wedge \ (y \cap x = \emptyset)) \ .$$

Cet axiome a pour conséquence que pour tout ensemble x on a $\neg(x \in x)$ et que l'opération successeur est injective, c'est-à-dire que, si $s(x) = s(y)$, alors $x = y$ (voir plus loin) [12].

Axiome 9 : (a. du choix) Cet axiome dit que si x est un ensemble non-vide d'ensembles non-vides, alors pour tout élément z de x on peut choisir un élément $c(z)$ dans x, plus précisément et avec le vocabulaire que nous allons introduire sous peu, il existe une fonction $c : x \to \bigcup x$ telle que pour tout z dans x on a $c(z) \in z$ (c est une "fonction choix").

Cet axiome a donné lieu a beaucoup de polémiques car il permet par exemple de montrer que *tout* ensemble admet un bon ordre, c'est-à-dire un ordre pour lequel tout sous-ensemble non-vide possède un plus petit élément (comme $\leq$ pour les entiers naturels). Pour un exemple d'utilisation de cet axiome voir les propriétés des fonctions décrites plus loin dans ce paragraphe et le chapitre sur l'aire des figures planes.

Aussi étonnant que cela puisse paraître les neufs axiomes précédents suffisent pour construire tous les objets mathématiques (d'usage courant) [13]. En particulier ils permettent de construire toute l'arithmétique élémentaire. Le système déductif ainsi obtenu est donc "assez riche" et à la lumière du théorème de

[10] A ce stade en principe on ne sait pas encore compter et ceci est une des définitions possibles de 2^n ; c'est une bonne définition, elle explique par exemple pourquoi 2^0 vaut 1.

[11] Vu que $(u, v) = \{\{u\}, \{u, v\}\}$, que $\{u\} \subset x \subset x \cup y$ et que $\{u, v\} \subset x \cup y$, on a bien $(u, v) \in \mathcal{P}(\mathcal{P}(x \cup y))$.

[12] Une autre conséquence de l'axiome de fondation est qu'il permet d'identifier l'univers (à l'univers de von Neumann).

[13] Comme pour tout système déductif, on peut donner pour la théorie des ensembles différents systèmes d'axiomes plus ou moins équivalents. Notons que certains choix entraînent des conséquences pour le moins bizarres : ainsi il a été calculé que le système proposé par N. Bourbaki dans son "Théorie des ensembles" (Hermann, Paris, 1971) demanderait un nombre immense de symboles pour écrire 1. Plus précisément on a calculé qu'il faudrait $4{\cdot}523{\cdot}659{\cdot}424{\cdot}929$ termes, soit un million de livres d'un millier de pages (avec 50 lignes par page et 80 symboles par ligne ; voir les papiers de A.R.D. Mathias sur le serveur *www.dpmms.cam.ac.uk/* $\sim$ *ardm*). En fait les recherches en théorie des ensembles continuent : ainsi par exemple à la suite des travaux de Gödel et de la démonstration par Cohen dans les années soixante de l'indépendance de ZFC de l'Hypothèse du continu, on cherche à trouver des "axiomes de grands infinis", qui soient non-contradictoires et qui permettraient de fournir d'autres outils de démonstration, plus puissants que ceux dont on dispose. Nous allons rencontrer l'Hypothèse du continu dans le chapitre sur les nombres réels.

Gödel on peut se demander si le système déductif ainsi obtenu est consistant, c'est-à-dire si on a la garantie que jamais on n'arrivera sur ces bases à démonter un énoncé et sa négation. On ne sait pas montrer que le système ZFC est consistant, mais on sait par exemple montrer que ZFC n'est pas plus inconsistant que ZF : le problème ne vient pas de l'axiome du choix.

4.3 Démonstrations par récurrence et applications.

Nous allons voir comment utiliser l'axiome de l'infini pour fonder le principe de démonstration par récurrence.

Le fait que tout ensemble héréditaire a contienne $\mathbf{N}$ se traduit par

$$\forall a \ (\emptyset \in a \ \wedge \ \forall y \ (y \in a \Rightarrow s(y) \in a)) \Rightarrow (\mathbf{N} \subset a) \ .$$

Si on applique ceci à l'ensemble défini par compréhension $a = \{x \in \mathbf{N} \ : \ P(x)\}$ pour P une propriété on obtient le

Principe de démonstration par récurrence. (1-ère forme)

$$[P(\emptyset) \wedge \forall y \in \mathbf{N} \ (P(y) \Rightarrow P(s(y)))] \Rightarrow (\forall x \in \mathbf{N} \ P(x)) \ [14]$$

A retenir : si on identifie 0 à l'ensemble vide et $s(n)$ à $n+1$, on obtient le principe qui consiste à dire que pour montrer qu'un énoncé $P(n)$ qui dépend d'un entier n est vérifié pour tout n, il suffit de vérifier deux choses : (1) que $P(0)$ est vérifié et (2) que pour tout entier n l'implication $P(n) \Rightarrow P(n+1)$ est vérifiée.

Exemples. Nous avons déjà rencontré ce type de démonstration dans la discussion de la formule du binôme. Voici d'autres exemples d'énoncés que l'on peut montrer par récurrence :

1) Soit $n \geq 1$ un entier naturel, alors

$$1^3 + 2^3 + \cdots + n^3 = (1 + \cdots + n)^2 \ .$$

(Indication : on pourra utiliser l'identité $2(1 + \cdots + n) = n(n+1)$, que l'on peut aussi démontrer par récurrence.)

2) Soit n un entier naturel, alors la somme des n premiers nombres impairs égale n^2.
3) Pour tout entier naturel $n \geq 1$ on a $2^n \geq n + 1$.
4) *Opérations sur les entiers.* On utilise une démonstration par récurrence pour montrer l'existence des opérations usuelles sur les entiers naturels et de leurs propriétés. Une présentation plus claire de ceci sera possible une fois introduite la notion de fonction définie récursivement (voir plus bas).

Une deuxième forme du principe de démonstration par récurrence est la suivante :

Principe de démonstration par récurrence. (2-ème forme)

$$(\forall x \in \mathbf{N} \ ((\forall y \in x \ P(y)) \Rightarrow P(x))) \Rightarrow (\forall x \in \mathbf{N} \ P(x)) \ .$$

Sous cette forme le principe dit que pour vérifier $P(n)$ pour tout n il suffit de voir que pour un m quelconque $P(m')$ pour tout $m' < m$ entraîne $P(m)$ [15].

La propriété du bon ordre.

[14] Ici et dans la suite on écrit $\forall x \in E \ Q(x)$ pour $\forall x (x \in E \wedge Q(x))$.

[15] On rédige la déduction du principe de démonstration par récurrence sous sa deuxième forme du principe sous sa première forme.

La proposition qui suit dit que **N** est bien ordonné, il s'agit là d'une ultérieure traduction du fait que **N** est le plus petit ensemble héréditaire.

Proposition. Tout sous-ensemble non-vide de l'ensemble des entiers naturels admet un plus petit élément.

Démonstration. Vu que l'ordre sur **N** est donné par $\in$ cela s'obtient comme suit. On prend la contraposée de la deuxième forme du principe de démonstration par récurrence (où l'on remplace P par S)

$$\neg(\forall x \in \mathbf{N}\ S(x)) \Rightarrow \neg(\forall x \in \mathbf{N}\ ((\forall y \in x\ S(y)) \Rightarrow S(x))) .$$

C'est-à-dire

$$\exists x \in \mathbf{N}\ \neg S(x) \Rightarrow \exists x \in \mathbf{N}(\neg S(x) \wedge (\forall y \in x\ S(y))) .$$

En particulier, si $P = \neg S$

$$\exists x \in \mathbf{N}\ P(x) \Rightarrow \exists x \in \mathbf{N}\ (P(x) \wedge (\forall y \in x\ \neg P(y))) ,$$

et pour $P(x) = x \in a$, avec $a \subset \mathbf{N}$

$$\neg(\emptyset = a) \wedge a \subset \mathbf{N} \Rightarrow \exists x \in a\ \wedge\ \forall y \in x\ \neg(y \in a) .$$

Ceci se lit bien comme "tout ensemble non-vide a contenu dans **N** contient un élément x tel que tout (autre) élément y de **N** plus petit que x n'est pas dans a".

Remarques. 1) On peut aussi montrer que la propriété du bon ordre implique le principe de démonstration par récurrence, mais cela n'a pas un grand intérêt à ce stade vu que ce principe est donné avec **N** (comme on l'a vu).

2) Le bon ordre sur **N** est ce qui est à la base de l'*algorithme d'Euclide* pour les entiers.

4.4 Relations et fonctions : vocabulaire.

Le but ici est d'abord de définir une fonction entre deux ensembles comme un cas particulier de relation entre ces ensembles et ensuite de mettre en évidence quelques types de fonctions. Ce travail nous permettra de préciser ce que l'on entend par cardinal/nombre d'éléments d'un ensemble. Dans toute la suite les mots "fonction" et "application" seront considérés comme synonymes.[16]

Relations. Une *relation* R entre des ensembles a et b est un sous-ensemble (quelconque) du produit cartésien $a \times b$, c'est donc un ensemble de couples. Le *domaine* (resp. l'*image*) d'une telle relation R est l'ensemble des premières (resp. deuxièmes) composantes des éléments de R. On note le domaine et l'image d'une relation R par

$$\mathrm{dom}(R) \quad \text{et} \quad \mathrm{im}(R) .$$

Soit $Q(x)$ la proposition $(\forall y \in x\ P(x))$. On doit montrer

$$(\forall x \in \mathbf{N}\ (Q(x) \Rightarrow P(x))) \Rightarrow (\forall x \in \mathbf{N} : P(x)) \quad (*) .$$

On observe que $(\forall x \in \mathbf{N}\ Q(x)) \Rightarrow (\forall x \in \mathbf{N}\ P(x))$. En effet, par définition $\forall x \in \mathbf{N}\ Q(x)$ signifie $\forall x \in \mathbf{N}\ \forall y \in x\ P(y)$, ce qui implique $\forall y \in \mathbf{N}\ P(y)$ car pour tout y de **N** il existe x de **N** tel que $y \in x$ (par exemple $x = s(y)$).

Il suffit donc de montrer $\forall x \in \mathbf{N}\ Q(x)$ à partir de l'antécédent de (*). On utilise l'antécédent de (*) pour faire une démonstration par récurrence de $Q(x)$. Tout d'abord $Q(\emptyset)$ est vrai car il n'y a aucun y élément de $\emptyset$. Il faut maintenant voir si le pas de récurrence $Q(x) \Rightarrow Q(s(x))$ est vrai. Par définition $s(x) = x \cup \{x\}$, donc $Q(s(x))$ est $(\forall y \in x \cup \{x\}\ P(y))$. On sait par l'antécédent de (*) que $\forall x \in \mathbf{N} : Q(x) \Rightarrow P(x)$, qui vérifie le pas de récurrence pour $y = x$, c'est-à-dire $y \in \{x\}$. Aussi, si $y \in x$, alors on a $P(y)$, car on suppose $Q(x)$. Ceci termine la démonstration.

[16] L'approche abstraite exposée ici est aussi justifiée par le fait que, par exemple en géométrie on est naturellement amené à utiliser des fonctions, ou correspondances, qui servent à classer des familles d'objets d'une même nature : voire la description paramétrique de l'ensemble des droites dans le plan.

On peut définir la *composition* de relations R et S comme étant la relation

$$S \circ R := \{(x,y) \in \mathrm{dom}(R) \times \mathrm{im}(S) \ : \ [\exists z \ (((x,z) \in R) \wedge ((z,y) \in S))]\} \ .$$

La *relation identité* sur un ensemble a est la relation

$$id_a := \{(x,x) \in a \times a : x \in a\} \ .$$

La *relation réciproque* (ou *inverse*) d'une relation R est la relation

$$R^{-1} := \{(x,y) \in \mathrm{im}(R) \times \mathrm{dom}(R) \ : (y,x) \in R\} \ .$$

Noter que $\mathrm{dom}(R^{-1}) = \mathrm{im}(R)$ et $\mathrm{im}(R^{-1}) = \mathrm{dom}(R)$.

Fonctions. Écrivons

$$\exists! x \ Fx$$

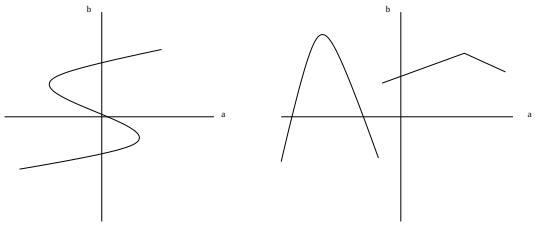
pour signifier "il existe un unique x tel que Fx". Une relation f entre deux ensembles a et b est une *fonction* [17] (de a dans b) si

$$\forall x \in \mathrm{dom}(f) \ \exists! \ y \ [((y \in b) \wedge \ ((x,y) \in f)] \ .$$

Autrement dit

$$\forall x \in \mathrm{dom}(f) \ \forall y \forall z \ \ ((x,y) \in f \wedge (x,z) \in f \Rightarrow (y = z)) \ .$$

La figure qui suit donne une représentation schématique de deux relations entre des ensembles a et b. La première n'est pas une fonction, la deuxième en est une [18].



Si $a = \mathrm{dom}(f)$ on écrira

$$f : a \to b \ \ \text{ou} \ \ \ a \overset{f}{\to} b \ ,$$

pour signifier que f est une fonction de a dans b *et* que $a = \mathrm{dom}(f)$. Par définition, pour chaque x élément de a il existe un unique élément y de b tel que (x,y) soit dans f. On écrira : $y = f(x)$ ou $x \mapsto y$.

---

[17] Avec le vocabulaire que nous introduisons ici nous pouvons énoncer l'*axiome de substitution*, qui généralise l'axiome de séparation. Un énoncé $E(x,y,x_1,\ldots,x_k)$ défini une *relation fonctionnelle* à un argument (ici entre x et y), si pour tout choix de x_1, ..., x_k et tout choix de x, y et y', l'égalité $E(x,y,x_1,\ldots,x_k) = E(x,y',x_1,\ldots,x_k)$ entraîne $y = y'$. Comme pour les fonctions, on peut définir le domaine de, et l'image d'un élément par, une telle relation fonctionnelle. L'axiome de substitution demande alors, que donné a un ensemble quelconque et E une relation fonctionnelle, il existe un ensemble b dont les éléments sont exactement les images par la relation fonctionnelle E des éléments de a, qui se trouvent dans le domaine de E. En symboles :

$$\forall a \exists b \ \forall y \, (y \in b \ \Leftrightarrow \ \exists x \ ((x \in a) \wedge E(x,y,x_1,\ldots,x_k))) \ .$$

L'axiome de séparation est le cas particulier où l'on considère au lieu de E un énoncé $P(x,x_1,\ldots,x_k)$ à une seule variable libre (avec en quelque sorte y constant).

[18] On voit que la définition que nous avons donné d'une fonction f revient à identifier f à ce que l'on appelle d'habitude son *graphe*. Du coup une fonction n'a pas un graphe elle *est* son graphe.

A retenir : $f = g$ si et seulement si $\mathrm{dom}(f) = \mathrm{dom}(g)$ et $\forall x \in \mathrm{dom}(f)\ :\ f(x) = g(x)$.

Injectivité, surjectivité et bijectivité. Soit $f : a \to b$ une fonction. Si $\mathrm{im}(f) = b$ on dira que f est *surjective*, noté

$$f : a \twoheadrightarrow b$$

On dira que $f : a \to b$ est *injective* si pour x et y dans $\mathrm{dom}(f)$ on a l'implication $f(x) = f(y) \Rightarrow x = y$. On note

$$f : a \hookrightarrow b \qquad \text{ou} \qquad f : a \rightarrowtail b$$

Une fonction f est *bijective* si elle est injective et surjective à la fois.

En termes imagés, une fonction $f : a \to b$ est injective si elle *jette a dans b*, c'est-à-dire si elle permet de retrouver a dans b : si $x \neq y$, alors $f(x) \neq f(y)$. Si la fonction f est injective, alors elle établit une bijection de a sur son image par f. Une fonction $f : a \to b$ est surjective si elle *jette a sur b*. Pour tout élément y de b, on a au moins un élément x de a qui est "envoyé sur" y par f, c'est-à-dire tel que $f(x) = y$.

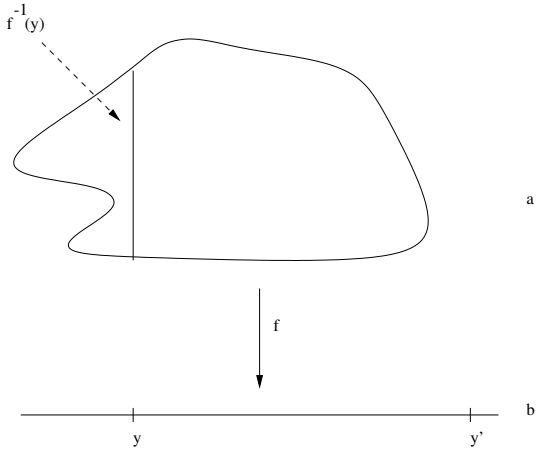
L'*image réciproque* d'un sous-ensemble b' de b par une fonction $f : a \to b$ est

$$f^{-1}(b') := \{x \in a\ :\ f(x) \in b'\}\ .$$

En particulier la *fibre* en $y \in b$ d'une fonction $f : a \to b$ est l'ensemble $f^{-1}(y) := f^{-1}(\{y\})$, en clair

$$f^{-1}(y) := \{x \in a\ :\ f(x) = y\}\ .$$

Une représentation graphique de la fibre d'une fonction est donnée sur la figure suivante, où la fibre de y' est vide.



4.5 Fonctions : propriétés.

On peut montrer que la composition de deux fonctions est encore une fonction.

A retenir : si $f : a \to b$ et $g : b \to c$ sont des fonctions, alors par définition $g \circ f : a \to c$ est la fonction telle que pour x élément de a

$$(g \circ f)(x) = g(f(x))\ .$$

Lemme. Soit $f : a \to b$ une fonction. Sont équivalents :

1) La fonction f est injective.
2) La relation réciproque f^{-1} est une fonction.
3) Chaque fibre de f contient au plus un élément.

Lemme. Soit $f : a \to b$ une fonction. Sont équivalents :
1) La fonction f est surjective.
2) Chaque fibre de f est non-vide.

Voici une liste de quelques autres propriétés des fonctions. Il va sans dire, que chacune d'entre elles peut se démontrer... Dans ce qui suit, soit $f : E \to F$, $g : F \to G$ et $h : G \to H$ des fonctions.

a) *Associativité de la composition* : $h \circ (g \circ f) = (h \circ g) \circ f$.
b) Si g et f sont injectives, alors $g \circ f$ est injective.
c) Si g et f sont surjectives, alors $g \circ f$ est surjective.
d) Si g et f sont bijectives, alors $g \circ f$ est bijective.
e) L'application $f : E \to F$ est injective si et seulement si il existe une application $\widetilde{f} : F \to E$ telle que $\widetilde{f} \circ f = id_E$.
f) L'application $f : E \to F$ est surjective si et seulement si il existe une application $\widetilde{f} : F \to E$ telle que $f \circ \widetilde{f} = id_F$.
g) L'application $f : E \to F$ est bijective si et seulement si il existe une application $\widetilde{f} : F \to E$ telle que $\widetilde{f} \circ f = id_E$ et $f \circ \widetilde{f} = id_F$ (dans ce cas l'application $\widetilde{f}$ s'appelle l'*inverse* de f ; c'est encore une bijection).

Remarques : i) Pour démontrer la nécessité de la condition du (f), on utilise l'*Axiome du choix* et pour définir $\widetilde{f}$ on choisit un élément dans chaque fibre de f.

ii) On peut généraliser le point (e) ainsi :

Lemme de factorisation. Soit $f : E \to F$ une surjection et soit $h : E \to H$ une application. Il existe $g : F \to H$, telle que $h = g \circ f$ si et seulement si pour tout x, x' de E $(f(x) = f(x') \Rightarrow h(x) = h(x'))$.

(Pour le voir : on définit g comme étant

$$g = \{(y, z) \in F \times H : \exists x \in E : f(x) = y \ \ , \ \ h(x) = z\} \ .)$$

Dans quel sens est-ce que ce lemme généralise le point (e) ?

4.6 Exemples.

Avec le matériel dont nous disposons, nous pouvons déjà considérer quelques fonctions intéressantes et construire des ensembles d'un type nouveau.

La fonction successeur.

L'axiome de fondation entraîne que la fonction successeur

$$\begin{array}{rcl} s : \mathbf{N} & \to & \mathbf{N} \\ n & \mapsto & s(n) =: n + 1 \end{array}$$

est *injective*. (Ici on a noté $s(n) = n \cup \{n\} = n+1$.) Vu que 0 n'est successeur d'aucun entier, la fonction successeur n'est pas surjective et donc pas bijective.

Ensembles d'applications.

Soient a et b des ensembles. Alors on peut considérer l'ensemble

$$b^a$$

de toutes les applications $f : a \to b$. C'est une partie de l'ensemble des parties de $a \times b$.

Une construction semblable est celle du produit d'une famille d'ensembles. On appelle *famille d'ensembles indexée par un ensemble I* une application a de domaine I. On note $a_i := a(i)$ et $(a_i)_{i\in I}$ pour la famille. A partir d'une telle famille on considère alors la *réunion de la famille* $(a_i)_{i\in I}$

$$\bigcup_{i\in I} a_i := \bigcup im(a)$$

et, si $I \neq \emptyset$, l'*intersection de la famille* $(a_i)_{i\in I}$

$$\bigcap_{i\in I} a_i := \{x \in a_{i_0} \ : \ \forall i \ (i \in I \Rightarrow x \in a_i)\}$$

où i_0 est un quelconque élément de I. Notons que x appartient à $\bigcup_{i\in I} a_i$ si et seulement si il existe i dans I tel que $x \in a_i$. Le *produit de la famille* $(a_i)_{i\in I}$ est alors défini comme étant l'ensemble

$$\prod_{i\in I} a_i$$

des fonctions $f : I \to \bigcup \mathrm{im}(a)$, telle que pour tout i on a $f(i) \in a_i$. C'est un sous-ensemble de l'ensemble $(\bigcup_{i\in I} a_i)^I$.

Exemples : a) soit n un entier et a un ensemble, alors a^n peut être vu comme l'ensemble des applications de n (identifié à un ensemble à n éléments) dans a, ou comme un produit de n copies de a (la famille est la famille constante $a_i = a$).

b) Soit a un ensemble. Les éléments de $a^{\mathbf{N}}$ s'appellent *suites dans a*. Une suite est une famille d'éléments de a et on note souvent $(a_n)_{n\in\mathbf{N}}$ la suite telle que $n \mapsto a_n$.

Fonctions définies récursivement.

Nous allons voir comment définir les opérations sur les entiers naturels. On utilise le résultat général suivant.

Théorème-Définition. Soit E un ensemble non-vide, a un élément de E et $\varphi : E \to E$ une fonction. Il existe alors une et une seule fonction

$$f : \mathbf{N} \to E$$

telle que

1) $f(0) = a$

2) Pour tout n on a $f(s(n)) = \varphi(f(n))$.

Une telle fonction f est dite *définie récursivement* (ou plus simplement *récursive* [19]).

Remarques : (i) on ne peut pas simplement définir $f(n)$ comme étant $\varphi^{n-1}(a)$, la composée de φ avec elle-même $(n-1)$-fois, évaluée en a. En effet un tel type de composition se définit à partir du théorème.

(ii) Observons que le théorème est valable en remplaçant partout 0 par 1 et $\mathbf{N}$ par $\mathbf{N}^* = \mathbf{N} \setminus \{0\}$.

[19] Les fonctions que nous venons de construire sont un cas particulier des très importantes fonctions récursives générales. Un principe, dû à A. Church, et qui n'a jamais été contredit, prédit qu'en fait toute fonction effectivement calculable est une fonction récursive générale. Une des difficultés avec cet énoncé de principe est de définir ce que l'on entend par effectivement calculable. Une possibilité de définition est offerte par les machines de Turing.

Démonstration. Donnons seulement les grandes lignes de la démonstration, qui se fait à grands coups de récurrence. Pour montrer l'unicité de f, soit aussi $g : \mathbf{N} \to E$ une fonction vérifiant les propriétés (1) et (2). On considère l'ensemble

$$M = \{n \in \mathbf{N} \ : \ f(n) = g(n)\} \ .$$

Par récurrence on voit que $M = \mathbf{N}$, ce qui signifie bien que f égale g.

Pour montrer l'existence de f, on doit exhiber un sous-ensemble de $\mathbf{N} \times E$ ayant certaines propriétés. Soit

$$S = \{R \in \mathcal{P}(\mathbf{N} \times E) : (0, a) \in R \ \wedge \ ((n, x) \in R \Rightarrow (s(n), \varphi(x)) \in R)\} \ .$$

Cet ensemble n'est pas vide car il contient $R = \mathbf{N} \times E$. On considère alors

$$f := \bigcap S = \bigcap_{R \in S} R \ .$$

On voit que $(0, a)$ appartient à f et que f appartient à S. En quelque sorte f est donc le "plus petit" élément de S (par inclusion). Ce qu'il reste à démontrer est donc que : (a) le domaine de (la relation) f est $\mathbf{N}$ et (b) f est une fonction.

Pour (a), soit $M' = \{n \in \mathbf{N} \ : \ \exists x \in E \ (n, x) \in f\}$, alors $0 \in M$ et par récurrence (en utilisant $f \in S$) on obtient $M' = \mathbf{N}$. Pour (b) on considère l'ensemble M'' formé des entiers $n \in \mathbf{N}$ tels qu'il existe au plus un $x \in E$ tel que $(n, x) \in f$. Par l'absurde on montre que $0 \in M''$: sinon il existerait x différent de a tel que $(0, x) \in f$, alors l'ensemble $R_0 := f \setminus \{(0, x)\}$ serait strictement contenu dans f, mais on vérifie que R_0 est élément de S, ce qui est une contradiction avec le fait que f est le "plus petit" élément de S. Pour voir que $n \in M''$ implique $s(n) \in M''$, soit x_0 l'unique élément de E tel que $(n, x_0) \in f$. Si $s(n)$ n'appartenait pas à M'' on aurait $(s(n), \varphi(x_0))$ et $(s(n), y)$ dans f, pour un élément y de E différent de $\varphi(x_0)$. Alors l'ensemble $R_1 := f \setminus \{(s(n), y)\}$ serait lui aussi strictement contenu dans f et élément de S, ce qui est une contradiction.

Addition sur $\mathbf{N}$.

Soit $m \in \mathbf{N}$; notons σ_m l'unique application donnée par le théorème

$$\sigma_m : \mathbf{N} \to \mathbf{N}$$

telle que (a) $\sigma_m(0) = m$ et (b) $\sigma_m(s(n)) = s(\sigma_m(n))$. On définit la *somme* des entiers m et n par

$$m + n := \sigma_m(n) \ .$$

Multiplication sur $\mathbf{N}$.

Pour définir la multiplication il est commode d'utiliser la remarque (ii) après le théorème. Soit $m \in \mathbf{N}^*$; notons μ_m l'unique application

$$\mu_m : \mathbf{N}^* \to \mathbf{N}$$

telle que (a) $\mu_m(1) = m$ et (b) $\mu_m(s(n)) = \sigma_m(\mu_m(n))$. (C'est pour formuler le (a) de manière naturelle que l'on utilise la remarque (ii).) On définit le *produit* des entiers m et n par :

$$mn := \mu_m(n) \quad \text{et} \quad mn = 0 \ \text{ si } m = 0 \text{ ou } n = 0.$$

Pour montrer les propriétés usuelles des opérations que nous venons de définir, on procède encore par récurrence. Ainsi pour vérifier l'associativité de la somme, à savoir l'égalité

$$(m+n)+k = m+(n+k)$$

on fait une récurrence sur k. Pour la commutativité

$$m+n = n+m$$

on montre d'abord que $1+n = s(n)$, que $s(m)+n = s((m+n))$ et on termine par récurrence. Pour montrer les lois distributives

$$m(n+p) = mn+mp \quad \text{et} \quad (n+p)m = nm+pm$$

on fait une récurrence sur p (resp. m).

Ordre sur $\mathbf{N}$.

Une fois que l'on dispose de la somme sur $\mathbf{N}$ on peut aussi voir que l'*ordre* sur $\mathbf{N}$ défini par $\in$ est donné par

$$m \leq n \quad \Leftrightarrow \quad \exists p \in \mathbf{N} \ (m+p=n) \ .$$

Factorielle.

La dernière fonction sur $\mathbf{N}$, que nous allons définir ici de manière récursive est la factorielle. (A nouveau nous travaillons avec la version du théorème avec 1 au lieu de 0.) Soit $E = \mathbf{N} \times \mathbf{N}$, $a = (1,2)$ et

$$\begin{array}{rccc} \varphi & : E & \to & E \\ & (m,n) & \mapsto & (mn, n+1) \end{array}$$

Si f dénote la fonction $f : \mathbf{N} \to E$, définie par le théorème, alors la *factorielle* est définie par

$$n! := (\pi_1 \circ f)(n) \quad \text{et} \quad 0! = 1 \ .$$

Ici π_1 est la "projection sur la première composante" qui à un couple (n,m) de E fait correspondre n. En clair : $n! = n(n-1)\cdots 2$.

Exemple : pour le coefficient binomial on a

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \ .$$

4.7 Dénombrement.

On a envie de dire que s'il existe une injection $f : E \to F$ alors E est "de taille inférieure" à F : ceci donnerait une relation d'ordre (de grandeur) sur les ensembles. Or on peut montrer que si on a une injection $f : E \to F$ et aussi une injection $g : F \to E$, alors il existe une *bijection* $h : E \to F$ (Théorème de Bernstein). On dit que des ensembles E et F ont *même cardinal* (ou sont *équipotents*) s'il existe une bijection entre E et F (noté : card(E) = card(F)). Un ensemble est dit *fini* s'il a le même cardinal qu'un entier naturel. Il est dit *infini* s'il n'est pas fini. On démontre qu'un ensemble est infini si et seulement si

il possède un sous-ensemble propre de même cardinal. Un ensemble héréditaire est infini. Un ensemble est (infini) *dénombrable* s'il a le même cardinal que $\mathbf{N}$.

Exemples. 1) Le produit cartésien de deux ensembles dénombrables est dénombrable.
2) L'ensemble $\mathbf{Q}$ des nombres rationnels est dénombrable.
3) $\mathbf{R}$ n'est pas dénombrable (argument de la diagonale de Cantor).
4) Il y a une bijection entre $\mathbf{R}$ et $\mathbf{R} \times \mathbf{R}$, et la courbe de Peano donne une surjection continue de l'intervalle $I = [0, 1]$ dans $I \times I$.
5) On s'attend à ce que tout sous-ensemble infini de l'ensemble $\mathbf{R}$ des nombre réels est soit dénombrable soit de cardinal égal au cardinal de $\mathbf{R}$. C'est l'*Hypothèse du continu*. On peut montrer que cette hypothèse est indépendante de la théorie ZFC (P. Cohen).

Deuxième partie

Nombres et limites.

Le nombre réel est une abstraction
qui, du segment, sépare et réunit ce qui est
indépendant des constructions géométriques.
B. Levi, "En lisant Euclide", Agone, 2003, p. 178

Si l'on n'utilise que des nombres rationnels et les quatre opérations de base $+$, $-$, $\times$, $/$, et un nombre fini d'opérations, on obtient une structure mathématique très stable, puisque chacune de ces opérations préserve l'ensemble des nombres rationnels.

Cependant, comme nous l'avons vu, d'abord par l'intermédiaire de la géométrie, dès l'époque des Grecs au moins, puis par le développement de la mécanique ou d'autres considérations physiques, il s'est avéré que les nombres rationnels ne suffisent pas à décrire le monde "réel", ou celui de la géométrie euclidienne. L'introduction des nombres réels qui permet de combler ce manque est allée de pair avec l'apparition, d'abord mystérieuse, d'une nouvelle opération fondamentale, le "passage à la limite".

D'abord vue comme une manière d'approcher des nombres irrationnels donnés (provenant de la géométrie par exemple) par des rationnels, cette opération se révèle indispensable et spectaculairement efficace pour "construire" de nouveaux nombres permettant de résoudre une incroyable quantité de problèmes : constructions de tangentes, calculs d'aires ou de volume, etc, pour ne parler que des plus évidentes.

Cette partie du cours va présenter les nombres réels de manière concrète, mais néanmoins précise et rigoureuse, par l'intermédiaire des développements décimaux illimités. Les propriétés fondamentales apparaîtront alors clairement, ainsi que l'intérêt d'en avoir d'autres caractérisations plus souples. En parallèle, la notion de limite de suites, puis de fonctions, sera introduite et développée.

Les nombres complexes seront introduits et leurs propriétés élémentaires seront détaillées.

Nous aurons alors suffisamment de matériel pour *définir* les fonctions usuelles, les notions d'aire et d'intégrale, et montrer leurs propriétés élémentaires.

Chapitre 5

Les nombres réels.

5.1 La droite géométrique.

Pour définir les nombres réels nous allons donner une "version numérique de la droite". C'est-à-dire que nous allons construire un ensemble à partir de l'ensemble des nombres rationnels dont les éléments ont les propriétés attendues des points sur une droite.

On se représente la droite (géométrique), comme une *longue règle*, permettant de faire des mesures, et sur laquelle en particulier les points sont ordonnés. On se dit qu'il n'y a pas de "trous" entre les points : la droite est *continue*. Si on enlève un point à la droite on la sépare/coupe en deux (elle est de dimension 1). Les points sur la droite ne sont pas infinitésimaux : avec les multiples d'une partie aussi petite que l'on veut, on peut recouvrir toute la droite.

5.2 Notions de calcul segmentaire

Nous avons déjà dit que les Grecs avaient développé un calcul sur les segments, qui leur permettait de résoudre de nombreux problèmes. Il faudra que notre théorie des nombres réels soit aussi performante que de calcul.

Nous suivons, D. Hilbert, "Les fondements de la géométrie", Ed. J. Gabay, Paris, 1997 ; réimpression de la traduction française de 1971. (La 1ère édition de l'original allemand date de 1899.)

Ce calcul géométrique peut être basé sur le théorème de Pascal (proche parent du théorème de Thalès, qui se déduit du théorème de Pascal en utilisant le calcul segmentaire, voir *loc. cit.* Théorème 41, page 86).

Théorème (de Pascal) *Soient A, B, C, et A', B', C' deux groupes de trois points appartenant respectivement à deux droites concourantes et tous différents de l'intersection de ces droites ; si CB' est parallèle à BC' et CA' est parallèle à AC', alors BA' est parallèle à AB' (voir figure 1).*

Nous allons maintenant voir comment, en utilisant ce théorème, on peut *définir* des opérations sur les segments et aussi comment on peut démontrer les *propriétés* fondamentales de ces opérations.

Il est clair comment *sommer deux segments* : on les juxtapose. Plus précisément, si trois points A, B, C sont alignés et B est entre A et C on dit que le segment $c = AC$ est la *somme* des deux segments $a = AB$ et $b = BC$. On écrit :

$$c = a + b \ .$$

On dit que les segments a et b sont *plus petits que* le segment c : on écrit

$$a < c \quad \text{et} \quad b < c \ .$$

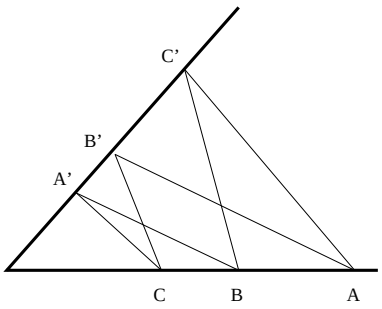


Figure 1

Les axiomes de la géométrie élémentaire montrent que *la somme des segments est associative et commutative* :

$$a + (b + c) = (a + b) + c \quad \text{et} \quad a + b = b + a \ .$$

Le *produit* ab de a et b est obtenu par la construction de la figure 2. En utilisant la construction de la figure 3, on montre que *le produit est commutatif* :

$$ab = ba \ .$$

En effet le théorème de Pascal montre que les extrémités de ab et de ba sont confondues (les droites pointillées sont parallèles). En utilisant la construction de la figure 4 on voit que *le produit est associatif* :

$$a(bc) = (ab)c \ .$$

A nouveau le théorème de Pascal est utilisé pour identifier l'extrémité de segments. La *distributivité*

$$a(b + c) = ab + ac$$

peut se voir sur la figure 5. Après avoir construit les segments ab, ac et $a(b + c)$, on mène une droite verticale par c. Les triangles hachurés sont congruents et le théorème de congruence des côtés opposés d'un parallélogramme permet de conclure.

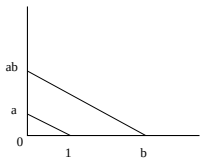


Figure 2

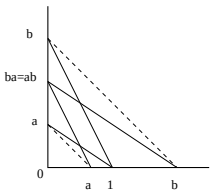


Figure 3

Voici les étapes pour la construction de la figure 3 : on part de la figure 2, on porte a et b sur l'autre côté, on relie 1 et b par un segment et on tire la parallèle à ce segment par a (sur le premier côté).
Pour la figure 4 : porter 1 et b, porter a et c, construire $d = ab$ et $e = cb$, porter d et e sur le premier côté, construire ae et cd, appliquer le théorème de Pascal.

Exercice. Montrer comment définir l'inverse $1/a$ et la racine carrée d'un segment a. (Indications : considérer la figure analogue à la figure 2, avec a à la place de b sur l'axe horizontal et 1 à la place de ab, alors $1/a$ est représenté par le point où se trouve a dans la figure 2 ; pour la racine carrée construire le cercle de diamètre a de centre $a/2$, puis considérer l'intersection P de ce cercle avec la droite verticale par 1 : le côté OP du triangle rectangle OPa représente la racine carrée de a.)

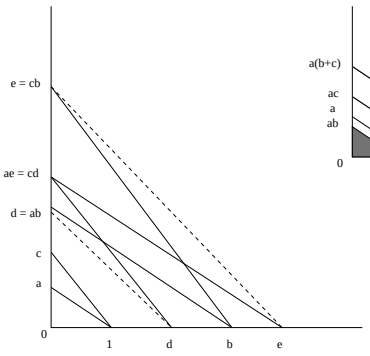


Figure 4

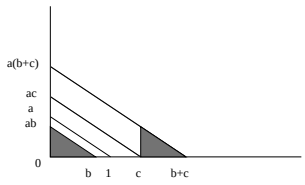


Figure 5

Avec ces considérations on voit que l'on peut représenter sur la droite toutes les longueurs rationnelles, c'est-à-dire commensurables avec l'unité, ainsi que les solutions (géométriques) des équations $x^2 = a$, avec a positif. A travers les siècles les mathématiciens avaient aussi développé des méthodes pour résoudre géométriquement d'autres équations (cubiques, quartiques, ...).

5.3 La droite numérique–développements décimaux illimités.

Le point de départ pour notre construction des réels est l'observation que *tout point sur la droite peut être approché avec un degré de précision quelconque par des points rationnels.* Ceci découle de considérations analogues à celle utilisées pour établir l'algorithme d'Euclide. Si on se donne A un point sur la droite et une unité 1, on trouve d'abord d_0 entier tel que $A = d_0 1 + A_1$ avec A_1 inférieur à 1. Puis on cherche d_1 entier tel que $10A_1 = d_1 1 + A_2$, avec $A_2 < 1$ et ainsi de suite on cherche d_k entier tel que

$$10A_k = d_k 1 + A_{k+1}, \quad \text{avec } A_{k+1} < 1 \ .$$

Alors A est approché par défaut par le point

$$d_0 1 + d_1 \frac{1}{10} + d_2 \frac{1}{10^2} + \cdots + d_k \frac{1}{10^k} \ .$$

Il se peut qu'avec cette construction on tombe pile sur A ; alors A est rationnel.

Remarque. Nous avons utilisé de manière essentielle la propriété qu'à partir des multiples d'un segment aussi petit que l'on veut on peut recouvrir toute la droite.

Definition. Un *développement décimal illimité (DDI)* est une suite de chiffres

$$\pm m, d_1 d_2 \ldots d_n \ldots$$

où $\pm$ est un signe, $m \geq 1$ est un entier positif et pour tout n,

$$d_n \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

est un chiffre décimal arbitraire.

L'*ensemble des nombres réels* $\mathbf{R}$ est l'ensemble des DDI. [1]

[1] Cette définition n'est pas vraiment correcte tant que nous n'avons pas identifié certains DDI entre eux : comme $0, \overline{9}$ et 1. Pour l'instant nous avons quelques DDI de trop. Voir plus bas.

La notion de suite d'entiers a un sens indépendamment de toute considération géométrique. Le fait que nous écrivons la suite comme "un nombre à virgule" est une question de commodité, pour nous rappeler du sens que l'on veut donner à la suite : nous aurions très bien pu écrire la suite comme $(\pm, m, d_0, d_1, \ldots)$. Au DDI $\pm m, d_1 d_2 \ldots d_n \ldots$ on associe une suite de nombres rationnels en reprenant la formule ci-dessus : c'est la suite des $\pm(m + d_1 \frac{1}{10} + d_2 \frac{1}{10^2} + \cdots + d_k \frac{1}{10^k})$ pour k croissant.

Terminologie. Nous allons appeler m la *partie entière* du développement décimal. Ceci ne devrait pas prêter à confusion avec un autre usage de ces mots : on appelle aussi partie entière de x le plus grand entier inférieur ou égal à x. Les deux notions coïncident si x est positif, mais la partie entière de $-1/2$ dans cet autre sens, par exemple, n'est pas 0, mais bien -1.

En résumé, nous avons donc adopté la manière usuelle de représenter les nombres, sur une calculatrice par exemple, comme définition des réels, sauf qu'une calculatrice ne pourra jamais représenter de manière complètement exacte un réel avec un DDI non fini. Ce qui nous reste à faire est de montrer que l'ensemble des DDI est un bon candidat pour une version numérique de la droite.

Regardons les DDI de plus près. Par exemple,

$$x = 0,2069261.$$

Un tel nombre peut se représenter aussi comme un rationnel en multipliant par 10^d, où d est le nombre de chiffres après la virgule. Dans l'exemple ci-dessus, $10^d = 10^7$ et on trouve que

$$10000000x = 2069261 \text{ donc } x = \frac{2069261}{10000000}.$$

Exercice. Cet exercice donne une présentation équivalente mais plus simple des nombres décimaux avec un nombre fini de chiffres après la virgule.

Soit $n \geq 0$, $m \geq 1$ des entiers. Pour chaque entier i tel que $-m \leq i \leq n$, soit d_i un chiffre $d_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, vu comme un entier. On définit le rationnel

$$x = 10^n d_n + 10^{n-1} d_{n-1} + \cdots + 10 d_1 + d_0 + \frac{1}{10} d_{-1} + \cdots + 10^{-m} d_{-m}.$$

Montrer que le développement décimal de ce nombre est

$$x = k, d_{-1} \ldots d_{-m}$$

où la partie entière est

$$k = 10^n d_n + 10^{n-1} d_{n-1} + \cdots + 10 d_1 + d_0.$$

On écrit plus rapidement

$$x = \sum_{i=-m}^{i=n} 10^i d_i.$$

Remarque. En s'inspirant de cet exercice, on est tenté d'écrire

$$x = \pm m + \sum_{i=-\infty}^{i=-1} 10^i d_i. \tag{5.3.0}$$

Une telle écriture est suggestive et tentante car elle fournit une expression compacte et qui permet des manipulations intuitives en utilisant les propriétés habituelles de sommes d'entiers. Lorsque l'on aura

défini ce que peut être une telle somme infinie (que l'on appelle une série), on verra que (5.3.0) est effectivement rigoureusement vraie avec cette interprétation.

Exercice. Montrer que placer le signe "dans la partie entière n" ne fournirait pas une définition satisfaisante. (Comment pourrait-on alors définir le développement décimal de $-1/2$?)

On sait qu'en effectuant la division de a par b "à la main", *tout nombre rationnel a/b possède un développement décimal périodique*, c'est-à-dire qu'après un certain nombres de chiffres, ceux-ci se répètent (le signe utilisé est le signe habituel de a/b). En utilisant la notation

$$n, d_1 \dots d_k \overline{d_{k+1} \dots d_n}$$

pour un développement décimal où la partie surlignée se répète ensuite périodiquement, on a par exemple

$$\begin{aligned} \frac{1}{3} &= 0,\overline{3} \\ -\frac{1}{9} &= -0,\overline{1} \\ \frac{55}{62} &= 0,88\overline{70967741935483 8} \end{aligned}$$

Il est possible que la partie périodique soit de longueur 1 et que le chiffre correspondant soit 0 : cela se produit si et seulement si le nombre rationnel a/b a la propriété que son dénominateur réduit (c'est-à-dire que l'on suppose que a et b sont premiers entre eux) n'est divisible que par les nombres premiers 2 et 5. Par exemple

$$\begin{aligned} \frac{1}{4} &= 0,25\overline{0} \\ -\frac{13}{10} &= -1,3\overline{0} \end{aligned}$$

Exercice. Justifier que, comme dit ci-dessus, un développement décimal n'ayant qu'un nombre fini de chiffres après la virgule correspond à un nombre rationnel qui peut s'écrire

$$x = \frac{a}{10^d}$$

pour un certain $d \geq 1$ (cette fraction n'étant pas forcément irréductible), et justifier que tout nombre rationnel du type

$$x = \frac{a}{2^a 5^b}$$

peut se ramener à cette forme.

On a donc vu que l'ensemble $\mathbf{Q}$ des rationnels peut "s'identifier" à un sous-ensemble de $\mathbf{R}$, formé de développements qui sont périodiques à partir d'un certain rang. Il n'est pas forcément évident que *tout développement périodique correspond à un nombre rationnel*, mais c'est le cas.

Par exemple, considérons le développement périodique

$$-5,12\overline{329}.$$

Si l'on veut essayer de l'identifier à un nombre rationnel, on fait les calculs suivants : si x désigne le nombre réel ainsi défini, on a

$$\begin{aligned}
-100x = &\quad 512,\overline{329}\\
-100x - 512 = &\quad 0,\overline{329}\\
-100x - 512 = &\quad 0,329\overline{329}\\
1000(-100x - 512) = &\quad 329,\overline{329}\\
1000(-100x - 512) - 329 = &\quad 0,\overline{329}\\
1000(-100x - 512) - 329 = &\quad -100x - 512\\
999(-100x - 512) = &\quad 329\\
-100x = &\quad 512 + \tfrac{329}{999}\\
x = &\quad -\tfrac{511817}{99900}.
\end{aligned}$$

Exercice. (1) En adaptant ce calcul, vérifier que tout développement illimité périodique correspond à un nombre rationnel. Peut-on prévoir un dénominateur possible à partir de la donnée de la période ?

(2) Écrire un développement périodique "au hasard", exprimer le sous la forme d'une fraction, et vérifier que le résultat est correct.

(3) Soit a/b est un nombre rationnel, avec b positif. On définit une suite d'entiers y_n et une suite de rationnels x_n comme suit : y_n est tel que $y_n b < a10^n < b(y_n + 1)$; on pose $x_n = y_n/10^n$. Finalement on pose $a_n = 10^n(x_n - x_{n-1})$. Vérifier que les a_n donnent un DDI pour a/b.

On peut se demander si le calcul précédent est bien une démonstration correcte du fait que $x = -\frac{511817}{99900}$, ou s'il est seulement suggestif. En fait, dans l'argument ci-dessus on a utilisé par exemple, que l'on puisse ajouter, multiplier, comparer, etc... des développements décimaux illimités. Ces opérations n'ont pas encore été définies (voir ci-dessous). Cependant, puisque l'on sait, par le procédé de division habituel, calculer le développement décimal périodique d'un nombre rationnel, une fois que l'on a "deviné" le résultat ($x = -\frac{511817}{99900}$), on peut calculer le développement décimal de cette fraction et vérifier qu'il s'agit bien de $5,12329$. En ce sens, savoir si le calcul est justifié n'est pas vraiment important dans ce cas particulier (sauf qu'une telle division n'est pas amusante à effectuer !)

Il est cependant évident qu'il n'est pas possible d'assimiler les développements décimaux à des "nombres" s'il n'est pas possible d'étendre les opérations élémentaires $+$, $-$, $*$, $/$. Par ailleurs, comme nous l'avons déjà observé, il faut également pouvoir comparer deux nombres réels, c'est-à-dire, étant donnés x et y, dire si $x < y$, $x = y$ ou $x > y$. De plus, ces opérations ne doivent pas être arbitraires, mais elles doivent se ramener aux opérations usuelles sur les nombres rationnels lorsque les nombres réels concernés sont dans $\mathbf{Q}$. Nous allons voir que mettre en œuvre ce programme n'est pas si évident.

Commençons par la comparaison. Il paraît évident de dire que $x = y$ si et seulement si les signes, les parties entières, et toutes les décimales de x et y coïncident. Il faut évidemment prendre garde à ne pas distinguer $+0,\overline{0}$ de $-0,\overline{0}$ (représentant tout les deux le nombre rationnel zéro). Il y a d'autres pièges plus cachés, illustrés par le calcul suivant :

$$\begin{aligned}
\frac{4}{9} = &\quad 0,4444\ldots = 0,\overline{4}\\
\frac{5}{9} = &\quad 0,5555\ldots = 0,\overline{5}\\
1 = 1,\overline{0} = \frac{4+5}{9} = &\quad 0,9999\ldots = 0,\overline{9}.
\end{aligned}$$

Autrement dit, on trouve ainsi deux expressions décimales naturelles pour le nombre (rationnel) 1.

Exercice. En utilisant la méthode de calcul de la fraction associée à un développement décimal illimité périodique, vérifier que $0,\overline{9} = \frac{1}{1}$.

De la même manière, on voit que les deux développements décimaux illimités

$$\begin{aligned} x_1 &= \quad \pm m, d_1 \dots d_n \overline{0} \quad \text{avec } d_n \neq 0, \\ x_2 &= \pm m, d_1 \dots d'_n \overline{9} \quad \text{avec } d'_n = d_n - 1, \end{aligned}$$

correspondent à des rationnels égaux (le signe étant identique dans les deux expressions).

Exercice. En choisissant le signe, l'entier $m \geq 0$, et $d_1, \dots, d_n$ au hasard, vérifier cette assertion sur des exemples.

On remarque qu'une telle construction n'est possible que pour des nombres rationnels. L'*égalité* de deux nombres réels x et y peut alors se définir ainsi : soit x et y sont rationnels, et égaux au sens usuel (si $x = a/b$ et $y = c/d$, cela signifie $ad = bc$), soit x et y ne sont pas rationnels, et le signe, la partie entière, et chaque décimale coïncident.

Passons à l'*inégalité* $x < y$. La manière qui s'impose en faisant des exemples est la suivante : on compare d'abord suivant les signes par la règle usuelle, puis si $x \geq 0$ et $y \geq 0$ (par exemple), on a $x < y$ si et seulement si : soit la partie entière de x est strictement inférieure à celle de y, soit (si les parties entières sont égales) il existe k chiffres $d_1, \dots, d_k$ qui coïncident et le $(k+1)$-ème chiffre de x est strictement inférieur à celui de y.

Exercice. Montrer qu'il faut aménager cette définition pour le cas des rationnels pour la même raison que pour l'égalité.

Exercice. Soit

$$x = m, d_1 \dots d_n \dots$$

un nombre réel. Montrer que $m \leq x \leq m + 1$.

Enfin on en vient aux opérations élémentaires ou algébriques. Lorsque les deux nombres x et y concernés ont un développement décimal limité, la méthode d'addition, soustraction, multiplication, ou division est la méthode usuelle. Mais pour des développements illimités, si l'on essaie de "poser l'opération", il y a des problèmes. Pour l'addition, on a l'habitude de commencer à additionner les chiffres le plus à droite, mais il n'y en a plus ! Pour la multiplication, il faudrait multiplier x par chaque décimale de y, puis faire une infinité d'additions !

Vu que les début des DDI sont à interpréter comme des approximations rationnelles par défaut, on pourrait penser que, par exemple, le produit de deux DDI s'obtient comme le DDI dont les débuts sont les approximations rationnelles données par les produits des approximations des débuts des DDI des termes du produit.

Pour traiter un cas concret, admettons que les racines carrées de 2 et de 3 sont données par les (débuts de) DDI suivants

$$\sqrt{2} = 1,414\dots \quad \text{et} \quad \sqrt{3} = 1,732\dots \ .$$

On s'attend à ce que $\sqrt{2}\sqrt{3} = \sqrt{6}$, mais, si nous procédons comme envisagé, voici ce qui arrive. Nous avons donc les DDI $(m, a_1, a_2, a_3, \dots) = (1, 4, 1, 4, \dots)$ et $(m', b_1, b_2, b_3, \dots) = (1, 7, 3, 2, \dots)$. Posons

$$A_k = m + a_1/10 + \dots + a_k/10^k \ \text{ et } \ B_k = m' + b_1/10 + \dots + b_k/10^k \ .$$

Alors

$$\begin{array}{rclrclrclrcl} A_0 & = & 1 & A_1 & = & 1,4 & A_2 & = & 1,41 & A_3 & = & 1,414\ , \\ B_0 & = & 1 & B_1 & = & 1,7 & B_2 & = & 1,73 & B_3 & = & 1,732\ , \end{array}$$

et donc

$$\begin{array}{rclrclrclrcl} A_0B_0 & = & 1 & A_1B_1 & = & 2,38 & A_2B_2 & = & 2,4393 & A_3B_3 & = & 2,449048\ . \end{array}$$

Or ceux-ci ne sont *pas* les débuts du DDI de $\sqrt{6} = 2,449\ldots$. On obtient néanmoins une suite d'approximations rationnelles, qui semble s'approcher de ce DDI.

Remarque. On peut comparer cette situation avec celle, où on essaierait de définir le produit de rationnels *sous forme réduite*, disons

$$\frac{2}{3} \times \frac{1}{4} = \frac{2 \times 1}{3 \times 4} = \frac{1}{6} \ .$$

La règle simple ne donne pas une fraction réduite (l'analogue du DDI), mais une autre fraction, que l'on peut réduire. Une façon de définir les opérations sur les DDI, va consister à les considérer comme (définissant) des suites de rationnels et de travailler avec des suites plus générales.

Exercice. (1) Soit

$$x = \pm 0, d_1 \ldots d_n \ldots$$

un nombre réel. Vérifier que

$$\begin{aligned} 10x = &\quad \pm d_1, d_2 \ldots d_n \ldots, \\ \frac{1}{10}x = &\quad \pm 0, 0 d_1 d_2 \ldots d_n \ldots \end{aligned}$$

c'est-à-dire que la multiplication ou la division par 10 correspondent à décaler les chiffres d'une unité.

(2) Déduire de (1) la propriété suivante : si $x \geq 0$ et si $y > 0$, alors il existe un entier m tel que $my > x$. (Cela s'appelle la propriété d'Archimède, voir plus bas).

(3) Montrer comment définir plus simplement $x \pm y$, $x \times y$ si x est un nombre réel et y a un développement décimal limité.

Exercice. Soit

$$x = 0, d_1 \ldots d_n \ldots$$

un nombre réel. Soit $n \geq 1$ quelconque et x_n le nombre rationnel ayant le développement décimal limité

$$x_n = 0, d_1 \ldots d_n \overline{0}.$$

(1) Montrer que

$$x_n \leq x \leq x_n + 10^{-n}.$$

(2) Soit

$$y = 0, e_1 \ldots e_n \ldots$$

un autre développement décimal illimité. Montrer que si

$$y \leq x \leq y + 10^{-n},$$

alors les n premiers chiffres de x et y sont égaux.

Exercice. Vérifier que si $x \leq y$ et $z \leq t$ on a $x + z \leq y + t$ et si $0 \leq x \leq y$ et $0 \leq z \leq t$, on a $xz \leq yt$.

Revenons aux approximations par défaut de manière plus générale. [2] L'exemple ci-dessus nous montre que pour définir les opérations sur les DDI on aimerait avoir la propriété suivante, dite *des suites croissantes*

[2] Nous allons utiliser un peu de vocabulaire : si E est un ensemble muni d'une relation d'ordre $\leq$ et si F est un sous-ensemble de E, on dit que l'élément m de E est un *majorant* de F, si pour tout élément f de F on a $f \leq m$. Si m est un élément de F, on dit que c'est un *élément maximal* de F. Le plus petit des majorants de l'ensemble F s'appelle le *supremum* de F. On définit de même les *minorants*, *éléments minimaux* et *infima*. Ici l'ensemble F est celui des valeurs d'une suite.

(SC) L'ensemble des majorants d'une suite croissante soit est vide, soit possède une plus petit élément.

On appelle ce plus petit élément le *supremum* de la suite, et on le note sup. La propriété des suites croissantes permet effectivement de définir les opérations : si $A = (a_0, a_1, \dots)$ et $B = (b_0, b_1, \dots)$ avec $A \geq 0$ et $B \geq 0$, alors on pose

$$\begin{aligned} A + B &:= \sup(A_n + B_n) \\ A \cdot B &:= \sup(A_n B_n) \end{aligned}$$

Observons que les suites $(A_n + B_n)$ et $(A_n B_n)$ sont bien croissantes et majorées, et que donc (SC) garanti(rait) l'existence de leurs sup (par exemple $(a_0 + 1)(b_0 + 1)$ majore la première suite).

Exercice. Vérifier que les règles de calcul suivantes restent valides entre nombres réels : $x + 0 = x$, $x \times 0 = 0$, $x \times 1 = x$, $x + y = y + x$, $x \times y = y \times x$, $x \times (y + z) = x \times y + x \times z$, ...

Pour l'instant on a introduit les nombres réels d'une manière concrète, qui montre que, certainement, il y en a "plus" que de nombres rationnels. Mais il n'est pas du tout clair qu'ils permettent de construire les "nombres géométriques" comme $\sqrt{2}$ et π.

Proposition. Il existe un nombre réel x – c'est-à-dire un développement décimal illimité – tel que $x^2 = 2$. L'ensemble des x vérifiant cette relation est réduit à $\{x, -x\}$. De plus on a

$$x = 1,414213562373095048801688724209698078569671875376948 0\dots$$

Démonstration. Il est assez facile de déterminer les décimales de $\sqrt{2}$ en utilisant les propriétés de stabilité des premiers chiffres d'un produit, et en procédant par "approximations successives" par des développements décimaux limités. Pour commencer, comme $1^2 < 2$ et $2^2 > 2$, la partie entière de $\sqrt{2}$ doit être égale à 1. Si on regarde les nombres $1, d_1\overline{0}$, où d_1 est un chiffre décimal, on constate que

$$1,4^2 = 1,96 < 2, \text{ mais } 1,5^2 = 2,25 > 2,$$

donc le premier chiffre décimal doit être 4. Puis si l'on regarde $1, 4d_2\overline{0}$, on a

$$1,41^2 = 1,9881 < 2, \text{ mais } 1,42^2 = 2,0164 > 2.$$

Il est naturel alors de procéder par récurrence pour montrer que le $(n+1)$-ème chiffre décimal peut être déterminé de manière unique si les n premiers chiffres sont connus. Autrement dit, on suppose que l'on a un développement décimal limité $x_n = 1,41d_3 \dots d_n$ tel que

$$x_n^2 < 2, \text{ mais } (x_n + \overbrace{0,000\dots0}^{n-1 \text{ zéros}}1)^2 > 2.$$

Noter que l'on écrit la seconde inégalité de cette manière plutôt que de la façon suivante

$$(1,41d_3 \dots d'_n)^2 > 2, \text{ où } d'_n = d_n + 1,$$

qui peut sembler plus intuitive, en raison de la possibilité de "retenue" si le n-ème chiffre est égal à 9 (ce qui arrive pour le 14-ème chiffre de $\sqrt{2}$ par exemple).

On peut alors choisir le $(n+1)$-ème chiffre d_{n+1} de la manière suivante : si $1,41d_3 \dots d_n 9^2 > 2$, on prend pour d_{n+1} le plus grand chiffre $0 \leq d \leq 9$ tel que $1,41d_3 \dots d_n d^2 < 2$: il existe nécessairement

puisque cette inégalité est vraie pour $d = 0$ par hypothèse, et fausse pour $d = 9$. Si au contraire $1,41d_3 \dots d_n 9^2 < 2$, on pose $d_{n+1} = 9$.

Ainsi, par récurrence on a construit un certain développement décimal illimité et l'on peut vérifier que les premiers chiffres correspondent à ceux donnés dans l'énoncé de la Proposition. Ce qu'il reste à faire est de montrer que le nombre réel x qui correspond par définition à ce développement décimal illimité vérifie $x^2 = 2$. Mais on peut prédire les premiers chiffres de x^2 à l'aide des premiers chiffres de x, c'est à dire des d_n, et ceux-ci ont la propriété de coïncider avec ceux de l'entier $2,\overline{0}$ par construction.

Exercice. (1) Discuter de la rigueur de l'argument précédent. êtes-vous convaincus ? Quelles propriétés de l'ensemble des nombres réels a-t-on utilisées implicitement ?

(2) Soit maintenant x un nombre réel quelconque. Existe-t-il y tel que $y^2 = x$?

(3) Existe-t-il un nombre réel x tel que $x^3 = 2$? Tel que $x^4 = 2$?

(4) Soit a un entier positif. On définit des suites y_n, x_n et a_n en posant : y_n l'entier tel que $y_n^2 \leq 10^{2n}a < (y_n + 1)^2$, puis $x_n = y_n/10^n$ et $a_n = 10^n(x_n - x_{n-1})$. Montrer que (a_n) donne un DDI de la racine carrée de a.

Comment faire pour π ? Par définition essentiellement, 2π est le périmètre d'un cercle de rayon 1. Archimède procède de la manière suivante, déjà décrite géométriquement dans la Section 2.5 : il construit des suites de polygones inscrits et circonscrits dans un cercle de rayon 1. Le périmètre d'un polygone inscrit est "évidemment" plus petit que le périmètre du cercle, et de même son aire est plus petite que celle du cercle. Similairement, pour les polygones circonscrits, le périmètre ou l'aire est plus grande que celle du cercle. Avec les notations de 2.5, pour des polygones réguliers à $3 \cdot 2^n$ côtés, b_n étant le demi-périmètre de celui qui est inscrit et a_n celui du polygone circonscrit, on a

$$b_n < \pi < a_n, \quad b_0 = 3,\ a_0 = 2\sqrt{3},\ \text{et } a_{n+1} = \frac{2}{\frac{1}{a_n} + \frac{1}{b_n}}, \quad b_{n+1} = \sqrt{a_{n+1}b_n}. \tag{5.3.0}$$

Comme le suggèrent les dessins, on peut constater en faisant les calculs que les premières décimales de a_n et b_n ne changent plus après une certaine valeur de n. Par exemple, les deux premières décimales sont fixes à partir de $n = 4$:

$$\begin{aligned}
b_0 &= \ 3,000\overline{0}, & a_0 &= 3,4641\dots \\
b_1 &= \ 3,1058\dots, & a_1 &= 3,21531\dots \\
b_2 &= \ 3,1326\dots, & a_2 &= 3,1596\dots \\
b_3 &= \ 3,1393\dots, & a_3 &= 3,1460\dots \\
b_4 &= \ 3,1410\dots, & a_4 &= 3,1427\dots \\
b_5 &= \ 3,1314\dots, & a_5 &= 3,1418\dots
\end{aligned}$$

Il est intéressant de noter que les formules (5.3.0) montrent que, par cette méthode en tout cas, il est beaucoup plus naturel d'approcher π par des nombres réels (déjà connus, comme $\sqrt{x}$ pour $x > 0$) plutôt que par le développement décimal illimité lui-même. En particulier, la méthode d'Archimède nécessite de connaître déjà l'existence des racines carrées.

Le processus d'approximation utilisé pour arriver à π nous amène à considérer la propriété *de Cantor* :

(C) si x_n et y_n sont des suites avec $x_n \leq x_{n+1} \leq y_{n+1} \leq y_n$,
alors il existe x tel que pour tout n on a $x_n \leq x \leq y_n$.

Notons que nous n'affirmons pas l'existence d'un x unique, qui ne sera garantie que si la distance entre les y_n et les x_n décroît arbitrairement. On appelle souvent cette propriété, la propriété "du gendarme" : x est "coincé" entre les deux suites x_n et y_n.

Une autre propriété que nous avons déjà rencontrée est la propriété d'*Archimède* :

(A) pour tous x et y, avec $x > 0$, il existe un entier $n \geq 0$ tel que $y < nx$.

Proposition.

a) L'ensemble des nombres réels a la propriété (SC) des suites croissantes.
b) La propriété (SC) des suites croissantes implique la propriété (C) de Cantor.
c) La propriété (SC) des suites croissantes implique la propriété (A) d'Archimède.
d) La propriété (A) d'Archimède implique que l'ensemble $\mathbf{Q}$ des rationnels est dense dans l'ensemble $\mathbf{R}$ des réels, c'est-à-dire : si x et y sont des réels avec $x < y$, alors il existe p rationnel tel que $x < p < y$.

On voit donc que la propriété (SC) est tout à fait fondamentale. Nous verrons plus loin, qu'elle est un cas particulier d'une propriété, qui en quelque sorte caractérise l'ensemble des nombres réels.

Démonstration. a) Pour montrer que toute suite croissante et majorée de DDI admet un supremum, nous devons considérer une suite de suites. Soit $s_1 \leq s_2 \leq s_3 \leq \ldots \leq M$ une suite des DDI $s_i = s_{i0}, s_{i1}s_{i2}\ldots$ (on néglige les signes). Le supremum doit être un réel, c'est-à-dire un DDI et nous allons en définir les décimales une à une. On commence par poser la partie entière égale à la plus grande des parties entières s_{i0}. Le fait que la suite des parties entières admet un plus grand élément découle de l'hypothèse que la suite est majorée. Ensuite on considère le plus grand des s_{i1} *pour* $i \geq i_0$, etc. en veillant à toujours faire en sorte que les nouvelles décimales ajoutées ne fassent pas baisser la valeur. On vérifie que le nombre ainsi obtenu est bien le supremum.

b-c) Le fait que (SC) implique (C) n'est pas difficile à montrer. Pour voir que (SC) implique (A) on peut procéder par l'absurde. Considérons la suite croissante $mx < mx + x = (m+1)x$. Si (A) était fausse, alors cette suite aurait un majorant, et donc par (SC) on aurait un plus petit majorant : soit α. Vu que $x > 0$, on a $\alpha - x < \alpha$ donc $\alpha - x$ n'est pas un majorant de la suite, c'est-à-dire qu'il existe m tel que $\alpha - x < mx$, mais alors $\alpha < (m+1)x$, ce qui contredit le fait que α est un majorant de la suite.

d) Utilisons (A) pour montrer la densité des rationnels. Par (A), d'une part il existe un entier $n > 0$, tel que $n(y-x) > 1$. D'autre part, il existe des entiers $m_1 > 0$ et $m_2 > 0$, avec $m_1 > nx$ et $m_2 > -nx$. D'où $-m_2 < nx < m_1$ et donc il existe m avec $-m_2 \leq m \leq m_1$ et $m - 1 \leq nx < m$, ce qui donne $ny > 1 + nx \geq m > nx$ et il suffit de poser $p = m/n$.

5.4 La propriété du sup.

Une propriété qui généralise la propriété (SC) des suites croissantes est la propriété du sup (ou, de manière plus élégante, de la borne supérieure) :

(SUP) tout sous-ensemble non-vide de $\mathbf{R}$ possède un ensemble de majorants, qui soit est vide, soit possède un plus petit élément.

On peut montrer, que $\mathbf{R}$ possède la propriété du sup (nous allons le faire ci-après). Il faut remarquer que (SC) et (SUP) sont les propriétés essentielles qui font la différence entre $\mathbf{Q}$ et $\mathbf{R}$.

Le supremum d'un ensemble E non-vide et majoré peut être caractérisé comme suit : *si s majore E et est tel que pour tout $\epsilon > 0$, il existe x dans E avec $s - \epsilon < x \leq s$, alors $s = sup(E)$*. En termes imagés cela veut dire, que le sup "colle à E depuis la droite".

Exercice. Montrer cette affirmation.

Exemple. 1) La borne supérieure de $X_a =]-\infty, a[$, $a > 0$, est égale à a. En effet, a est clairement un majorant, et pour tout $b < a$, on voit que $b \in X_a$, donc b n'est pas un majorant de X_a.

2) La borne supérieure de $X = [a, b]$, $a < b$, est égale à b. Dans ce cas, b est dans X, et on dit que b est le maximum de X.

Exercice. Définir la borne inférieure d'un ensemble X, notée $\inf X$ quand elle existe, et en donner une caractérisation analogue à celle du supremum.

5.5 Sur la construction des rationnels.

Les constructions de l'ensemble des nombres réels présentées en cours partent toutes de la donnée de l'ensemble des nombres rationnels muni des opérations et de l'ordre usuels. Voyons rapidement comment on peut construire l'ensemble des rationnels à partir de celui des naturels, qui est une donnée des axiomes de la théorie des ensembles. Il s'agit d'une construction assez longue à faire en détail, mais les idées de base sont claires.

Quelques étapes : les entiers sont donnés comme un ensemble avec une application particulière, l'application successeur. De plus les propriétés des entiers fondent le principe de démonstration par récurrence. La somme et le produit d'entiers naturels est définie de façon récursive : l'idée est que la somme $m + n$ de deux entiers m et n signifie "ajouter n fois 1 à m", c'est-à-dire appliquer n fois la fonction successeur à m (ici m et n sont traités de façon asymétrique : m est fixé et n varie). Le problème à résoudre est de montrer que ceci définit une opération sur tous les entiers : comme on l'a vu dans la Partie I, on le montre par récurrence ! Ensuite on vérifie les propriétés usuelles de la somme (en particulier que $m+n = n+m$). Pour définir le produit on procède de même en utilisant l'idée, que faire le produit mn des entiers m et n revient à sommer n fois m avec lui-même.

Ensuite on définit les *entiers relatifs*, c'est-à-dire les entiers avec signe. Moralement ce sont les solutions x des équations $b + x = a$, pour a et b entiers naturels. Rappelons que l'on suppose à ce stade que le seul ensemble de nombres connu est celui des entiers naturels, on ne sait donc pas où chercher les solutions... Voici le truc : on observe d'abord que si, par exemple, x est solution de $5 + x = 3$ il est aussi solution de $8 + x = 6$ et d'une infinité d'autres telles équations. Le couple $(5, 3)$ peut définir -2, mais alors $(8, 6)$ aussi. Par conséquent on définit -2 comme étant l'ensemble de *tous* les couples $(5, 3)$, $(8, 6)$ et *équivalents*. Ici on dira que (b, a) et (b', a') sont équivalents si on a l'égalité d'entiers $b + a' = a + b'$. Un nombre entier relatif est donc un ensemble infini de couples équivalents d'entiers naturels ! On note $\mathbf{Z}$ l'ensemble des entiers relatifs. Les entiers naturels se retrouvent parmi les entiers relatifs comme l'ensemble des couples équivalents à ceux de la forme $(0, n)$, on a l'injection :

$$\begin{array}{ccc} \mathbf{N} & \to & \mathbf{Z} \\ n & \mapsto & [(0, n)] \end{array}$$

L'*opposé* de l'entier n (identifié à l'ensemble contenant $(0, n)$) est l'ensemble des couples équivalents à $(n, 0)$. On définit la *somme d'entiers relatifs* composante par composante : si x est l'ensemble des couples équivalents à (b, a) et y l'ensemble des couples équivalents à (b', a'), alors $x+y$ est par définition l'ensemble des couples équivalents à $(b + b', a + a')$ (ici on utilise la somme des entiers naturels). On définit de même le *produit d'entiers relatifs*, on vérifie que ces définitions ne dépendent pas du choix de

(b,a) et (b',a') et on montre les propriétés usuelles de ces opérations. On étend aussi la *relation d'ordre*. On peut aussi vérifier que tout entier relatif contient un couple de la forme $(0,n)$ ou de la forme $(n,0)$: il est soit positif, soit négatif, soit nul. D'où la notion de *signe* d'un entier relatif.

L'idée pour définir l'*ensemble des nombres rationnels* à partir de celui des entiers relatifs est semblable. On considère un rationnel x comme la solution d'une équation $bx = a$ avec a et b entiers relatifs. On définit donc $1/2$ comme l'ensemble des couples équivalents au couple $(2,1)$, où ici (b,a) et (b',a') sont équivalents si on a l'égalité d'entiers $ba' = b'a$. (On ne considère pas tous les couples : on évite les couples de la forme $(0,m)$. Noter que l'équation $0x = a$ n'a pas de solutions si $a \neq 0$.) Si on écrit $[(b,a)]$ pour l'ensemble des couples équivalents à (b,a) au sens qui vient d'être dit, alors l'ensemble des rationnels est l'ensemble $\mathbf{Q}$ des ensembles $[(b,a)]$ pour b entier relatif non-nul et a entier. On a l'injection :

$$\begin{array}{ccc} \mathbf{Z} & \to & \mathbf{Q} \\ m & \mapsto & [(1,m)] \end{array}$$

Les *opérations* sur les rationnels sont définies comme suit :

$$[(b,a)] + [(d,c)] = [(bd, ad+cb)] \quad \text{et} \quad [(b,a)][(d,c)] = [(bd,ac)] \ .$$

On peut aussi définir un *ordre* sur les rationnels. Pour a et b de même signe on a :

$$[(b,a)] \leq [(d,c)] \Leftrightarrow ad \leq bc \ .$$

Il s'agit là des règles usuelles, que l'on retrouve en écrivant a/b pour $[(b,a)]$. Ces opérations prolongent les opérations définies sur les entiers relatifs.

On voit donc que la construction des rationnels est essentiellement algébrique. Le réels ne sont pas définis pour résoudre des équations entre rationnels ! La propriété que les réels ont, et que les rationnels n'ont pas, est qu'ils "contiennent toutes les limites de suites de *rationnels*, qui devraient converger".

Exercice. Vous êtes évidemment familiers avec la définition de la somme de deux rationnels. Peut-être vous êtes-vous déjà demandés pourquoi on ne somme pas deux rationnels en utilisant la règle $a/b + c/d = (a+b)/(c+d)$. On vous propose de réfléchir à cette question en observant que cette "mauvaise" somme semble bien être celle qu'il faut utiliser pour, par exemple, résoudre le petit problème suivant. Lors d'une tournée de deux jours dans une ville, un représentant commercial visite onze clients, six le premier et cinq le deuxième. Il réussit à vendre son produit à un client sur six le premier jour, et à deux clients sur cinq le deuxième jour. Quel est son taux de réussite ? (Réponse : $(1+2)/(5+6)$?)

5.6 Il y a (beaucoup) plus de nombres réels que de rationnels.

Ce n'est qu'un rappel d'un énoncé déjà abordé dans la première partie, mais qui devrait maintenant prendre tout son sens. Ce que l'on veut montrer est que *l'ensemble des nombres réels n'est pas dénombrable.*

Comme on a vu on suppose, par l'absurde avoir numéroté les réels entre 0 et 1, que l'on représente par leur développement décimal illimité :

$$\begin{array}{ll} 1 & : \ 0, a_{11}a_{12}a_{13}... \\ 2 & : \ 0, a_{21}a_{22}a_{23}... \\ 3 & : \ 0, a_{31}a_{32}a_{33}... \\ \text{etc.} & \end{array}$$

Puis on considère le nombre b défini par le développement décimal

$$0, b_1 b_2 b_3 ...$$

où l'on définit b_i comme étant égal à 0 ou 1 suivant que a_{ii} est égal à 1 ou pas. Ceci définit bien un réel b : noter par exemple que b_i ne vaut pas identiquement 9 à partir d'un certain rang. Ce nombre ne peut pas être dans la liste, car sinon il existerait i avec $b_i = a_{ii}$; à b ne correspond donc aucun entier dans la numérotation, qui n'en est donc pas une.

5.7 Valeur absolue, intervalles.

La *valeur absolue* du réel x est

$$|x| = \begin{cases} x & \text{si } x \geq 0 \\ -x & \text{si } x < 0 \end{cases}$$

Elle mesure la distance de x à l'"origine" 0 et jouit des propriétés suivantes, qui seront démontrées dans un cas plus général plus tard.

1) *Inégalité du triangle* : $|x + y| \leq |x| + |y|$ ou plus généralement

$$|x_1 + \cdots + x_n| \leq |x_1| + \cdots + |x_n| \ .$$

2) $||x| - |y|| \leq |x \pm y|$
3) si $a \geq 0$, on a $|x| \leq a$ si et seulement si $-a \leq x \leq a$.

Pour a et b réels on pose

$$[a, b] = \{x \in \mathbf{R} : a \leq x \leq b\}$$

l'*intervalle fermé* entre a et b, et

$$(a, b) = \{x \in \mathbf{R} : a < x < b\}$$

l'*intervalle ouvert* entre a et b. On a aussi les versions semi-ouvertes ou semi-fermées $[a, b)$, $(a, b]$.

On montre que toute suite $I_n = [a_n, b_n]$ d'intervalles fermés emboîtés (donc $I_n \subset I_{n-1}$) et telle que la différence $b_n - a_n$ prend des valeurs aussi petites que l'on veut pour n grand, a une intersection non-vide, réduite à un point. Cette conséquence de la propriété de Cantor est connue comme la propriété des *intervalles emboîtés*.

5.8 Fractions continues et autres bases.

Fractions continues. Il existe d'autre méthodes pour définir ou paramétrer les nombres réels. L'une des plus fascinantes est celle des fractions continues. Dans cette formulation tout réel $x \geq 0$ s'écrit

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \cfrac{1}{\cdots}}}}$$

où les a_i sont des entiers positifs.

On peut montrer que chaque nombre réel peut s'écrire sous cette forme pour une suite (a_n) d'entiers, de manière unique si x n'est pas rationnel. Si $x \neq 0$ est rationnel il y a deux représentations, par exemple

$$\frac{12}{35} = 0 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{11}}} = 0 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{10 + \cfrac{1}{1}}}}.$$

Il est possible de montrer que, dans un certain sens, les nombres rationnels obtenus en "arrêtant la fraction continue" après n termes sont les meilleurs approximations possibles de x par des nombres rationnels. Par exemple, pour $x = \pi$, les cinq premiers termes sont

$$\pi = 3 + \frac{1}{7 + \frac{1}{15 + \frac{1}{1 + \frac{1}{292 + \cdots}}}},$$

et si l'on ne conserve que les quatre premiers coefficients (3, 7, 15, 1) on trouve le nombre rationnel

$$\begin{aligned} 3 + \frac{1}{7 + \frac{1}{15 + \frac{1}{1}}} &= \frac{355}{113} \\ &= 3,\overline{14159292035398230088495575221238938053097345132743362831858407079646017699115044247787610619469026548672566371681}. \end{aligned}$$

Exercice. A quoi la longueur de la partie répétée de ce développement décimal semble-t-elle reliée ?

On remarque que cette fraction continue à quatre termes (et le nombre rationnel $\frac{355}{113}$) approchent π à 10^{-6} près, et sont bien plus simples que la fraction évidente

$$\frac{3141592}{1000000} = \frac{392699}{125000}.$$

De même que les développements décimaux périodiques sont "spéciaux", les fractions continues périodiques sont caractérisées par une condition simple : elles correspondent aux réels (positifs) qui sont solution d'une équation polynomiale de degré 2

$$ax^2 + bx + c = 0$$

à coefficients entiers. Le cas le plus fameux est

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

Dans ce cas, on est tenté d'écrire que

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

donc $x^2 - x - 1 = 0$. Le discriminant de cette équation est $\Delta = 5$ et comme l'une des deux solutions est négative on "trouve"

$$x = \frac{1 + \sqrt{5}}{2} = 1,61803398874989484820458683\ldots$$

Ce nombre étant irrationnel, le développement décimal est illimité.

Exercice. Justifier intuitivement que

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

(Nous avons déjà rencontré ce développement dans la première partie.)

Autres bases. Le choix de la base 10 (autrement dit, le choix d'utiliser les chiffres usuels 0, 1, 2, 3, 4, 5, 6, 7, 7, 9) pour exprimer les nombres réels par des développements décimaux illimités n'est pas une nécessité. Certaines cultures et civilisations ont développé leurs mathématiques à l'aide d'autres bases, notamment les bases 12 et 60 (qui subsistent dans les calendriers), et bien entendu les ordinateurs utilisent des calculs en base 2 de manière interne (on parle de chiffres binaires, 0 ou 1, dans ce cas). Cela signifie simplement qu'au lieu d'écrire

$$x = \sum_{n=-k}^{n=\infty} 10^{-n} d_n$$

avec d_n un chiffre décimal, on pose

$$x = \sum_{n=-k}^{n=\infty} b^{-n} d_n$$

où (si la base b vérifie $2 \leq b \leq 10$), d_n appartient à l'ensemble $\{0, 1, \ldots, b-1\}$.

Il est important de remarquer que le choix de la base n'affecte pas l'ensemble des nombres réels que l'on obtient : autrement dit, d'un certain point de vue, l'ensemble des développements binaires "est le même" que celui des développements décimaux. Peano utilise la base 3 dans la construction de sa courbe.

En général, du point de vue théorique, il n'y a guère de raison de privilégier une base plutôt qu'une autre (il n'en est évidemment pas de même pour l'implémentation informatique !) Mais Bailey, Borwein et Plouffe ont découvert assez récemment (1997) qu'il est possible de calculer le n-ème chiffre binaire de π, sans connaître les précédentes ! Leur méthode est basée sur la formule

$$\pi = \sum_{n \geq 0} \frac{1}{16^n} \left(\frac{4}{8n+1} - \frac{2}{8n+4} - \frac{1}{8n+5} - \frac{1}{8n+6} \right),$$

et en particulier sur la présence du facteur $16^{-n} = 2^{-4n}$. Ainsi, le 10^{12}-ème chiffre binaire de π est-il égale à 1, et il est suivi de 0000111, etc...[3]

Exercice. 1) Soit x un réel entre 0 et 1, donné par un développement décimal illimité. Montrer qu'il est possible d'exprimer x en base 2 par un développement binaire illimité.

2) Trouver les premiers chiffres du développement binaire de $\sqrt{2}$ et de π.

3) Soient x et y des développements binaires illimités. Définir l'addition $x + y$ en tant que développement binaire. Montrer que $x + y$ correspond bien au même nombre réel que celui obtenu en effectuant d'abord une "conversion" de x et de y en chiffres décimaux et en additionnant ces développements décimaux.

[3] Cette formule est par exemple démontrée dans P. Eymard et J-P. Lafon : Autour du nombre π, Hermann, Paris 1999.

Chapitre 6

L'ensemble des nombres complexes.

Les nombres complexes sont des nombres tout aussi utiles, que les nombres réels, et pas seulement aux mathématiciens ! Ceci parce qu'ils fournissent le domaine naturel, où beaucoup de problèmes importants trouvent leur solution : essentiellement la résolution d'équations algébriques, et la définition de fonctions élémentaires. [1] Une des difficultés pour comprendre les nombres complexes est due au fait, qu'ils ne sont pas la mesure de quelque chose (d'ailleurs, il n'y a pas de manière utile de dire si un nombre complexe est plus grand qu'un autre). Cette difficulté avait déjà empêché la reconnaissance des nombres négatifs comme des nombres à part entière. Comme les nombres négatifs, les nombres complexes ont d'abord été perçus comme des "fictions", des nombres "imaginaires". Personne ne saurait nier de nos jours leur importance, mais pour les raisons évoquées, leurs compréhension reste souvent difficile au premier contact.[2]

Les nombres complexes présentent deux aspects complémentaires : un *aspect algébrique* et un *aspect géométrique*. Pour comprendre le premier il faut se dire qu'il n'y a rien de plus à comprendre, que... ce qui est défini : un ensemble de couples de réels ayant la propriété cruciale, qu'il admet des opérations algébriques, qui étendent celles des nombres réels, et qui font que de nombreuses équations algébriques, qui n'avaient pas de solution dans les réels en trouvent dans ce nouvel ensemble. Le deuxième aspect est plus facile comprendre, surtout si on pense aux nombres complexes comme à des transformations du plan. Ceci généralise une propriété des réels : multiplier par un nombre réel positif, peut s'interpréter comme faire une homothétie sur la droite, et multiplier par -1 correspond à retourner/inverser (les sens de parcours sur) la droite.

6.1 Pas tous les nombres sont réels (loi des signes).

Rappelons pourquoi les nombres réels ne suffisent pas.

> *"Più via più fa più. Meno via meno fa più. Più via meno fa meno. Meno via più fa meno."*
> (R. Bombelli, L'Algebra, (1562), ed. Feltrinelli, Milano 1966, Libro I, p. 62)

Déjà Diophante avait énoncé cette "loi des signes", mais ce ne furent que les mathématiciens de l'époque de Bombelli, qui commencèrent à utiliser avec aisance les nombres négatifs et d'autres, qui à la plupart semblaient encore bien imaginaires.

[1] Ainsi le problème de savoir quelle valeur donner à $\log(-1)$ amène à définir la fonction logarithme dans le domaine complexe.

[2] Le fait que l'on persiste à les appeler imaginaires n'aide évidemment pas. On devrait peut-être suivre une idée attribuée à Gauss, qui suggérait d'appeler i l'unité latérale (voir l'interprétation géométrique). Noter qu'à une certaine époque il y avait aussi des nombres sourds, etc.

La loi des signes a pour conséquence que l'équation $x^2 = -1$ ne peut pas avoir de solution dans un système de nombres où la loi est valable. L'ordre sur les réels empêche donc d'y trouver une solution : il faut étendre l'ensemble des nombre réels à l'ensemble des nombres complexes pour pouvoir en trouver une ; mais nous verrons que le plus dur est déjà fait.

Exercice. Montrer à partir des propriétés de l'ordre sur les réels, que $-1 < 0$.

6.2 Définition comme couple de réels.

On considère l'ensemble $\mathbf{R} \times \mathbf{R}$ des couples (a, b) de réels, et on identifie $\mathbf{R}$ au sous-ensemble formé des couples $(a, 0)$, ayant la deuxième composante nulle. Lorsque nous allons définir les opérations sur les couples, nous allons veiller à ce que celles-ci étendent les opérations sur $\mathbf{R}$.

La définition de l'opération *somme* est la suivante :

$$(a, b) + (c, d) := (a + c, b + d) \ .$$

On vérifie que $(a, 0) + (c, 0) = (a + c, 0)$ redonne la somme usuelle sur les réels. On pourrait, de manière analogue, définir un produit en posant $(a, b)(c, d) = (ac, bd)$, mais, même si ce produit étend celui des réels, ce n'est pas ce que nous voulons : ce produit est trop semblable à celui dans les réels et donc ne va pas donner des résultats trop intéressants.

La bonne définition du *produit* est celle-ci :

$$(a, b)(c, d) := (ac - bd, bc + ad) \ .$$

Notons tout de suite, qu'en posant $i := (0, 1)$, cette règle pour le produit donne $i^2 = i \cdot i = (0, 1)(0, 1) = (-1, 0)$, que nous identifions à -1. Donc, avec ces définitions on a bien

$$i^2 = -1 \ .$$

(En fait le carré de $-i = (0, -1)$ est aussi -1.) Nous sommes sur la bonne voie. On vérifie aussi que $(a, 0)(c, 0) = (ac, 0)$. On note $\mathbf{C}$ l'ensemble des couples de réels muni de ces deux opérations : c'est l'ensemble des nombres complexes. Les opérations vérifient les propriétés usuelles : $uv = vu$, $u(v + w) = uv + uw$, etc. Dans la suite on notera $a + ib$ le couple (a, b) et on appellera $z = a + ib$ le nombre complexe de *partie réelle* $\mathrm{Re}(z)$ et de *partie imaginaire* $\mathrm{Im}(z)$. Les parties réelles et imaginaire sont donc des nombres réels. Pour c un nombre réel on écrit $c(a, b) = (ca, cb)$.

Exercice. Montrer que toute équation de degré 2 à coefficients réels a une solution dans $\mathbf{C}$. Qu'en est-il des équations à coefficients complexes ? (Indications : résoudre les équations de la forme $x^2 = d$; se ramener à ce type d'équation en "complétant le carré", c'est-à-dire en écrivant $ax^2 + bx + c = a(x + b/2a)^2 + (c - b^2/4a)$.)

Exercice. Trouver l'inverse multiplicatif de $z = a + ib$. (Indication : il s'agit de résoudre l'équation $(a, b)(c, d) = 1 = (1, 0)$.)

Un théorème très important, affirme que *toute équation algébrique de la forme* $z^n + a_1 z^{n-1} + a_2 z^{n-2} + \cdots + a_0 = 0$*, avec* a_i *complexe, a toutes ses solutions dans* $\mathbf{C}$. On résume ce résultat en disant que $\mathbf{C}$ est algébriquement clos, c'est-à-dire que l'on ne peut pas l'étendre et trouver des solutions de ce type d'équations, qui n'y sont pas déjà.

Si $z = (a, b)$ on appelle $\overline{z} = (a, -b)$ le *conjugué* de z. La conjugaison complexe a les propriétés :

1) $\overline{\overline{z}} = z$

2) $\overline{zw} = \overline{z}\,\overline{w}$ et $\overline{z + w} = \overline{z} + \overline{w}$

3) les nombres réels sont exactement les nombres complexes z tels que $\overline{z} = z$
4) si $z = a + ib$, alors $z\overline{z} = a^2 + b^2$ est un nombre réel

On appelle *module* du nombre complexe $z = a + ib$ le nombre réel

$$|z| = (a^2 + b^2)^{1/2} .$$

5) $|zw| = |z||w|$

Vu que $z\overline{z}$ est réel, nous savons qu'il admet un inverse multiplicatif dans $\mathbf{R}$, et vu la propriété multiplicative du module, on s'attend à ce que l'inverse multiplicatif de z vérifie $|z^{-1}| = 1/|z|$. Ces considérations nous amènent à la formule

6)

$$z^{-1} = \frac{\overline{z}}{(z\overline{z})} = \frac{a - ib}{a^2 + b^2}$$

Toutes ces propriétés se vérifient par calcul direct à partir des définitions, mais elles deviennent plus compréhensibles sur la représentations géométrique, que nous allons considérer plus bas. Notons encore quelque propriétés :

7) $|w| = |\overline{w}|$
8) $\mathrm{Re}(z) = (z + \overline{z})/2$ et $\mathrm{Im}(z) = (z - \overline{z})/2i$
9) $|\mathrm{Re}(z)| \leq |z|$
10) *Inégalité du triangle* dans $\mathbf{C}$: si z et w sont des nombres complexes, alors

$$|z + w| \leq |z| + |w| .$$

Montrons ceci. On calcule :

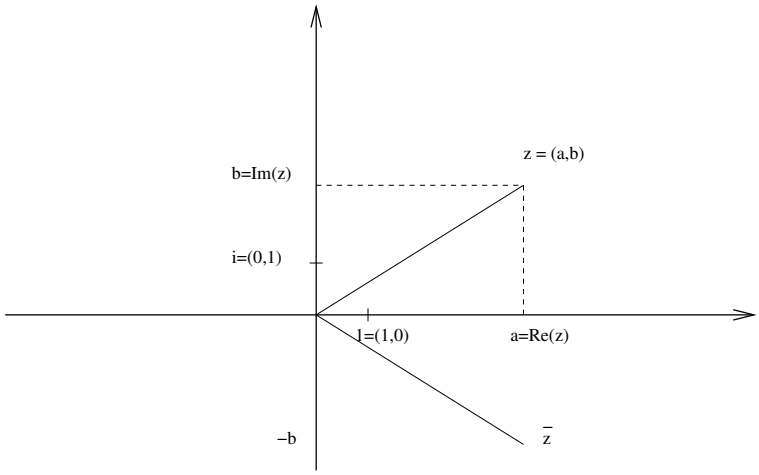
$$\begin{aligned} |z+w|^2 &= (z+w)\overline{(z+w)} \\ &= (z+w)(\overline{z}+\overline{w}) \\ &= z\overline{z} + w\overline{z} + z\overline{w} + w\overline{w} \\ &= |z|^2 + (w\overline{z} + \overline{(w\overline{z})}) + |w|^2 \\ &= |z|^2 + 2\mathrm{Re}(w\overline{z}) + |w|^2 \\ &\leq |z|^2 + 2|w\overline{z}| + |w|^2 \\ &\leq (|z| + |w|)^2 \end{aligned}$$

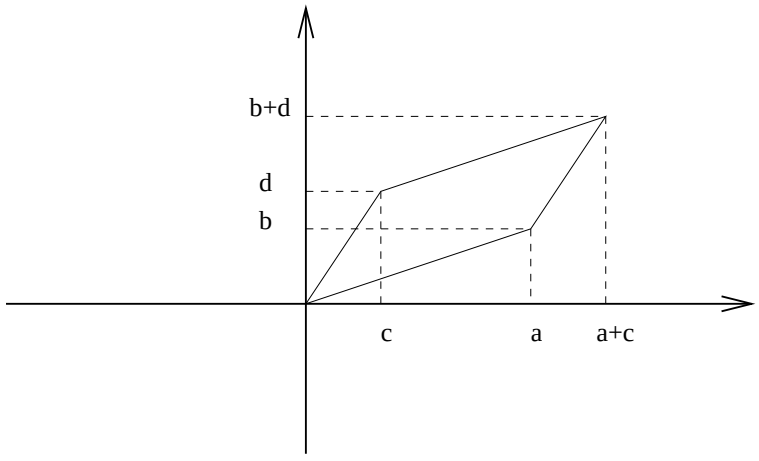
Notons, que nous n'avons pas utilisé l'inégalité du triangle pour les réels, qui donc en résulte comme cas particulier.

6.3 Représentation géométrique.

Pour justifier les considérations qui vont suivre, on revient sur la question de trouver une solution à l'équation $x^2 = -1$. Si on raisonne en termes géométriques et on se souvient que l'on peut interpréter la multiplication par -1 dans les réels comme l'opération consistant à effectuer une symétrie de la droite par rapport à l'origine, alors la question devient : trouver une opération qui répétée deux fois donne la symétrie par rapport à l'origine sur la droite réelle. Nous savons qu'il faut que nous cherchions en dehors de la droite pour trouver une solution. Il se trouve que la solution n'est pas très loin. Il suffit de considérer le plan et l'opération *dans le plan* consistant à effectuer une rotation d'angle droit de centre l'origine. Cette formulation montre que l'on a en effet deux solutions : la rotation dans le sens contraire à celui des aiguilles d'une montre (le sens habituel en géométrie, dit positif) et la rotation en sens opposé. Notons donc $i = (0, 1)$ le point du plan en coordonnées cartésiennes, qui est l'image de $1 = (1, 0)$ par la rotation d'angle droit dans le sens positif, alors il s'agit de définir des opérations sur les points du plan qui donneront $i^2 = -1$, ou plus généralement que la multiplication par i soit la rotation

d'angle droit (dans les sens positif). La solution à ce problème est reportée sur les figures. La première illustre le codage des complexes dans le plan. Il est clair que l'on obtient le conjugué d'un point en considérant la symétrie par rapport à l'axe horizontal des réels. Le module d'un point correspond à la distance du point à l'origine. La deuxième figure donne l'interprétation géométrique de la somme en termes de la "règle du losange", et la troisième précise que le produit de deux points z et w est obtenu en ajoutant les angles à partir de l'axe réel et en faisant le produit des modules des points.

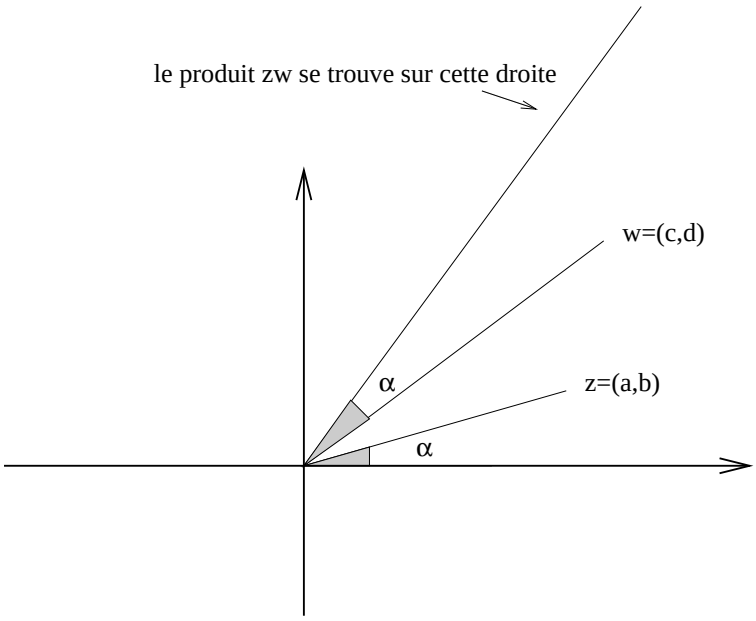




L'angle (en radians) entre la droite portée par $0z$ et l'axe réel s'appelle l'*argument* de z et est noté $\arg(z)$. On a donc :

$$\arg(zw) = \arg(z) + \arg(w) \ .$$

Exercice. Montrer que le module de i vaut 1 et que son argument vaut $\pi/4$. Trouver géométriquement les (trois) solutions complexes de l'équation $z^3 = 1$.



6.4 Coordonnées polaires.

Il est clair que tout point z du plan peut être repéré par sa distance $r = |z|$ à l'origine et par l'angle $\theta = \arg(z) \in \mathbf{R}$ entre la droite portée par $0z$ et l'axe horizontal (par exemple). C'est là une intuition géométrique. La définition de l'argument d'un nombre complexe sera précisée avec l'étude de la fonction exponentielle complexe. Une fois cela fait, nous pourrons donner un sens précis à l'égalité

$$z = a + ib = re^{i\theta} \ .$$

Le complexe $e^{i\theta}$ est un nombre de module 1, un point du cercle unité, comme il résulte par exemple de la fameuse *formule d'Euler*

$$e^{it} = \cos t + i \sin t \ ,$$

qui sera aussi expliquée avec l'étude de l'exponentielle, que nous allons faire plus bas. Le nombre θ est donc défini à 2π près, et les *coordonnées polaires* de z sont $(\theta, r) \in [0, 2\pi[\times\mathbf{R}_{>0}$.

6.5 Distances, boules.

Nous avons déjà observé que les nombres complexes ne peuvent pas être utilisés pour mesurer des grandeurs : on montre qu'il n'y a pas d'ordre sur les complexes, qui soit compatible avec les opérations somme et produit, on ne peut pas partager les complexes en deux sous-ensembles (plus 0) dont l'un jouerait le rôle des nombres positifs et l'autre celui des nombres négatifs. Il est important de noter que cela ne nous empêche pas de développer la théorie des fonctions de la variable complexe, au même titre que celle des fonctions de la variable réelle. En effet *ce qui compte est que l'on sache dire si deux complexes sont proches.* Pour ça il suffit d'avoir une notion de distance, qui dans le cas des réels se confond (partiellement) avec la notion d'ordre.

Il est clair sur le modèle géométrique, que l'on peut définir une notion de *distance* entre des nombres complexes z et w : c'est le nombre réel positif ou nul

$$\mathrm{dist}(z, w) = |z - w| \ .$$

On vérifie facilement que $\mathrm{dist}(z, w) = \mathrm{dist}(w, z)$ et, à l'aide de l'inégalité du triangle, que

$$\mathrm{dist}(z, w) \leq \mathrm{dist}(z, a) + \mathrm{dist}(a, w) \ .$$

(Écrire $|z - w| = |(z - a) + (a - w)|$.) De plus, $\mathrm{dist}(z, w) = 0$ si et seulement si $z = w$.

L'analogue des intervalles, ce sont les *boules*.[3] La *boule fermée* de centre complexe a et rayon (réel) $r > 0$ est l'ensemble

$$B(a, r) := \{z \in \mathbf{C} : \mathrm{dist}(a, z) \leq r\} \ .$$

La *boule ouverte* de centre a et rayon r est

$$B(a, r)^\circ := \{z \in \mathbf{C} : \mathrm{dist}(a, z) < r\} \ .$$

Exercice. Vérifier que si a est réel, alors l'intersection avec l'axe réel des boules de rayon r s'identifie aux intervalles $[a - r, a + r]$ pour la boule fermée, et à $(a - r, a + r)$ pour la boule ouverte.

[3] On pourrait les appeler des disques, mais on pense ici aux généralisations en dimension supérieure.

Chapitre 7

Limites, continuité, dérivabilité.

Comment savoir que le procédé d'Archimède va vraiment permettre de calculer les décimales de π ? Comment sait-on qu'après un grand nombre d'étapes, le périmètre du polygone utilisé ne va pas cesser d'approcher les bonnes décimales ? N'est-il vraiment pas possible que, entre le n-ème découpage et le suivant, le 1000-ème chiffre, par exemple, ne change pas entre 8 et 9 ? Plus généralement, si l'on a une suite de nombres rationnels (ou réels) dont plus en plus de décimales semblent coïncider avec celles d'un nombre réel x fixé, comment savoir si cette propriété reste valide "à la limite" ?

Dans un autre ordre d'idées, comment donner un sens à l'idée de "vitesse instantanée" ou de "variation continue" d'un ensemble de données.

Pour traiter de telles questions il faut définir précisément ce que l'on entend par "approximations de plus en plus précises" et par "limite".

Nous allons ici le faire de deux manières, d'abord en nous basant sur la contemplation des décimales des nombres réels, puis de manière plus synthétique en nous basant sur la notion de distance. Cette deuxième approche permet de traiter en même temps les fonctions d'une variable complexe.

7.1 Le cathé des limites.

Pour insister d'entrée sur l'importance de la notion de limite, nous proposons d'apprendre par coeur et de réciter à haute voix, sans se prendre trop au sérieux, la litanie suivante [1] :

Q1 *Qu'est-ce qu'une dérivée–vraiment ?*
R1 Une limite.
Q2 *Qu'est-ce qu'une intégrale–vraiment ?*
R2 Une limite.
Q3 *Qu'est-ce qu'une somme infinie comme* $1 + 1/4 + 1/9 + \cdots$ *–vraiment ?*
R3 Une limite.
Q4 *Mais alors, qu'est-ce qu'une limite ?*
R4 Un nombre. Et ça on connaît !

[1] Inspiré du livre de E. Hairer et G. Wanner, Analysis by its history, Springer, New York, 1996, Chap. III ; une litanie est une prière formée d'une suite de courtes invocations.

7.2 Définitions I : limites via les DDI.

Commençons par définir la convergence des suites de nombres réels, vu que celles-ci sont le plus proches des DDI.

Définition. Soit $N \geq 1$ un entier. Une suite (x_n) de nombres réels, définie pour $n \geq 1$, *converge avec une précision de N décimales* vers un réel x donné si, à partir d'un certain rang, le signe, la partie entière et les N premières décimales de x_n sont identiques à celles de x.

Autrement dit, si l'on écrit les développements décimaux concernés sous forme d'un tableau

$$\begin{aligned} x_1 &= \pm m_1, d_1 \dots d_N \dots \\ x_2 &= \pm m_2, e_1 \dots e_N \dots \\ \vdots &= \vdots \\ x_n &= \pm m_n, z_e \dots z_N \dots \\ x &= \pm m, \delta_1 \dots \delta_N \dots, \end{aligned}$$

il faut que si n est plus grand qu'un certain entier n_0, les N premières colonnes soient identiques.

Si (x_n) vérifie cette propriété, et si l'entier n_0 (tel que les N premières décimales coïncident pour $n \geq n_0$) est connu, on "connaît" automatiquement les N premières décimales du réel x : il suffit de regarder celles de x_{n_0}, ou de x_n pour n'importe quel $n \geq n_0$.

La définition que nous avons donnée est intuitive mais telle quelle, elle peut poser problème du point de vue des développements décimaux du type $0, \overline{9} = 1, \overline{0}$. On la rend parfaitement rigoureuse, et plus maniable, en procédant ainsi : pour chaque réel

$$x = \pm m, d_1 \dots d_n \dots$$

et chaque entier $N \geq 1$, on définit l'approximation de x avec N décimales comme le nombre donné par le développement décimal limité

$$x = \pm m, d_1 \dots d_N \overline{0}$$

qui est donc un nombre rationnel. Si l'on effectue cette opération pour tout les éléments d'une suite (x_n), on obtient des rationnels (y_n), et de même on a le rationnel y correspondant à x. La définition est alors simplement que $y_n = y$ pour tout $n \geq n_0$, autrement dit, mis à part certains termes initiaux, la suite (y_n) est constante.

Définition. Une suite (x_n) de nombres réels, définie pour $n \geq 1$, *converge* vers un réel x donné si, pour tout $N \geq 1$, elle converge vers x avec une précision de N décimales. On note cela sous les formes suivantes

$$x_n \to x \quad , \quad \lim x_n = x \quad , \quad \lim_{n \mapsto \infty} x_n = x \ .$$

On dit aussi que x est la limite de la suite (x_n).

Si la suite (x_n) ne converge pas, on dit qu'elle *diverge*. Il n'y a pas de notation standard pour cela.

Pour un calcul numérique sur un ordinateur donné, il ne serait *a priori* pas nécessaire de définir d'autre notion d'approximation que celle à N décimales, pour N fixé, car la précision de chaque processeur est forcément limitée. Mais puisque la puissance des machines évolue, il n'est pas possible véritablement de fixer une précision de manière définitive !

Exemple. L'exemple le plus simple et le plus important est celui de la suite des approximations décimales elle-même. Soit x un nombre réel

$$x = \pm m, d_1 \dots d_n \dots$$

On pose

$$\begin{aligned} x_0 &= \pm m, \overline{0} \\ x_1 &= \pm m, d_1\overline{0} \\ x_2 &= \pm m, d_2 d_2\overline{0} \end{aligned}$$

et plus généralement

$$x_n = \pm m, d_1 \dots d_n\overline{0}.$$

Soit $N \geq 1$. On voit que pour tout $n \geq N$, les N premières décimales de x_n et x sont identiques par construction. Donc, par définition, x_n converge vers x avec une précision de N décimales. Mais N est quelconque, donc on en déduit que

$$\lim x_n = x.$$

En particulier, cela montre que tout nombre réel peut s'écrire comme la limite d'une suite de nombres rationnels. Cette propriété est une des propriétés fondamentales de l'ensemble $\mathbf{R}$ des nombres réels.

Exercice. Soit $x = 0, d_1 \dots d_n \dots$ un nombre réel. On pose

$$\begin{aligned} x_0 &= 0 \\ x_1 &= 0, \overline{d_1} \\ x_2 &= 0, d_1\overline{d_2} \\ x_n &= 0, d_1 \dots \overline{d_n}, \end{aligned}$$

et ainsi de suite. Montrer que $\lim x_n = x$.

Exercice. Soit a un nombre réel. On note $x_0 = 1$, $x_n = a^n$ pour $n \geq 1$. 1) Montrer que x_n converge vers 1 pour $a = 1$.

2) Montrer que x_n diverge pour $a = -1$, ou pour $|a| > 1$.

3) Montrer que x_n converge vers 0 pour $|a| < 1$.

Exercice. Soit (x_n) une suite de nombres entiers. Montrer que la suite (x_n) ne peut converger que si elle est constante à partir d'un certain rang.

On peut aussi définir les suites de nombres complexes, mais pour cela il est plus simple de se baser sur la définition en termes de distances, qu'en termes de DDI.

Considérons maintenant une fonction f définie sur $\mathbf{R}$ ou sur un sous-ensemble et, étant donné un réel x_0, pour lequel $f(x_0)$ peut être ou ne pas être défini, demandons nous quel est le comportement de $f(x)$ lorsque x est proche de x_0. Par exemple, soit $f(x) = \sin(x)/x$; cette fonction est définie sur $\mathbf{R} - \{0\}$ (si l'on sait comment définir $\sin(x)$ pour $x \in \mathbf{R}$!), mais pas en 0. Cependant, si l'on calcule $f(x)$ pour x très petit (à la calculatrice par exemple), on voit que $f(x)$ est proche de 1.

Définition. Soit f une fonction définie sur $\mathbf{R}$, sauf éventuellement en $x_0 \in \mathbf{R}$, à valeurs réelles. On dit que f *admet la limite* $\ell \in \mathbf{R}$ quand $x \to x_0$ si, pour toute précision N donnée, il existe un entier N_1 tel que les N premiers chiffres de $f(x)$ et de ℓ sont les mêmes lorsque les N_1 premiers chiffres de x et de x_0 sont les mêmes. On note alors

$$\ell = \lim_{x \to x_0} f(x).$$

Parfois f n'est pas définie sur $\mathbf{R} - \{x_0\}$ entier, mais sur $]x_0, +\infty[= \{x \mid x > x_0\}$, par exemple. Dans ce cas on doit adapter la définition en ne calculant que $f(x)$ pour $x > x_0$, ce qui définit la *limite de* f *en* x_0 *par valeurs* $> x_0$, notée

$$\lim_{x \to x_0, x > x_0} f(x)$$

si elle existe. De même pour

$$\lim_{x \to x_0, x < x_0} f(x)$$

Exercice. Montrer (ou essayer de comprendre) le critère suivant qui ramène les limites de fonctions à des limites de suites : la limite de $f(x)$ quand $x \to x_0$ existe et vaut ℓ si et seulement pour toute suite (x_n) telle que (x_n) converge vers x_0, la suite $f(x_n)$ converge également et sa limite est ℓ.

7.3 Définitions II : limites via les distances.

Soit $X \subset \mathbf{C}$ un sous-ensemble quelconque, par exemple $X = \mathbf{R}$, ou X une partie de $\mathbf{R}$, comme $\mathbf{N}$. Soit

$$f : X \to \mathbf{C}$$

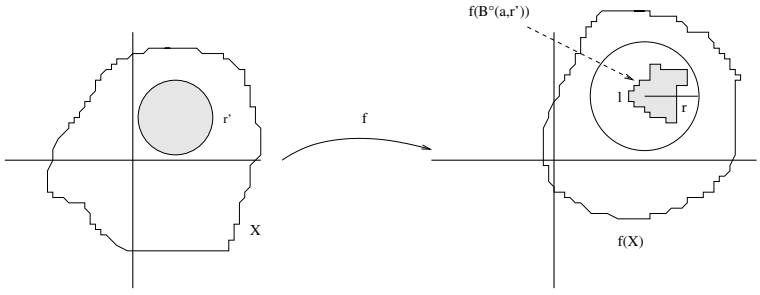
une application. Rappelons que cela signifie que X est le domaine de f. Enfin, soit a et ℓ deux éléments de $\mathbf{C}$. On veut définir ce que signifie "f a pour limite ℓ en a", sans utiliser les DDI. On écrira, comme précédemment,

$$\lim_{x \to a} f(x) = \ell \ .$$

La définition que nous avons donnée via les DDI signifie que *les valeurs de f s'approchent arbitrairement de ℓ au voisinage de a*. Aussi, si on pense aux approximations successives, que *pour tout $r > 0$ la distance $d(f(x), \ell)$ est inférieure à r pour x assez proche de a*. Si on veut préciser ce que l'on entend par "assez proche" on dit mieux, que *pour tout $r > 0$ on a $d(f(x), \ell) < r$ dès que $(x \in X)$ et $d(x, a) < r'$ pour un r' convenable*. Le r est la tolérance, la précision, qui mesure la distance à la limite, et qui doit pouvoir être choisie quelconque, et en particulier aussi petite que l'on veut. Le r' peut évidemment dépendre de r. La version formelle de la définition est la suivante.

Définition.

$$\forall r > 0, \ \exists r' > 0 \ : \ x \in X, \ |x - a| < r' \Rightarrow |f(x) - \ell| < r \ .$$



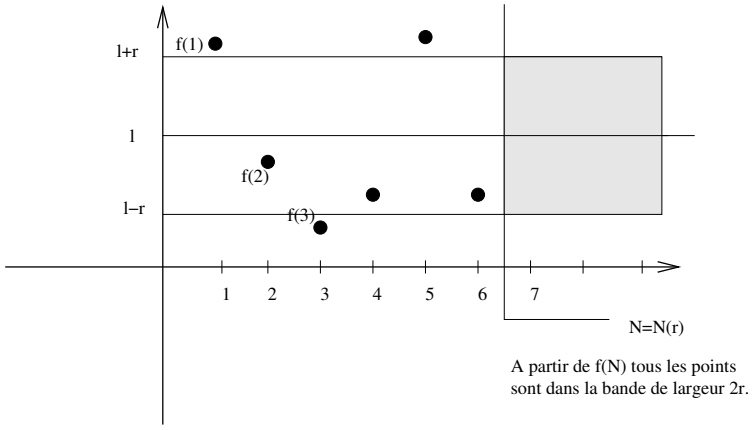
Celle-ci est la définition classique (qui remonte à Weierstrass). D'habitude elle est formulée en termes du duo $\delta - \epsilon$ au lieu du duo $r' - r$

Montrons que la formulation donnée ici est équivalente à la définition précédente pour les fonctions de la variable réelle. Comme nous l'avons noté le nombre r tient le rôle de la précision (de N chiffres) demandée, puisque le fait de demander que $|f(x) - \ell| < r$ revient à demander, si r est $< 10^{-N}$ par exemple, que les N premiers chiffres de $f(x)$ et ℓ coïncident. De même, r' tient le rôle de la précision nécessaire sur les valeurs de x pour connaître $f(x)$ avec une précision de r. Comme r est arbitraire, on

peut prendre $r = 10^{-N}$ pour N quelconque ; si alors N_1 est un nombre quelconque tel que $10^{-N_1} < r'$, la condition d'avoir les N_1 premiers chiffres égaux sur x et x_0 implique $|x-a| < r'$, donc $|f(x)-\ell| < r$. Ainsi la définition avec $r' - r$ implique la définition donnée auparavant.

Réciproquement, étant donné $r > 0$, on peut trouver une précision de N chiffres telle que si $f(x)$ et ℓ coïncident avec une telle précision, alors $|f(x) - \ell| < r$, et on en déduit que la première définition implique celle en termes de $r' - r$.

Exercice. Écrire formellement que f ne tend pas vers ℓ quand x tend vers a.



Pour $X \subset \mathbf{R}$ on définit ce que signifie

$$\lim_{x\to\infty} f(x) = \ell \ ,$$

que l'on lit "f tend vers ℓ pour x tendant vers l'infini". On entend par là, que *les valeurs de f s'approchent arbitrairement de ℓ pour x assez grand*. On écrit $-\infty$ pour signifier la même chose avec "petit" à la place de "grand" dans la dernière phrase. Nous traduisons "assez grand" (resp. "assez petit"), par "plus grand qu'un entier naturel N arbitraire" (resp. "plus petit que l'inverse d'un entier naturel N arbitraire"). Ce qui donne, formellement la définition

$$\forall r > 0,\ \exists N \in \mathbf{N}\ :\ x \in X,\ x > N \Rightarrow |f(x) - \ell| < r \ .$$

Exercice. Écrire la définition formelle de $\lim_{x\to-\infty} f(x) = \ell$.

Exercice. On définit une suite de points dans $\mathbf{C}$ par

$$z_n = 1 + i + i^2/2 + i^3/2\cdot 3 + i^4/2\cdot 3\cdot 4 + \cdots + i^n/2\cdots n \ .$$

Deviner la limite de cette suite.

7.4 Exemples de suites numériques.

Pour faire le calcul explicite de limites de suites, on peut évidemment essayer de vérifier directement la définition. En pratique il est beaucoup plus commode d'utiliser quelques résultats généraux, qui permettent de se ramener à des cas élémentaires. Ce sont des énoncés du type suivant.

Soit (a_n) et (b_n) des suites numériques. Supposons

$$\lim_{n\to\infty} a_n = a \quad \text{et} \quad \lim_{n\to\infty} b_n = b \ .$$

Alors :

$$\lim_{n\to\infty}(a_n + b_n) = a + b \quad \lim_{n\to\infty}(a_n b_n) = ab$$

et si $b \neq 0$, alors $b_n \neq 0$ pour n suffisamment grand et

$$\lim_{n\to\infty} \frac{1}{b_n} = \frac{1}{b} \ .$$

Si $M \in \mathbf{R}$ est tel que $x_n \leq M$ pour tout n alors

$$\lim x_n \leq M \ .$$

Si $M \in \mathbf{R}$ est tel que $|x_n| \leq M$ pour tout n alors

$$|\lim x_n| \leq M \ .$$

Exercice. Montrer qu'il est possible que $x_n < M$ pour tout n mais $\lim x_n = M$. (Indications : considérer $x_n = M - 1/n$, $n = 1, 2, ...$)

Les démonstrations des affirmations ci-dessus se font comme pour les limites de fonctions, que nous allons traiter plus loin.

Dans toutes les manipulations de limites, il faut prendre soin de d'abord vérifier l'existence de celles-ci. Dans le cas contraire, on est rapidement mené à des paradoxes inextricables (des mathématiciens fameux s'y sont fait prendre...) Par exemple "posons"

$$S = \lim(1 + (-1) + \cdots + (-1)^n).$$

On est tenté d'une part d'écrire

$$-S = -1 + 1 + \cdots + (-1)^{n+1} + \cdots = S - 1$$

donc $S = 1/2$, mais d'autre part

$$S = 1 - 1 + 1 + \cdots + (-1)^n + \cdots = (1 - 1) + (1 - 1) + \cdots = 0.$$

L'erreur est que la limite n'existe pas : en effet par définition S est la limite de la suite x_n où x_n est la somme partielle de n termes. On a $x_{2n} = 0$ et $x_{2n-1} = 1$ pour $n \geq 1$. Donc même la partie entière n'est pas constante à partir d'un certain rang, ce qui veut dire que (x_n) diverge.

Considérons maintenant quelques exemples de *calculs de limites de suites* numériques de base.

1) Soit $a_n = 1$ la suite constante de valeur 1, alors $\lim_{n\to\infty} a_n = 1$. En effet pour tout $r > 0$ on peut prendre $N = 0$ et on aura bien que, si $n \geq N$, alors $|a_n - 1| = 0 < r$.

2) Pour $n \geq 1$, soit (a_n) la suite définie par $a_n = 1/n$, alors $\lim_{n\to\infty} a_n = 0$. Pour $r > 0$ donné on cherche N tel que, si $n > N$, alors $|1/n - 0| = 1/n < r$. On voit qu'il suffit de prendre $N > 1/r$. Un tel N existe par la propriété archimédienne de $\mathbf{R}$. Ici N dépend de r.

3) **Exercice.** Montrer que la suite $a_n = \sqrt{n}$ n'admet pas de limite lorsque n tend vers l'infini.

4) Pour $n \geq 1$, soit (a_n) la suite définie par $a_n = (-1)^n + 1/n$. Cette suite ne converge pas. Par exemple, dès que r est tel que $0 < r < 1$, il existe une infinité de termes de la suite a_n en dehors de l'intervalle ouvert $(1-r, 1+r)$ (prendre n impair).

5) Soit q un nombre réel avec $|q| < 1$. Alors

$$\lim_{n\to\infty} q^n = 0 \ .$$

Supposons $q > 0$. Pour $r > 0$ donné il faut trouver N tel que, si $n > N$, alors $|q^n| < r$. On réécrit l'hypothèse $|q| < 1$ comme $1 = q + t$ avec $t > 0$ ($t \in \mathbf{R}$). Alors, en utilisant la formule du binôme :

$$1 = 1^{n+1} = (q+t)^{n+1} = q^{n+1} + (n+1)q^n t + \dots \ .$$

On en déduit, vu que $q > 0$ et $t > 0$, l'inégalité $0 < (n+1)q^n t < 1$ ou encore $0 < q^n < 1/t(n+1)$. Par conséquent, pour r donné (et t fixé par q), si n est tel que

$$\frac{1}{t(n+1)} < r \quad \text{ou encore} \quad \frac{1}{tr} - 1 < n \ ,$$

alors $q^n < r$. Notons, que le résultat est aussi valable pour q complexe, car q complexe est proche de 0 si et seulement si son module $|q|$ est proche de 0.

6) Soit x un nombre réel strictement positif. Alors

$$\lim_{n\to\infty} x^{1/n} = 1 \ .$$

Ceci est certainement vrai si $x = 1$. Supposons d'abord $x > 1$ et écrivons $x^{1/n} = 1 + x_n$, pour un certain x_n tel que $x_n > 0$. Par la formule du binôme on a

$$x = (1 + x_n)^n = 1 + nx_n + (\dots) \ ,$$

où le terme $(\dots)$ est positif. Donc $0 < x_n < (x-1)/n$ et vu que $1/n$ tend vers 0 lorsque n tend vers l'infini on a bien que x_n tend vers 0, ce qui entraîne le résultat.

7) **Séries.** Donné une suite (a_n) on considère la *série associée*, c'est la suite définie par

$$\begin{aligned} s_0 &= a_0 \\ s_1 &= a_0 + a_1 \\ &\vdots \\ s_k &= a_0 + a_1 + \cdots + a_k \end{aligned}$$

Si la limite $\lim_{n\to\infty} s_n$ existe on la note

$$\sum_{n=0}^{\infty} a_n \ .$$

Par exemple les *développements décimaux illimités* sont des séries.

Exercice. 1) On considère une suite arbitraire de chiffres (d_n), et on construit la série $\sum s_n$ avec $s_n = d_n 10^{-n}$ (donc s_n est un nombre rationnel). Montrer que la série $\sum s_n$ converge et que

$$\sum_{n=1}^{\infty} d_n 10^{-n} = 0, d_1 \dots d_n \dots ,$$

et en déduire que (1) est valide en interprétant le terme de droite comme la somme de la série correspondante.

2) Calculer $0,\overline{9}$ de nouveau en utilisant (1)).

Un autre exemple important est donné par les *séries géométriques*. Ce sont les séries obtenues de suites de la forme $a_n = q^n$. On parle alors de la série de raison q. En utilisant l'égalité :

$$s_n(q) = 1 + q + q^2 + \cdots + q^n = \frac{1 - q^{n+1}}{1 - q}$$

et le (5) on voit que pour la série géométrique de raison q *avec* $|q| < 1$

$$\sum_{n=0}^{\infty} s_n(q) = \lim_{n\to\infty} s_n(q) = \frac{1}{1-q} \ .$$

La *série exponentielle* $\exp(x)$ est obtenue à partir de la suite des puissances divisées $x^n/n! =: x^{[n]}$. Nous allons voir que pour tout x réel positif cette série converge et l'on écrira :

$$\exp(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!} \ .$$

Montrons d'abord

$$\lim_{n\to\infty} \frac{x^n}{n!} = 0 \ .$$

Écrivons $a_n = x^n/n!$. Remarquons que si $n > p$, alors

$$a_n = a_p \frac{x}{(p+1)} \cdots \frac{x}{n} \ .$$

Si p est tel que $p \leq 10x < p+1$, pour $q \geq p+1$ on a $\frac{x}{q} < \frac{1}{10}$ et donc

$$a_n \leq a_p \left(\frac{1}{10}\right)^{n-p} \ .$$

Soit $r > 0$. Pour x fixé choisissons p comme ci-dessus, alors, vu que la suite $(1/10)^{n-p}$ tend vers 0, on voit que l'on peut choisir N tel que $n > N$ implique $|x^n/n!| < r$.

En utilisant le fait qu'une suite convergente est bornée, on obtient de ce qui précède que si $x \geq 0$, alors il existe un entier M tel que pour tout n on a $(2x)^n/n! \leq M$. Par conséquent

$$s_n(x) = 1 + x + \cdots + \frac{x^n}{n!} \leq M\left(1 + \frac{1}{2} + \cdots + \frac{1}{2^n}\right) \leq 2M \ .$$

En conclusion on voit que la série $s_n(x)$ est croissante et majorée et par conséquent elle admet une limite dans $\mathbf{R}$ (par la propriété (SC) du sup des suites croissantes). On note $\exp(x)$ sa limite.

On peut montrer qu'en fait la série exponentielle converge pour tout x complexe.

Exercice. Donner un exemple de suites (x_n) et (y_n) qui divergent telles que $(x_n + y_n)$ converge.

Exercice. Comment savoir si une suite converge si l'on ne connaît pas sa limite à l'avance ? (Par exemple, si les éléments de la suite sont fournis par une expérience). *Si* on sait quand même que la suite converge, on peut déterminer des approximations de la limite puisque par définition les N premiers chiffres de celle-ci seront égaux à ceux de tout terme de la suite d'indice assez grand. En particulier,

pour toute précision de N chiffres donnée, il existe N_1 tel que les N premiers de x_n et x_m coïncident, pour tout n et m tout les deux $\geq N_1$.

Réciproquement, on peut remarquer qu'un tel "test" peut être effectué sans savoir si la suite converge. C'est ainsi que Cauchy avait défini la limite des suites. Disons donc qu'une *suite* (x_n) *est de Cauchy* si, pour toute précision N, il existe N_1 tel que les N premiers chiffres de x_n et de x_m coïncident dès que $n \geq N_1$ et $m \geq N_1$.

Montrer (ou essayer du moins de comprendre) que toute suite de Cauchy est convergente.

Voici le critère le plus simple d'existence de limites de suites, qui ne nécessite pas de connaître la limite par avance et qui n'est rien d'autre qu'une version de la propriété (AC). La démonstration reprend la même idée que celle du fait que $\mathbf{R}$ a la propriété (AC).

Proposition. 1) Soit (x_n) une suite croissante et majorée par $A \in \mathbf{R}$, c'est à dire que $x_n \leq A$ pour tout n. Alors la suite (x_n) converge vers une limite a telle que $a \leq A$.

2) Soit (x_n) une suite décroissante et minorée par $A \in \mathbf{R}$, c'est à dire que $x_n \geq A$ pour tout n. Alors la suite (x_n) converge vers une limite a telle que $a \geq A$.

Démonstration. La preuve de (2) est similaire à celle de (1), et peut même s'en déduire en appliquant (1) à $y_n = -x_n$, qui sera majorée par $-A$.

Considérons donc une suite croissante (x_n). Pour simplifier, on suppose $x_n \geq 0$ pour tout n. Soit m_n la partie entière de x_n. On a $m_n \leq m_{n+1}$ et $m_n \leq A$ pour tout n. Mais il s'agit d'une suite d'entiers et donc, nécessairement, la valeur de m_n doit être constante à partir d'un certain rang. Soit m cette valeur commune.

On a donc $m \leq x_n \leq m+1$ pour tout n plus grand qu'un indice N_0. Regardons la suite (d_n) des premiers chiffres du développement décimal de x_n pour $n \geq N_0$. On a $d_n \leq d_{n+1}$ car $x_n \leq x_{n+1} \leq m+1 \leq x_n+1$. Mais d_n est une suite d'entiers entre 0 et 9. Donc d_n doit être constante à partir d'un certain rang $N_1 \geq N_0$, égale à e_1 disons. Cela montre déjà que la suite (x_n) converge avec une précision de une décimale vers $a_1 = m, e_1\overline{0}$.

Par récurrence, supposons que (x_n) converge avec une précision de K décimales avec

$$a_K = m, e_1 e_2 \dots e_K \overline{0} \ .$$

Soit N_K tel que les K premiers chiffres de x_n sont égaux à $e_1 e_2 \dots e_K$ pour $n \geq N_K$. La suite (f_n) des $(K+1)$-ème chiffres de x_n est encore croissante, et donc f_n sera constant, égal à e_{K+1} pour tout $n \geq N_{K+1}$ pour un certain $N_{K+1} \geq N_K$. Cela montre que x_n converge vers $m, e_1 e_2 \dots e_K e_{K+1}\overline{0}$ avec une précision de $(K+1)$ décimales.

En définitive, ceci construit par récurrence le développement décimal illimité $a = m, e_1 e_2 \dots e_n \dots$ et par définition on a bien

$$\lim x_n = a.$$

Comme $x_n \leq A$ pour tout n, on sait que $\lim x_n = a \leq A$ (voir la ci-dessus).

Exercice. Soit (x_n) une suite croissante, (y_n) une suite décroissante. On suppose que $x_n \leq y_n$ pour tout n.

1) Montrer que (x_n) converge vers une limite x et (y_n) converge vers une limite y.

2) On suppose que la suite $y_n - x_n$ converge vers 0. Montrer alors que

$$x = \lim x_n = \lim y_n = y$$

(on dit que (x_n) et (y_n) sont des *suites adjacentes.*)

Exercice. Montrer qu'on peut ramener toutes les limites à l'étude de suites positives tendant vers 0 : précisément, montrer que (x_n) converge et

$$\lim x_n = a$$

si et seulement si $(|x_n - a|)$ converge et

$$\lim |x_n - a| = 0.$$

7.5 Existence de la borne supérieure.

Le travail que nous avons effectué jusqu'ici nous permet déjà de montrer que l'ensemble des réels a la propriété du sup.

Proposition. Soit $X \subset \mathbf{R}$ un sous-ensemble non-vide et majoré, c'est à dire que $X \subset]-\infty, M]$ pour un certain $M \geq 0$. Alors X admet une borne supérieure.

Démonstration. Pour simplifier on suppose que $X \subset [0, 1/2]$. Pour tout entier $k \geq 1$, découpons l'intervalle $[0, 1[$ en intervalles du type $I_{i,k} = [i/10^k, (i+1)/10^k[$ de longueur 10^{-k}. Il y a un nombre fini de tels intervalles pour k donné. Regardons parmi ceux-ci les intervalles $I_{i,k}$ tels que $I_{i,k}$ ne rencontre pas X. Comme $X \subset [0, 1/2]$ et $1/10 < 1/2$, il en existe. Considérons le plus petit i, disons i_k, tel que $I_{i,k}$ ne rencontre pas X, et tel que tout les intervalles à sa droite (c'est à dire $I_{j,k}$ avec $j > i$) ne rencontrent pas X. Soit $x_k = i_k/10^k$ la borne inférieure de l'intervalle $I_{i_k,k}$. Par définition, x_k est un majorant de X.

Maintenant, on voit facilement que la suite (x_k) est décroissante. Puisque X est non-vide, il existe $y_0 \in X$, donc $x_k \geq y_0$ pour tout k, et (x_k) est minorée. D'après la proposition du paragraphe précédent, la suite (x_k) converge. Notons

$$x = \lim x_k.$$

On va voir que x est la borne supérieure de X. Tout d'abord, puisque x_k est un majorant pour tout k, on a $y \leq x_k$ pour tout $y \in X$ fixé, $k \geq 1$, et donc par les propriétés énoncées, on a $y \leq \lim x_n = x$, ce qui montre que x est un majorant de X.

Soit $z < x$. On veut trouver $y \in X$ tel que $z < y \leq x$. Si $z < 0$, c'est évident (tout $y \in X \subset [0, 1/2]$ conviendra). Sinon, puisque $z < x$, on peut trouver pour tout k assez grand des intervalles $I_{i,k}$ pour un certain k assez grand (si $10^{-k} < z - y$) tels que

$$z \in I_{i,k} \subset I_{i+1,k} \subset I_{i_k,k},$$

avec des inclusions strictes car la distance entre z et $x_k \in I_{i_k,k}$ est plus grande que $2 \cdot 10^{-k}$. D'après la définition de i_k, il n'est pas possible qu'aucun des intervalles intermédiaires $I_{j,k}$, $i + 1 \leq j < i_k$, ne rencontre X ; donc il existe un $y \in X$ dans un de ceux-ci, et alors $z < y$, et $y \leq x$ puisque x est un majorant de X.

Exercice. En utilisant la propriété de la borne supérieure, montrer que toute suite croissante et majorée (x_n) converge et que sa limite est

$$\lim x_n = \sup\{x_n\}.$$

Voici un exemple utilisant la propriété du sup.

Proposition. Pour tout $x \geq 0$, il existe $y \in \mathbf{R}$ tel que $y^2 = x$.

Démonstration. Soit X l'ensemble des $z \geq 0$ tels que $z^2 \leq x$. Puisque $0 \in X$, X n'est pas vide. Puisque $x \mapsto x^2$ est croissante, X est majoré : par exemple, si $x \geq 1$, on a $(x+1)^2 > x^2 \geq x$ donc $X \subset [0, x]$, et si $x \leq 1$, on a $X \subset [0, 1]$.

On note alors $y = \sup X$ la borne supérieure de cet ensemble X. Il s'agit de montrer que $y^2 = 2$. Cela tient essentiellement à la continuité de la fonction $f(x) = x^2$, que nous montrerons à la Sect. 7.8.

Soit $\epsilon = x - y^2$. On va montrer $\epsilon = 0$. Supposons d'abord $\epsilon > 0$. D'après la continuité de f par $\epsilon - \delta$, il existe $\delta > 0$ tel que si $|z - y| < \delta$, on a $|z^2 - y^2| < \epsilon$. Prenons $z = y + \delta/2$, donc $z > y$, $|z - y| < \delta$. Par monotonie on a $z^2 > y^2$ et donc $z^2 < y^2 + \epsilon < x$. Alors $z \in X$, mais $z > y$, ce qui contredit le fait que y soit un majorant de X. L'hypothèse $\epsilon > 0$ est donc intenable.

Similairement, supposons $\epsilon < 0$. Il existe $\delta > 0$ tel que si $|z - y| < \delta$, on a $|z^2 - y^2| < |\epsilon|$. Soit $z = y - \delta/2$ donc $z < y$. On a $z^2 < y^2$ et $z^2 - y^2 < \epsilon$ donc $z^2 > x$. Cela implique que z est un majorant de X : en effet, si $y \in X$, on a $y^2 \leq x < z^2$ donc $y \leq z$. Mais z est alors un majorant de X tel que $z < y$, donc il n'existe pas de $y \in X$ tel que $z < y \leq x$, et cela contredit la seconde partie de la définition de la borne supérieure.

7.6 Continuité via les DDI.

La méthode utilisée pour montrer que, si deux suites convergent, leur somme converge également et sa limite est la somme des deux limites individuelles, est beaucoup plus générale et s'applique à toute "opération" portant sur des nombres réels a, b, ..., disons $f(a, b, \ldots)$ qui a la propriété que les N premiers chiffres de $f(a, b, \ldots)$ peuvent être déterminés en utilisant seulement un nombre fini de chiffres de a, b, ... (Dans l'exemple précédent, les N premiers chiffres de $f(a, b) = a + b$ ne dépendent que des $N + 2$ premiers chiffres de a et b). Cette propriété n'est autre que la continuité de l'opération f.

Dans cette section, on s'intéressera pour simplifier aux applications f ne dépendant que d'une variable réelle.

Définition. Soit f une fonction quelconque définie sur $\mathbf{R}$ et à valeurs réelles.

1) On dit que f est *uniformément continue* si pour tout $N \geq 1$, il existe $M \geq 1$ tel que les N premiers chiffres de $f(x)$ sont identiques à ceux de $f(y)$ si les M premiers chiffres de y sont identiques aux M premiers chiffres de x.

2) Plus généralement, soit $x_0 \in \mathbf{R}$ fixé. On dit que f est *continue* en x_0 si pour tout $N \geq 1$, il existe $M \geq 1$, qui peut dépendre de x_0, tel que les N premiers chiffres de $f(x)$ sont identiques à ceux de $f(x_0)$ lorsque les M premiers chiffres de x sont identiques à ceux de x_0. On dit que f est continue si elle est continue en x_0 pour tout x_0.

On étend également ces définitions aux fonctions définies seulement sur un sous-ensemble de $\mathbf{R}$.

Les entiers M qui apparaissent peuvent être compris comme la précision nécessaire à avoir pour un argument x de sorte que l'on puisse calculer exactement les N premiers chiffres de $f(x)$. Pour une fonction uniformément continue, cette précision ne dépend pas de x, mais pour une fonction simplement continue, elle peut en dépendre.

Exercice. Montrer que la fonction $f(x) = x^2$ est continue sur $\mathbf{R}$ mais pas uniformément continue.

Exercice. Pour $x = 0, d_1 \ldots d_n \ldots$ on pose

$$f_1(x) = 0, d_2 d_1 d_4 d_3 \ldots d_{2n+1} d_{2n} \ldots$$

Est-ce que $f_1 : [0, 1] \to \mathbf{R}$ est continue ? Si oui, déterminer la précision M nécessaire pour calculer $f(x)$ avec une précision de N chiffres.

On pose

$$f_2(x) = 0, d_2 d_4 d_8 \ldots d_{2^n} \ldots$$

Est-ce que $f_2 : [0, 1] \to \mathbf{R}$ est continue ?

On pose

$$f_3(x) = \begin{cases} 0 & \text{si } x \text{ est rationnel} \\ 1 & \text{sinon.} \end{cases}$$

Est-ce que f_3 est continue ?

Remarque. Un résultat important dit que toute fonction continue sur un intervalle fermé borné $[a, b]$ est uniformément continue. Pouvez-vous essayer de comprendre pourquoi ? (Ou le démontrer !)

On a maintenant le résultat très utile suivant.

Théorème. Soit f une fonction continue en $x \in \mathbf{R}$. Pour toute suite (x_n) telle que $\lim x_n$ existe et vaut x, la suite $(f(x_n))$ converge et

$$\lim f(x_n) = f(x).$$

Démonstration. Soit $N \geq 1$. On veut montrer que $(f(x_n))$ converge vers $f(x)$ avec une précision de N chiffres. Mais par continuité de f en x, il existe $M \geq 1$ tel que les N premiers chiffres de $f(x_n)$ sont ceux de $f(x)$ si les M premiers chiffres de x_n sont identiques à ceux de x. D'après la convergence de x_n vers x avec M chiffres, ceci sera le cas pour tout n à partir d'un certain rang, d'où le résultat.

Voici une application classique des fonctions continues.

Corollaire. Soit f une fonction continue et $x_0 \in \mathbf{R}$. On pose par récurrence

$$x_{n+1} = f(x_n)$$

pour $n \geq 0$. Alors soit la suite (x_n) diverge, soit la limite $\lim x_n = x$ vérifie l'équation

$$f(x) = x.$$

Démonstration. Si la suite (x_n) diverge, c'est fini. Supposons donc que $\lim x_n$ existe. Posons $y_n = f(x_n)$. D'après la continuité de f et la convergence de (x_n), le théorème précédent s'applique, donc la suite (y_n) converge également et $\lim y_n = f(x)$. Mais on peut remarquer que $y_n = x_{n+1}$, autrement dit la suite (y_n), à la numérotation près, est identique à la suite (x_n). Les limites de (x_n) et (y_n) sont donc forcément identiques et

$$x = \lim x_n = \lim y_n = f(x).$$

Exercice. 1) Soit $f(x) = 1/x$ pour $x \neq 0$. On pose $x_0 = 2$ et $x_{n+1} = f(x_n)$. Montrer que (x_n) diverge.

2) Soit $f(x) = x/2 + 1/x$ pour $x > 0$ et $x_0 = 2$. On pose $x_{n+1} = f(x_n)$. Montrer que $x_n \geq 0$ pour tout n, que x_n est décroissante, et que la suite (x_n) est convergente et

$$\lim x_n = \sqrt{2}$$

(c'est à dire que $(\lim x_n)^2 = 2$).

Exercice. On peut interpréter la notion de continuité à l'aide simplement de celle de limite de fonction. Montrer que f est continue sur $\mathbf{R}$ si et seulement si, pour tout x_0, la limite de $f(x)$ quand $x \to x_0$ existe et vaut $f(x)$, c'est à dire

$$\lim_{x \to x_0} f(x) = f(x_0).$$

D'après le critère permettant de ramener les limites de fonctions aux limites de suites (voir un exercice ci-dessus), cela signifie que f est continue si et seulement si, pour tout x_0 et toute suite (x_n) telle

que (x_n) converge vers x_0, la suite $(f(x_n))$ admet une limite et celle-ci vaut $f(x_0)$. Autrement dit, la réciproque du théorème ci-dessus est valide.

Exercice. 1) Montrer qu'une suite (x_n) converge vers a peut s'écrire de la manière suivante : pour tout $\epsilon > 0$, il existe $N \geq 1$ tel que $|x_n - a| < \epsilon$ pour tout $n \geq N$.

2) Soit ϵ_k une suite quelconque telle que (ϵ_k) converge vers 0. Montrer que la définition précédente peut aussi s'écrire ainsi : pour tout $k \geq 1$, il existe N tel que si $n \geq N$ on a $|x_n - a| \leq \epsilon_k$.

(3) Montrer que f est uniformément continue si et seulement si pour tout $\epsilon > 0$ il existe $\delta > 0$ tel que $|x - y| < \delta$ implique $|f(x) - f(y)| < \epsilon$.

(4) Montrer que f est continue en x_0 si et seulement si pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que si $|x - x_0| < \delta$, alors $|f(x) - f(x_0)| < \epsilon$.

Exercice. Définir ce que signifie qu'une fonction $f(x, y, z \ldots)$ de plusieurs variables est continue. On pose $f(x, y) = x + y$. Montrer que f est continue.

L'infini. En plus des limites définies dans la section précédente, il est souvent utilise de définir ce que signifie, pour une suite ou pour une fonction, de "tendre vers l'infini". On note cet infini par le symbole ∞, parfois $+\infty$ ou $-\infty$ s'il y a un signe bien déterminé. Il s'agit simplement de dire que la suite (ou la fonction) devient "de plus en plus grande".

Définition. 1) Soit (x_n) une suite de nombres réels. On dit que (x_n) tend vers $+\infty$ ou converge vers $+\infty$ si, pour tout $N \geq 1$, il existe N_1 tel que $x_n \geq N$ pour tout $n \geq N_1$. On note

$$\lim x_n = +\infty.$$

2) Soit (x_n) une suite de nombres réels. On dit que (x_n) tend vers $-\infty$ ou converge vers $-\infty$ si, pour tout $N \geq 1$, il existe N_1 tel que $x_n \leq -N$ pour tout $n \geq N_1$. On note

$$\lim x_n = -\infty.$$

3) Soit f une fonction définie sur $\mathbf{R}$, sauf en x_0, à valeurs réelles. On dit que f tend vers $+\infty$ quand $x \to x_0$ si, pour tout $N \geq 1$, il existe N_1 tel que $f(x) \geq N$ pour tout x tel que les N_1 premiers chiffres de x et de x_0 coïncident. On note

$$\lim_{x \to x_0} f(x) = +\infty.$$

4) Soit f une fonction définie sur $\mathbf{R}$, sauf en x_0, à valeurs réelles. On dit que f tend vers $-\infty$ quand $x \to x_0$ si, pour tout $N \geq 1$, il existe N_1 tel que $f(x) \leq -N$ pour tout x tel que les N_1 premiers chiffres de x et de x_0 coïncident. On note

$$\lim_{x \to x_0} f(x) = -\infty.$$

5) Soit f une fonction définie sur $\mathbf{R}$. On dit que f admet une limite $\ell \in \mathbf{R}$ quand $x \to +\infty$ si pour toute précisions de N chiffres donnée il existe $N_1 \geq 1$ tel que les N premiers chiffres de $f(x)$ et ℓ coïncident lorsque $x \geq N_1$. On note

$$\lim_{x \to +\infty} f(x) = \ell.$$

6) Soit f une fonction définie sur $\mathbf{R}$. On dit que f tend vers $+\infty$ quand $x \to +\infty$ si pour tout $N \geq 1$, il existe $N_1 \geq 1$ tel $f(x) \geq N$ pour $x \geq N_1$. On note

$$\lim_{x \to +\infty} f(x) = +\infty.$$

Exercice. Définir ce que signifie : la fonction $f(x)$ définie sur $]-\infty, x_0[= \{x \mid x < x_0\}$ tend vers $-\infty$ lorsque $x \to x_0$ par valeurs $x < x_0$.

La notion d'infini, et les symboles $\pm\infty$, sont très pratiques car ils s'incorporent aisément aux propriétés déjà connues des limites "finies". Par exemple on a

Exercice. 1) Soit (x_n) une suite de nombres réels positifs. Montrer que la suite (x_n) converge vers 0 si et seulement si la suite $(1/x_n)$ converge vers $+\infty$.

2) Soit f une fonction définie pour $x \neq x_0$. Montrer que

$$\lim_{x \to x_0} f(x) = 0$$

si et seulement si

$$\lim_{x \to x_0} \frac{1}{|f(x)|} = +\infty.$$

Exercice. Montrer que pour tout $n \geq 1$ on a

$$\begin{aligned} \lim_{x \to +\infty} x^n &= +\infty, \\ \lim_{x \to 0, x>0} \frac{1}{x^n} &= +\infty \\ \lim_{x \to 0, x<0} \frac{1}{x^n} &= -+\infty. \end{aligned}$$

Exercice. Soit $a_n \geq 0$ pour tout n. On considère la série $\sum a_n$. Montrer que deux cas seulement sont possibles :

1) Si les sommes partielles

$$s_N = \sum_{n=1}^{N} a_n$$

tendent vers $+\infty$, la série diverge.

2) Si la suite des sommes partielles est majorée, alors la série converge.

On utilise souvent des tests de comparaison pour prouver que des limites existent.

Proposition. 1) Soient (u_n) et (v_n) des suites. On suppose que $u_n \geq v_n$ pour tout n et que (v_n) tend vers $+\infty$. Alors (u_n) tend vers $+\infty$.

2) Soient (u_n) et (v_n) des suites, $a \in \mathbf{R}$. On suppose que $|u_n - a| \leq v_n$ pour tout n et que (v_n) converge vers 0. Alors (u_n) converge vers a.

Exemple. On a $2^n \geq n$ pour $n \geq 1$, donc $\lim 2^n = +\infty$.

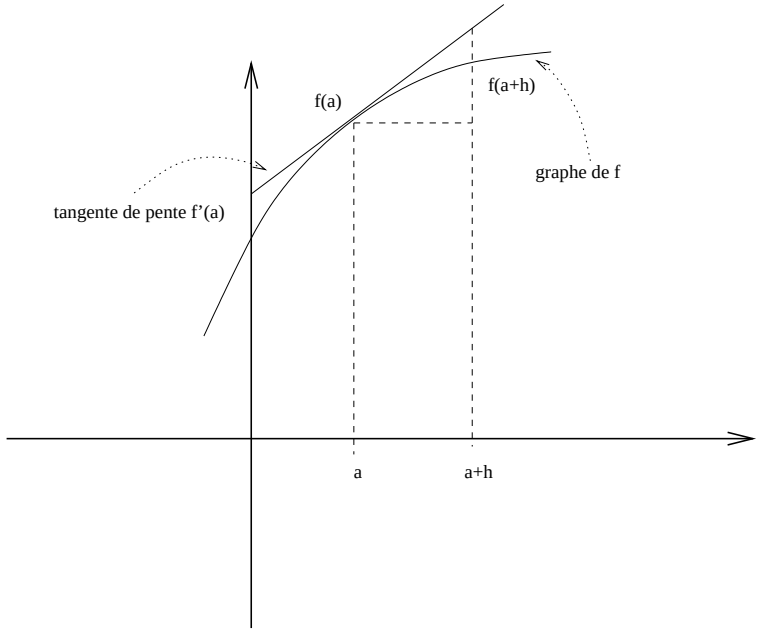
7.7 Continuité *vs.* dérivabilité.

La continuité d'une fonction f en un point ne signifie pas autre chose, que le fait que la fonction a pour limite en ce point la valeur de la fonction. En reprenant la définition en termes de distances, on voit donc que ça ne pose aucun problème que de définir la continuité d'une fonction de la variable complexe. On dira donc qu'une fonction $f : X \to \mathbf{C}$ est *continue* en $a \in X$ si

$$\lim_{x \to a} f(x) = f(a) \ .$$

Insistons sur deux points : ici il faut que a appartienne au domaine de f ; la définition est simplement une traduction de l'idée, explicitée dans les cas réel avec les degrés d'approximation des DDI, que "pour tout $r > 0$ la fonction f est *constante à r près* au voisinage de a".

Passons à la notion de dérivabilité en un point. Avec la courbe de Peano présentée dans la première partie, nous avons déjà vu que même si elle traduit une idée simple, la notion de continuité réserve des surprises. Il se trouve en fait que les fonctions les plus familières ne varient pas seulement de manière continue, mais d'une manière encore plus lisse. Leur graphe peut (localement, au voisinage d'un point) être approché par une droite. Dans le cas réel la situation est bien décrite par la figure usuelle, montrant la tangente en $f(a)$ comme la limite des sécantes de pente $(f(a+h) - f(a))/h$.



En symboles cela se traduit comme suit. Une fonction $f : X \to \mathbf{C}$ est *dérivable* en $a \in X$, s'il existe $\ell \in \mathbf{C}$ tel que

$$\lim_{x \to a} \frac{f(x) - f(a)}{x - a} = \ell \ .$$

Si f est dérivable en a on note $f'(a)$ la limite ℓ : c'est un nombre (peut-être complexe si f est à valeurs complexes). [2] Une formulation équivalente, qui nous rapproche de l'interprétation géométrique de la figure, est, que f est dérivable en a, si la limite

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

existe. Ici il faut que $a + h$ appartienne au domaine de f pour h petit, ce qui est garanti si f est définie sur une boule ouverte (ou intervalle) de centre a.

Montrons que la condition de dérivabilité est plus forte que celle de continuité. Comme nous l'avons déjà vu, il existe des fonctions partout continues et nulle part dérivables. La réciproque de la proposition qui suit est donc (très) fausse.

[2] A strictement parler il aurait fallu considérer la fonction $P_a(x) = (f(x) - f(a))/(x - a)$, définie sur $X \setminus \{a\}$, et s'intéresser à la limite $\lim_{x \to a} P_a(x)$.

Proposition. Soit $f : X \to \mathbf{C}$ une fonction définie en a. Si f dérivable en a, alors f est continue en a.

Démonstration. Donné un réel $r > 0$, nous devons trouver un r' tel que, si $|x - a| < r'$, alors $|f(x) - f(a)| < r$; ceci sous l'hypothèse que f est dérivable en a. Étudions donc la distance $|f(x) - f(a)|$: il faut la relier à la différence $(f(x) - f(a))/(x - a) - f'(a)$, qui contrôle la dérivabilité. Nous avons

$$\begin{aligned} |f(x) - f(a)| &= \left|\frac{f(x) - f(a)}{x - a}\right| |x - a| \\ &= |G(x)(x - a)| \quad \text{où on écrit } G(x) = (f(x) - f(a))/(x - a). \\ &= |(x - a)(G(x) - f'(a)) + f'(a)(x - a)| \\ &\leq |x - a||G(x) - f'(a)| + |f'(a)||x - a| \ , \end{aligned}$$

en utilisant l'inégalité du triangle (et l'astuce courante d'ajouter et enlever un même terme, ici le nombre $f'(a)$). Nous voyons donc que, si nous arrivons à trouver une condition sur x garantissant des majorations de ces derniers termes, alors nous aurons un majoration de $|f(x) - f(a)|$. Il faudra de plus que cette majoration donne l'inégalité $< r$ souhaitée. Or, vu que f est dérivable en a, nous savons que pour x assez proche de a, nous pouvons rendre $G(x) = P_a(x)$ aussi proche que nous voulons de $f'(a)$. Disons qu'il existe r'' tel que $|x - a| < r''$ implique $|G(x) - f'(a)| < r/2$. Alors on voit qu'en choisissant $r' = \min(1, r'', r/2|f'(a)|)$, on a bien $|f(x) - f(a)| \leq 1 \cdot r/2 + r/2 = r$.

7.8 Continuité et dérivabilité de $x^{[n]} = x^n/n!$.

On peut définir *toutes* les fonctions élémentaires à partir des fonctions "élévation à une puissance" $x \mapsto x^n$, à l'aide d'opérations algébriques (sommes et produits finis) et de passages à la limite (sommes infinies). Nous verrons plus loin comment les limites se comportent par rapport aux opérations algébriques.

Ici nous montrons directement à partir des définitions, que les fonctions de base, en quelque sorte les briques élémentaires, sont des fonctions partout continues et partout dérivables. Vu le résultat du paragraphe précédent, nous pourrions nous restreindre à une démonstration de la dérivabilité, mais ce ne serait pas très naturel.

Soit n un entier naturel. En fait, nous allons considérer la fonction *n-ième puissance divisée*

$$\begin{array}{rccl} f : \mathbf{C} & \to & \mathbf{C} \\ x & \mapsto & x^{[n]} := \frac{x^n}{n!} \ . \end{array}$$

Ceci car la démonstration est entièrement basée sur la formule du binôme, qu'il est pratique d'écrire sous la forme sans dénominateurs

$$(a + h)^{[n]} = \sum_{p+q=n} a^{[p]} h^{[q]} \ .$$

Exercice. Vérifier que cette formule est équivalente à la formule du binôme usuelle démontrée dans la Partie I.

Continuité. Soit $a \in \mathbf{C}$. Pour une tolérance r donnée il faut trouver r' tel que $|x - a| < r' \Rightarrow |f(x) - f(a)| < r$, c'est-à-dire

$$|x^{[n]} - a^{[n]}| < r \ . \quad (*)$$

On pose $h = x - a$, c'est-à-dire $x = a + h$, et $(*)$ devient

$$|(a+h)^{[n]} - a^{[n]}| < r \ .$$

La formule du binôme nous permet de mettre en évidence un facteur $|h|$:

$$\begin{aligned} |(a+h)^{[n]} - a^{[n]}| &= |a^{[n-1]}h + a^{[n-2]}h^{[2]} + \cdots + h^{[n]}| \\ &= |h||a^{[n-1]} + a^{[n-2]}\frac{h}{2!} + \cdots + \frac{h^{[n-1]}}{n!}| \ . \end{aligned}$$

L'espoir est que pour h petit cette façon d'écrire permet de montrer que $|(a+h)^{[n]} - a^{[n]}|$ est petit. Ce sera le cas si nous arrivons à majorer $|a^{[n-1]} + \cdots + h^{[n-1]}/n!|$ de manière indépendante de h. En appliquant l'inégalité du triangle on trouve que

$$|(a+h)^{[n]} - a^{[n]}| \leq |h|(|a^{[n-1]}| + |a^{[n-2]}||\frac{h}{2!}| + \cdots + |\frac{h^{[n-1]}}{n!}|) \ .$$

Le deuxième terme du membre de droite ressemble à $(|a| + |h|)^{[n-1]}$, la différence étant que les dénominateurs des facteurs ne sont pas les bons : on a $(k+1)!$ où on devrait avoir $k!$. Mais $(k+1)! > k!$, donc,

$$\begin{aligned} |(a+h)^{[n]} - a^{[n]}| &\leq |h|(|a| + |h|)^{[n-1]} \ \text{ et si } |h| < 1 \\ &\leq |h|(|a|+1)^{[n-1]} \ . \end{aligned}$$

De ce calcul on tire, que si $r'' = r/(|a|+1)^{[n-1]}$ et $r' = \min(1, r'')$, alors

$$|h| < r' \Rightarrow |(a+h)^{[n]} - a^{[n]}| < r \ .$$

Dérivabilité. On garde les mêmes notations et on s'intéresse à

$$\frac{f(a+h) - f(a)}{h} = a^{[n-1]} + h(\frac{a^{[n-2]}}{2!} + \cdots + \frac{h^{[n-2]}}{n!}) \ .$$

On se dit que pour h tendant vers 0 cette expression tend vers $a^{[n-1]}$. Donc on essaie de montrer que

$$(a^{[n]})' = a^{[n-1]} \ .^{3}$$

On procède comme ci-dessus : la formule du binôme donne

$$\begin{aligned} |\frac{f(a+h) - f(a)}{h} - a^{[n-1]}| &= |h|^{-1}|f(a+h) - f(a) - ha^{[n-1]}| \\ &= |h|^{-1}|h|^2|\frac{a^{[n-2]}}{2!} + \cdots + \frac{h^{[n-2]}}{n!}| \ . \end{aligned}$$

En appliquant l'inégalité du triangle et en utilisant le fait que $k! \geq 2!(k-2)!$, on obtient que cette expression est $\leq |h|(|a| + |h|)^{[n-2]}$, et on termine comme tout à l'heure.

Exercice. Compléter la démonstration, que nous venons d'esquisser.

[3] Cette formule est équivalente à la formule usuelle pour la dérivée de x^n : il suffit de multiplier par les constantes appropriées. Ainsi, si $(x^{[n]})' = x^{[n-1]}$, alors $(x^n)' = n!(x^{[n]})' = n!x^{[n-1]} = nx^{n-1}$.

7.9 Autres notions de limite.

Nous avons déjà dit que la notion de limite doit être maniée avec précaution et qu'elle n'est pas seulement formelle. Une illustration de ceci est qu'il est possible de définir les limites d'autres façons, avec parfois des "réponses" différentes. Par exemple, si (x_n) est une suite, on dit qu'elle *converge en moyenne (de Césaro)* vers un nombre réel x si la suite y_n définie par

$$y_n = \frac{x_1 + \cdots + x_n}{n}$$

converge elle-même vers x (au sens de la définition du cours). Autrement dit, on ne regarde pas les termes individuellement, mais seulement en moyenne.

On montre que si $\lim x_n$ existe au sens "usuel", alors (x_n) converge aussi en moyenne vers $\lim x_n$. Par contre si on prend

$$x_n = 1 - 1 + \cdots + (-1)^{n-1}$$

pour $n \geq 1$, on a $x_{2n} = 0$ et $x_{2n-1} = 1$ donc (x_n) diverge, mais on calcule que

$$y_{2n} = \frac{n}{2n} = \frac{1}{2}$$

et

$$y_{2n-1} = \frac{n-1}{2n} = \frac{1}{2} - \frac{1}{2n}.$$

Comme $\lim 1/n = 0$, on en déduit que $\lim y_n = 1/2$. Donc (x_n) converge en moyenne vers $1/2$.

Par rapport à la limite ordinaire, la limite au sens de Césaro est, en quelque sorte, plus tolérante de certaines oscillations de la suite. Elle peut "détecter" certaines régularités même dans le comportement d'une suite classiquement divergente.

Il peut être très utile d'utiliser une telle notion de limite par exemple parce que cela permet de "construire" des nombres réels vérifiant une certaine équation. Parfois aussi, même si la suite originale converge déjà, les moyennes de Césaro peuvent converger "plus vite" : pour des questions de calcul numérique, on comprend que cela peut être crucial.

Chapitre 8

Opérations sur les limites. Propriétés de la dérivation.

Nous avons défini ce que signifie pour une fonction d'admettre une limite en un point. Il n'est pas trop difficile de montrer que si l est un nombre tel que pour $f : X \to \mathbf{C}$

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

alors l est l'*unique* nombre à avoir cette propriété.

On commence par étudier comment se comportent les limites de fonctions par rapport à différentes opérations sur les fonctions.

Soit f et g des fonctions de domaine X à valeurs dans $\mathbf{C}$. Soit a un élément de X et soit l et m des éléments de $\mathbf{C}$. On suppose que

$$\lim_{x \to a} f(x) = l \quad \text{et que} \quad \lim_{x \to a} g(x) = m \ .$$

Alors :

i) $\lim_{x \to a}(f+g)(x) = l + m$

ii) $\lim_{x \to a}(fg)(x) = lm$

iii) Si $m \neq 0$, alors il existe r' tel que $g(x) \neq 0$ pour tout x avec $|x - a| < r'$.

iv) Si $m \neq 0$, alors $\lim_{x \to a}(1/g)(x) = 1/m$

Avant de nous lancer dans la preuve de ces affirmations remarquons les inégalités

$$|x| - |y| \leq |x - y| \quad \text{et} \quad |y| - |x| \leq |x - y| \qquad (*) \ .$$

L'affirmation (i) est une conséquence immédiate de l'inégalité du triangle : on écrit

$$|(f+g)(x) - (l+m)| = |(f(x) - l) + (g(x) - m)| \leq |f(x) - l| + |g(x) - m|$$

et donné $r > 0$ on cherche r' tel que à la fois $|f(x) - l|$ et $|g(x) - m|$ soient strictement inférieurs à $r/2$.

La démonstration de l'affirmation (ii) est basée sur l'identité

$$(fg)(x) - lm = f(x)(g(x) - m) + m(f(x) - l) \ ,$$

qui est obtenue en rajoutant et en retranchant $f(x)m$ au membre de gauche. Cette identité avec l'inégalité du triangle donne

$$|(fg)(x) - lm| \leq |f(x)||(g(x) - m)| + |m||(f(x) - l)| \ .$$

Soit $r > 0$ fixé. Vu que f admet pour limite l quand x tend vers a on peut certainement trouver $r_1^{'}$ tel que si $|x-a| < r_1^{'}$, alors

$$|f(x) - l| < 1 \quad \text{ou encore (par (*))} \quad |f(x)| < (1 + |l|) \ .$$

De même on peut trouver $r_2^{'}$ tel que si $|x-a| < r_2^{'}$, alors

$$|g(x) - m| < \frac{r}{2}\frac{1}{(1+|l|)} \ .$$

Donc pour $r_3^{'} = \min\{r_1^{'}, r_2^{'}\}$, si $|x-a| < r_3^{'}$, alors

$$|f(x)||(g(x) - m)| < r/2 \ .$$

On peut aussi trouver $r_4^{'}$ tel que si $|x-a| < r_4^{'}$, alors

$$|f(x) - l| < \frac{r}{2}\frac{1}{|m|} \ ,$$

et par conséquent, avec $r^{'} = \min\{r_3^{'}, r_4^{'}\}$, si $|x-a| < r^{'}$, alors on a bien que

$$|(fg)(x) - lm| < r \ ,$$

ce qui démontre le (ii).

Pour montrer le (iii), soit $r = m/2$, alors par l'hypothèse sur g et (*) il existe r' tel que, si $|x-a| < r'$, alors

$$|m| - |g(x)| \leq |g(x) - m| \leq \frac{|m|}{2} \quad \text{d'où} \quad 0 < \frac{|m|}{2} \leq |g(x)|$$

comme il fallait le démontrer.

Pour le (iv) choisissons à nouveau r' tel que $|x-a| < r'$ implique $0 < \frac{|m|}{2} \leq |g(x)|$. Alors

$$\frac{1}{|g(x)|} < \frac{2}{|m|}$$

et

$$\left|\frac{1}{g(x)} - \frac{1}{m}\right| = \left|\frac{m - g(x)}{mg(x)}\right| = \frac{|g(x) - m|}{|m||g(x)|} < \frac{2}{|m|}\frac{1}{|m|}|g(x) - m| \ .$$

Or, pour x assez proche de a on peut assurer $|g(x) - m| < |m|^2 r/2$, d'où le résultat.

Continuité. On déduit de ce qui précède, que si les fonctions f et g sont continues en un point a de leur domaine commun, alors les fonctions $f + g$ et fg le sont aussi. De plus, si $g(a) \neq 0$, alors il en est de même de la fonction f/g.

Exercice. Montrer que toute fonction polynôme $f(x) = a_0 + a_1 x + \cdots + a_k x^k$, où les a_i sont fixés, est continue.

Exercice. Étudier la continuité de la composition de deux fonctions continues.

Dérivabilité. On s'intéresse maintenant aux énoncés analogues au sujet de fonctions dérivables. Soit donc f et g des fonctions de domaine X et à valeurs complexes. Supposons que f et g soient dérivables en a et notons $f'(a)$ et $g'(a)$ les dérivées. Alors les fonctions $f + g$ et fg sont aussi dérivables en a. De plus si $g(a) \neq 0$ il en est de même de f/g. Les valeurs des dérivées sont les suivantes :

i) $(f + g)'(a) = f'(a) + g'(a)$

ii) $(fg)'(a) = f'(a)g(a) + f(a)g'(a)$
iii)

$$(\frac{f}{g})'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{g^2(a)}$$

La démonstration de l'énoncé concernant la fonction somme $f+g$ est laissée en exercice. Pour le produit, soit $h = fg$ et remarquons l'égalité

$$h(x) - h(a) = f(x)(g(x) - g(a)) + g(a)(f(x) - f(a)) \ ,$$

qui est obtenue en ajoutant et en retranchant $f(x)g(a)$ au membre de gauche. On obtient l'énoncé en divisant cette égalité par $x - a$ et en prenant la valeur absolue. Noter que, comme montré plus haut, vu que f est supposée dérivable en a elle est continue en a, donc pour x assez proche de a le terme $f(x)$ sera arbitrairement proche de $f(a)$.

La démonstration de l'assertion sur la fonction quotient $k = f/g$ est analogue et repose sur l'identité :

$$\frac{k(x) - k(a)}{x - a} = \frac{1}{g(x)g(a)} \left(g(a)\frac{f(x) - f(a)}{x - a} - f(a)\frac{g(x) - g(a)}{x - a} \right) \ .$$

Un résultat de base sur la dérivabilité est la règle de *dérivation d'une fonction composée*. Soit $f : X \to \mathbf{C}$ une fonction dérivable en un point a de X. Soit Y un sous-ensemble de $\mathbf{C}$, qui contient l'image de f et soit $g : Y \to \mathbf{C}$ une fonction dérivable en $b = f(x)$. Alors la fonction composée $h = g \circ f : X \to \mathbf{C}$ est dérivable en a et l'on a

$$(g \circ f)'(a) = g'(f(a))f'(a) \ .$$

Pour la démonstration de ces affirmations nous allons d'abord traduire la propriété de dérivabilité.

Affirmation. Soit $f : X \to \mathbf{C}$ une fonction et soit a élément de X. Sont équivalents :
i) f est dérivable en a.
ii) Il existe un nombre, noté $f'(a)$, et une fonction $r_a : X \to \mathbf{C}$ *continue en* a, tels que $r_a(a) = 0$ et tels que pour x dans X

$$f(x) = f(a) + f'(a)(x - a) + r_a(x)(x - a) \ .$$

iii) Il existe une fonction $\phi_a : X \to \mathbf{C}$ *continue en* a telle que pour x dans X

$$f(x) = f(a) + \phi_a(x)(x - a) \ .$$

La démonstration de ces équivalences n'est pas trop difficile : pour voir que (i) et (ii) sont équivalents, on définit r_a par l'égalité de (ii) et on interprète la définition de la dérivabilité en termes de la continuité de r_a. Pour voir que (ii) et (iii) sont équivalents on pose ϕ_a égal à $f'(a) + r_a(x)$. Ainsi $\phi_a(a) = f'(a)$.

Revenons à la dérivabilité de la composition de deux fonctions. Nous allons travailler avec la caractérisation (iii) de l'affirmation. Par hypothèse on a :

$$f(x) - f(a) = \phi_a(x)(x - a) \quad \text{et} \quad g(y) - g(b) = \psi_b(y)(y - b)$$

pour une certaine fonction ϕ_a (resp. ψ_b), qui est continue en a (resp. b) et est telle que $\phi_a(a) = f'(a)$ (resp. $\psi_b(b) = g'(b)$). Posons $y = f(x)$ et $b = f(a)$ dans la deuxième égalité et puis utilisons la première. Il vient :

$$\begin{aligned} g(f(x)) - g(f(a)) &= \psi_b(f(x))(f(x) - f(a)) \\ &= \psi_b(f(x))\phi_a(x)(x - a) \ . \end{aligned}$$

Or par les hypothèses $\psi_b(f(x))\phi_a(x)$ est continue en a et on déduit le résultat de ce que $\psi_b(f(a))\phi_a(a) = g'(f(a))f'(a)$.

On peut procéder de manière analogue pour traiter la *dérivabilité de la fonction réciproque* d'une fonction (disons bijective). Soit X et Y des sous-ensembles de $\mathbf{C}$ et soit $f : X \to Y$ une fonction bijective et dérivable en a, élément de X. Supposons que $f'(a) \neq 0$. Alors la fonction réciproque $f^{-1} : Y \to X$ est dérivable en $b = f(a)$ et

$$(f^{-1})'(b) = \frac{1}{f'(a)} \ .$$

Écrivons à nouveau $f(x) - f(a) = \phi_a(x)(x - a)$ avec ϕ_a continue en a et telle que $\phi_a(a) = f'(a)$. Alors si $y = f(x)$ et $b = f(a)$ on a

$$y - b = \phi_a(f^{-1}(y))((f^{-1}(y)) - (f^{-1}(b)) \ .$$

Comme par hypothèse $f'(a) = \phi_a(a)$ est non-nul on a par la continuité de ϕ_a, que pour x assez proche de a le terme $\phi_a(f^{-1}(y))$ est non-nul. On peut donc considérer son inverse et on obtient le résultat voulu.

Chapitre 9

L'exponentielle et d'autres fonctions élémentaires.

Nous définissons la valeur en un nombre complexe z de la fonction exponentielle par

$$\exp(z) = \sum_{n=0}^{\infty} z^{[n]} = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \cdots \ .$$

Voir le paragraphe sur les suites numériques pour une (idée de) démonstration du fait que cette série converge. Nous allons détailler ici les démonstrations de quelques propriétés de base de cette fonction et nous verrons comment l'étude de la restriction de cette fonction à l'axe $\mathbf{R}$ des réels mène à une définition de la fonction logarithme (réelle) et des fonctions puissance ; de même l'étude de la restriction de l'exponentielle à l'axe $i\mathbf{R}$ des imaginaires purs mène à une définition des fonctions trigonométriques.

Les démonstrations seront essentiellement complètes à cela près que par endroits nous allons manipuler la série définissant l'exponentielle comme s'il s'agissait d'une somme finie. Cela est légitime mais nécessiterait une justification, qui ferait appel aux propriétés des séries dites absolument convergentes, dont la série exponentielle est un exemple. Nous allons indiquer les endroits nécessitant une justification ultérieure par le signe !. Noter que les démonstrations où vont apparaître des ! fournissent en tout cas des énoncés sur les valeurs approchées de l'exponentielle.

Additivité. Soit z et w des nombres complexes. Alors

$$\exp(z + w) = \exp(z)\exp(w) \ .$$

En particulier

$$\exp(z)\exp(-z) = \exp(0) = 1 \ .$$

Avant de montrer l'additivité faisons la remarque, que par exemple le produit de deux sommes finies comme $a_0 + a_1 + a_2$ et $b_0 + b_1 + b_2$ comporte 9 termes, qui peuvent être regroupés comme suit

$$\begin{aligned}(a_0 + a_1 + a_2)(b_0 + b_1 + b_2) \quad &= \quad a_0b_0 + (a_0b_1 + a_1b_0) + (a_0b_2 + a_1b_1 + a_2b_0) \\ &\quad +(a_1b_2 + a_2b_1) + a_2b_2 \ .\end{aligned}$$

Ici on a regroupé dans une même parenthèse les produits dont la somme des indices est constante : ainsi $0 = 0 + 0$, $1 = 0 + 1 = 1 + 0$, etc. Pour démontrer l'additivité nous allons faire comme si cette

manière de réordonner les termes d'un produit de sommes pouvait s'appliquer aux séries. On écrit

$$\begin{aligned}\exp(z)\exp(w) &= (1+z+z^{[2]}+z^{[3]}+\cdots)(1+w+w^{[2]}+w^{[3]}+\cdots)\\ &= (\sum_{p=0}^{\infty} z^{[p]})(\sum_{q=0}^{\infty} w^{[q]})\\ &\stackrel{!}{=} \sum_{n=0}^{\infty}(\sum_{p+q=n} z^{[p]}w^{[q]})\\ &= \sum_{n=0}^{\infty}(z+w)^{[n]}\\ &= \exp(z+w)\end{aligned}$$

où l'avant-dernière égalité est donnée par le théorème du binôme.

Continuité et dérivabilité. La fonction exponentielle

$$\begin{aligned}\exp : \mathbf{C} &\to \mathbf{C}\\ z &\mapsto \exp(z)\end{aligned}$$

est continue et dérivable en tout point de $\mathbf{C}$. De plus :

$$\exp'(z) = \exp(z)\ .$$

On démontre la dérivabilité comme suit.

$$\begin{aligned}\lim_{h\to 0}\frac{\exp(z+h)-\exp(z)}{h} &= \exp(z)\lim_{h\to 0}\frac{\exp(h)-1}{h}\\ &= \exp(z)\cdot 1\end{aligned}$$

où la première égalité suit de l'additivité et la seconde directement de la définition de l'exponentielle.

Motivation. Le fait que la fonction exponentielle soit égale à sa propre dérivée est la propriété qui la rend incontournable dans les applications (à la physique, en biologie, *etc.*). C'est-à-dire que la traduction mathématique de bon nombre de problèmes se ramène à démontrer l'existence d'une fonction f égale à sa dérivée : $f = f'$. Or, *si* on cherche une telle fonction sous la forme

$$f(z) = \sum_n a_n z^n$$

et que l'on admet que la dérivée d'une telle fonction est donnée par la dérivation terme-à-terme (comme s'il s'agissait d'une somme finie)

$$f'(z) = \sum_n n a_n z^{n-1}\ ,$$

alors la seule possibilité pour avoir $f = f'$, c'est-à-dire $a_{n-1} = n a_n$, si on fixe par exemple $a_0 = 1$, est bien $a_n = 1/n!$.

Puissances du nombre e. On définit le nombre e par

$$e := \exp(1) = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \cdots$$

Il est clair d'après cette définition que e est un nombre réel positif, strictement plus grand que 2. Soit n un entier naturel. Par l'additivité on a

$$\exp(n) = \exp(1+\cdots+1) = \exp(1)\cdots\exp(1) = e^n\ .$$

De même, si $q = a/b$ est un nombre rationnel positif, avec a et b positifs, on voit que

$$(\exp(q))^b = e^a \quad \text{et} \quad \exp(q) = e^q \ .$$

On écrit souvent e^z au lieu de $\exp(z)$, mais il faut bien comprendre qu'en général e^z ne signifie pas "e multiplié avec lui-même z fois". On verra plus loin comment définir les puissances x^α.

On peut montrer (assez facilement) que e est irrationnel.

Valeurs de la restriction à l'axe réel. Résumons en un énoncé les propriétés de la fonction e^x obtenue en restreignant exp à l'axe réel. Observons que les valeurs de e^x sont *positives*. En effet, si $x > 0$, alors on voit sur la définition que e^x est positif (car limite de sommes à termes positifs, qui croissent). Pour $x < 0$ on utilise le résultat pour le nombre positif $-x$ et l'identité $e^x e^{-x} = 1$.

Théorème. *La fonction*

$$e^x : \mathbf{R} \to \mathbf{R}_{>0}$$

a les propriétés suivantes.

a) e^x *est continue et différentiable en tout point de* $\mathbf{R}$.
b) La dérivée de e^x *est donnée par* $(e^x)' = e^x$.
c) La fonction e^x *est strictement croissante.*
d) $e^{x+y} = e^x e^y$
e) $\lim_{x\to\infty} e^x = \infty$ *et* $\lim_{x\to-\infty} e^x = 0$
f) Pour tout entier naturel n

$$\lim_{x\to\infty} x^n e^{-x} = 0 \ .$$

Nous avons déjà montré les propriétés (a), (b) et (d). Le fait que e^x soit strictement croissante peut se voir sur la définition. Le fait que, pour x grand, e^x n'est pas bornée supérieurement suit du résultat pour les puissances entières positives de e. Avec l'égalité $e^x e^{-x} = 1$ ceci donne la valeur de la limite pour x tendant vers $-\infty$. Pour montrer (f) on remarque que par la définition, si $x > 0$, alors

$$e^x > \frac{x^{n+1}}{(n+1)!}$$

et que par conséquent $x^n e^{-x} < (n+1)!/x$.

La propriété (f) traduit le fait que e^x croît plus vite que n'importe quelle puissance de x (voir l'expression "*croissance exponentielle*").

La fonction logarithme. Nous allons énoncer un résultat général, qui appliqué à la fonction exponentielle e^x, montrera l'existence de la fonction logarithme et nous permettra de montrer les propriétés de base de cette fonction.

Théorème (des fonctions réciproques). *Soit* I *un intervalle dans* $\mathbf{R}$ *(*$I = \mathbf{R}$ *est permis). Soit* $f : I \to \mathbf{R}$ *une fonction, que l'on suppose* strictement monotone *(donc soit strictement croissante, soit strictement décroissante). Posons* $J = f(I)$. *Alors sont équivalents :*

a) f *est continue en tout point de* I.
b) J *est un intervalle.*

Si (a) et/ou (b) sont satisfaits, alors la relation g *réciproque de* f *est une fonction et* $g : J \to I$ *est continue et strictement monotone. De plus, si* f *est dérivable en tout point de* I, *alors* g *est dérivable en tout point de* J .

Nous n'allons pas démontrer ce théorème (qui serait pourtant à notre portée ; voir le Théorème des valeurs intermédiaires utilisé plus bas). On peut appliquer le théorème au cas de $f = e^x$, $I = \mathbf{R}$ et $J = \mathbf{R}_{>0}$ et on peut poser la définition suivante.

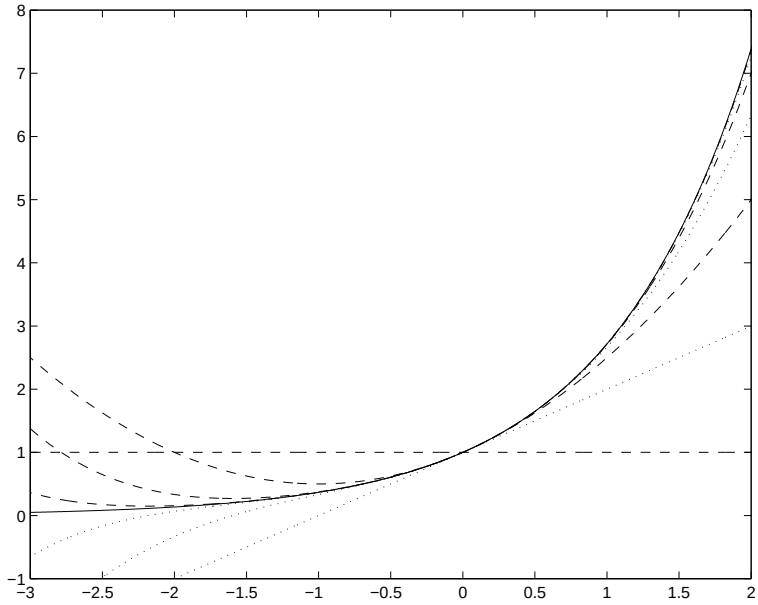


FIG. 9.1 – Exponentielle réelle et premières approximations polynomiales.

Définition. On appelle fonction *logarithme* (naturel ou népérien) la fonction

$$\log : \mathbf{R}_{>0} \to \mathbf{R}$$

réciproque de la fonction exponentielle.

La fonction logarithme est donc définie pour $x > 0$ par $\log(x) = u$ avec $x = \exp(u)$ c'est-à-dire, si u est réel et $x > 0$

$$\log(\exp(u)) = u \quad \text{et} \quad \exp(\log(x)) = x \ .$$

En utilisant l'additivité de l'exponentielle on a immédiatement la propriété caractéristique du logarithme : il transforme produits en sommes. C'est cette propriété qui l'a rendu indispensable pendant des siècles, pour permettre des calculs avec de grands nombres. On a pour x et y positifs

$$\log(xy) = \log(x) + \log(y)$$

(écrire $x = \exp(u)$ et $y = \exp(v)$ de manière à ce que

$$\log(xy) = \log(\exp(u)\exp(v)) = \log(\exp(u+v)) = u + v = \log(x) + \log(y) \ ).$$

Par les règles de dérivation d'une composée de fonctions on obtient

$$\log'(x) = \frac{1}{x}$$

(utiliser $\exp'(u) = \exp(u)$ et écrire $\log'(\exp(u))\exp'(u) = 1$ avec $x = \exp(u)$).

On peut montre que la fonction logarithme elle aussi admet une représentation comme série de fonctions. En effet pour $|x| < 1$ on a l'égalité

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \pm \cdots \ ,$$

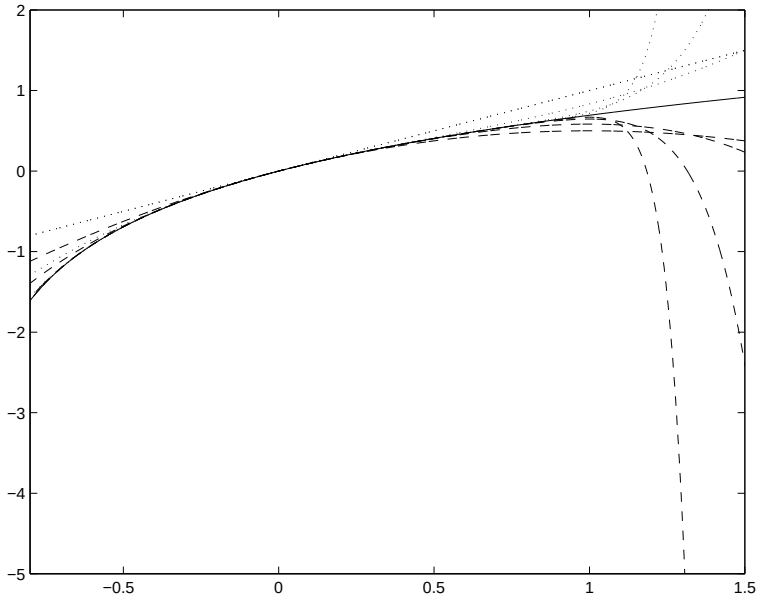


FIG. 9.2 – Logarithme et premières approximations polynomiales.

qui en fait est aussi valable pour $x = 1$ et donne la belle formule

$$\log 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} \pm \cdots \ .$$

Une manière de deviner ces formules est la suivante : la dérivée de $\log(1+x)$ est $1/(1+x)$; en utilisant par exemple la formule du binôme pour l'exposant négatif -1, on trouve

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 \pm \cdots \ .$$

Nous pouvons trouver une fonction, qui dérivée donne x^n, c'est $x^{n+1}/(n+1)$. Ainsi, si ici il était légitime d'interchanger dérivation et somme infinie on aurait bien la formule annoncée pour le logarithme.

Fonctions puissance. Pour α réel et x *positif* on définit

$$x^\alpha := \exp(\alpha \log x) \ .$$

On peut vérifier en utilisant ce qui précède, que cette définition donne bien le résultat voulu pour, par exemple, α rationnel. Par les propriétés des dérivées de fonctions composées on obtient

$$(x^\alpha)' = \exp'(\alpha \log x)(\alpha \log)'(x) = x^\alpha \cdot \frac{\alpha}{x} = \alpha x^{\alpha-1} \ .$$

En utilisant la propriété de la croissance exponentielle on a pour tout $\alpha > 0$

$$\lim_{x \to \infty} x^{-\alpha} \log x = 0 \ ,$$

c'est-à-dire que le logarithme $\log x$ croît moins vite que n'importe quelle puissance positive de x ($\log x$ n'est pas pour autant bornée).

Valeurs de la restriction à l'axe $i\mathbf{R}$. Fonctions trigonométriques. On s'intéresse ici aux nombres de la forme $\exp(iy)$ avec y réel. En supposant encore une fois que l'on peut opérer avec la série exponentielle comme avec une somme finie on obtient pour le conjugué de $\exp(z)$

$$\overline{\exp(z)} = \exp(\overline{z}) \ .$$

En effet

$$\begin{aligned} \overline{\exp(z)} &= \overline{1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \cdots} \\ &\overset{!}{=} 1 + \overline{z} + \frac{\overline{z}^2}{2} + \frac{\overline{z}^3}{6} + \cdots \\ &= \exp(\overline{z}) \ . \end{aligned}$$

De ceci on tire, pour y réel (et donc égal à son conjugué $\overline{y}$)

$$|\exp(iy)|^2 = \exp(iy)\overline{\exp(iy)} = \exp(iy)\exp(-iy) = 1 \ ,$$

et donc

$$|\exp(iy)| = 1 \ ,$$

c'est-à-dire que sur $i\mathbf{R}$ l'exponentielle prend ses valeurs dans le cercle unité.

Définition. Pour y réel on définit les *fonctions trigonométriques* cosinus et sinus par

$$\begin{aligned} \cos(y) &:= Re(\exp(iy)) = \frac{1}{2}(\exp(iy) + \exp(-iy)) \\ \sin(y) &:= Im(\exp(iy)) = \frac{1}{2i}(\exp(iy) - \exp(-iy)) \end{aligned}$$

Il n'est pas du tout clair sur cette définition que ce sont là les fonctions trigonométriques utilisées dans la mesure du cercle ! Mais nous allons nous en convaincre après un peu de travail.

En utilisant le fait que pour tout n entier naturel i^n ne prend que les quatre valeurs 1, i, -1 et $-i$ suivant le reste de la division de n par 4, on peut montrer que

$$\begin{aligned} \exp(iy) &= 1 + iy + (iy)^{[2]} + (iy)^{[3]} + (iy)^{[4]} \cdots \\ &= 1 + iy - y^{[2]} - iy^{[3]} + y^{[4]} \cdots \\ &\overset{!}{=} (1 - y^{[2]} + y^{[4]} - y^{[6]} \pm \cdots) + i(y - y^{[3]} + y^{[5]} \pm \cdots) \end{aligned}$$

Ce qui permet d'écrire

$$\begin{aligned} \cos y &:= 1 - y^{[2]} + y^{[4]} - y^{[6]} \pm \cdots \\ \sin y &:= y - y^{[3]} + y^{[5]} \pm \cdots \end{aligned}$$

Il est clair que cos et sin ainsi définies prennent leurs valeurs entre 1 et -1. Aussi la définition montre que

$$\cos(0) = 1 \quad \text{et} \quad \sin(0) = 0 \ .$$

Les fonctions cos et sin sont continues et dérivables en tout point et les règles de dérivation donnent

$$\cos'(y) = -\sin(y) \quad \text{et} \quad \sin'(y) = \cos(y) \ .$$

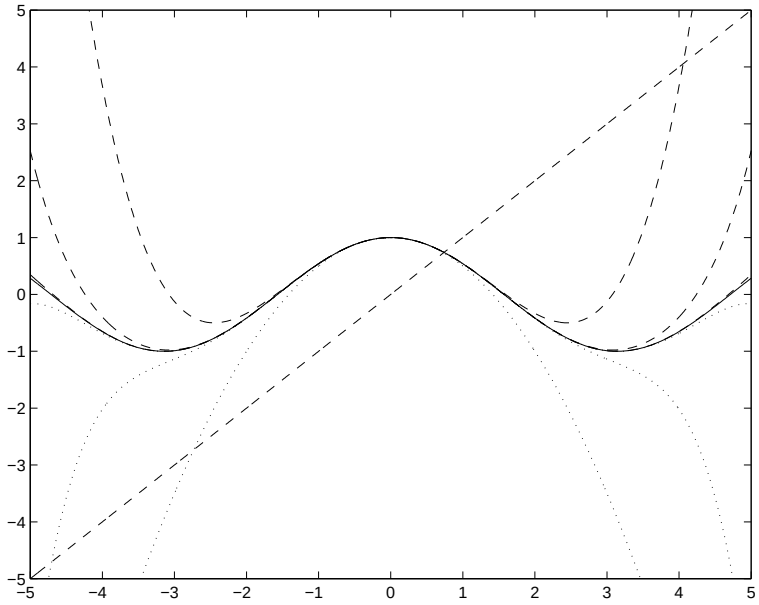


FIG. 9.3 – Cosinus et premières approximations polynomiales.

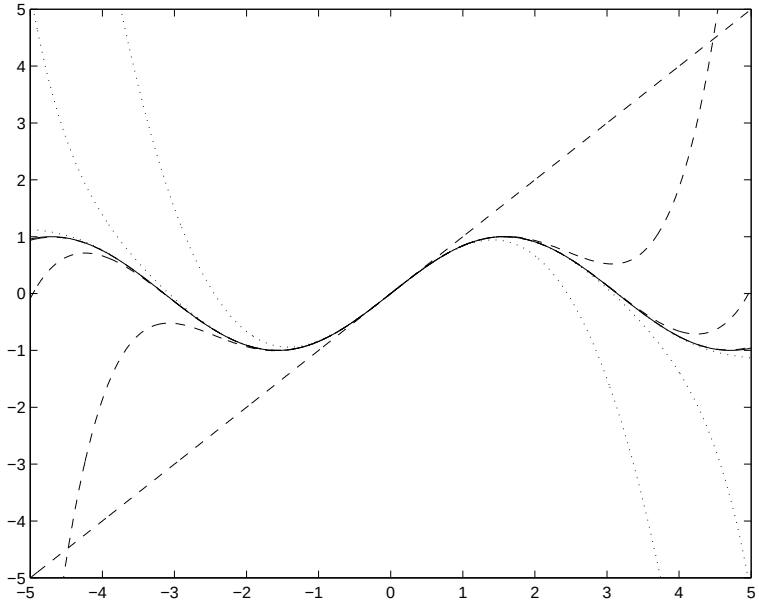


FIG. 9.4 – Sinus et premières approximations polynomiales.

C'est un bon début ! Il faudrait maintenant que l'on arrive à voir la périodicité de ces fonctions, mais pour cela il faudrait que l'on sache comment définir leur période, qui devrait être en principe 2π. Nous avons défini π comme étant la longueur du demi-cercle (unité). Ce que nous allons faire est de donner une

autre définition de π, puis, après avoir démontré les propriétés voulues des fonctions trigonométriques nous allons dire comment montrer que avec cette nouvelle définition on retrouve bien la longueur du demi-cercle.

Proposition-Définition. *Il existe un nombre réel y compris entre 0 et 2 pour lequel* $\cos(y) = 0$. *Vu que la fonction* cos *est continue en tout point, il existe un plus petit réel positif y_0 tel que* $\cos(y_0) = 0$. *On pose*

$$\pi = 2y_0 \ .$$

Noter que l'on a besoin de la continuité de cos pour avoir l'existence de y_0 : l'infimum de l'ensemble des y (positifs) annulant une fonction f n'annule pas forcément la fonction. Ici on utilise la continuité pour voir que toute limite de suites (a_n) formée de solutions de l'équation $f(x) = 0$ est encore une solution de cette équation.

Pour voir que cos s'annule (au moins une fois) sur l'intervalle $(0, 2)$ on peut procéder de plusieurs manières, aucune n'est vraiment élémentaire et toutes s'appuient sur des résultats que nous n'allons pas démontrer en détail.

Première manière. La première manière est basée sur des propriétés que possèdent les *fonctions dérivables sur un intervalle*. Plus précisément si a et b sont des réels avec $a < b$ et $f : [a, b] \to \mathbf{R}$ est une fonction continue sur l'intervalle fermé $[a, b]$ et dérivable en tout point de l'intervalle ouvert (a, b), alors il existe un point c dans (a, b) tel que

$$f(b) - f(a) = (b - a)f'(c) \ , \qquad (AC)$$

(c'est là une forme du *Théorème des accroissements finis*). En particulier, si pour tout c de (a, b) on a $f'(c) > 0$, alors la fonction f est strictement croissante.

Pour montrer la proposition à partir de (AC) supposons par l'absurde, que cos ne s'annule pas. Alors, vu que $\cos(0) = 1$ on aurait $\cos(y) > 0$ et donc $\sin'(y) > 0$. (AC) montre alors que $\sin(0) = 0$ entraîne $\sin(y) > 0$ pour $y > 0$. A nouveau par (AC) et vu que $\cos' = -\sin$, si $0 < y < x$, alors

$$\sin(y)(x - y) = \cos(y) - \cos(x) \leq 2 \ .$$

Ceci donne une contradiction pour x choisit suffisamment grand.

Deuxième manière. On déduit du développement en série de la fonction cos, que $1 - \cos y = y^2/2 - y^4/2 \cdot 3 \cdot 4 + y^6/2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 - \cdots$. On voit que pour y compris entre 0 et 3 on passe d'un terme à l'autre de cette série en multipliant par un facteur strictement inférieur à 1 : en effet alors $y^2 < 3 \cdot 4$. Ceci implique que la valeur de $1 - \cos y$ est comprise entre les deux premiers termes $y^2/2$ et $y^2/2 - y^4/2 \cdot 3 \cdot 4$, d'où pour $0 \leq y \leq 3$

$$1 - y^2/2 \leq \cos y \leq 1 - y^2/2 + y^4/24 \ .$$

En particulier

$$\cos 2 < -\frac{1}{3} < 0 \ .$$

On applique alors le *Théorème des valeurs intermédiaires* qui dit que si $f : [a, b] \to \mathbf{R}$ est une fonction *continue en tout point de l'intervalle* $[a, b]$ telle que $f(a) < 0$ et $f(b) > 0$, alors il existe un c dans l'intervalle (a, b) avec $f(c) = 0$. (Voici l'idée pour montrer le Théorème des valeurs intermédiaires : on considère l'ensemble E des x dans l'intervalle $[a, b]$ tels que $f(x') < 0$ *pour tout* x' avec $a \leq x' \leq x$. C'est un ensemble non-vide et majoré, il possède donc un supremum c. En utilisant la continuité de f sur l'intervalle on montre que $f(c)$ ne peut ni être < 0, ni être > 0.)

Périodicité. Par définition on a donc $\cos(\pi/2) = 0$ et vu que $\cos^2(y) + \sin^2(y) = 1$ on a $\sin^2(\pi/2) = 1$. Donc $\sin(\pi/2) = \pm 1$. Par le Théorème des accroissements finis, vu que $\cos y > 0$ pour y entre 0 et $\pi/2$, on a que sin croît sur l'intervalle $(0, \pi/2)$ et donc $\sin(\pi/2) = 1$. D'où

$$\exp(\frac{\pi i}{2}) = i \ ,$$

et par additivité

$$\exp(\pi i) = -1 \quad \text{et} \quad \exp(2\pi i) = 1 \ .$$

Ceci a la conséquence étonnante, si on pense aux propriétés de l'exponentielle réelle, que l'exponentielle complexe est périodique : pour tout z complexe

$$\exp(z + 2\pi i) = \exp(z) \ .$$

Théorème.

a) Les fonctions cos *et* sin *sont périodiques, de période* 2π.

b) Pour tout nombre complexe z de module 1*, il existe un unique y dans l'intervalle* $[0, 2\pi)$ *tel que* $\exp(iy) = z$.

Une manière pour traduire le (b) est de dire que l'exponentielle établit une bijection entre l'intervalle $[0, 2\pi)$ et le cercle unité. Ceci permet de définir l'*argument d'un nombre complexe z*, comme étant le réel $\arg(z)$, défini à un multiple de 2π près, tel que $z/|z| = \exp(i \arg(z))$ (il s'agit donc plutôt d'un ensemble de nombres).

Le (a) du théorème suit directement de la définition des fonctions cos et sin et de la périodicité de exp. Pour montrer (b), observons d'abord que si y est tel que $0 < y < 2\pi$, alors $\exp(iy) \neq 1$. En effet, si on écrit

$$\exp(iy) = a + ib \ ,$$

avec a et b réels, alors $a^2 + b^2 = 1$ et $0 < a, b < 1$. De plus si (par l'absurde) $\exp(iy) = 1$, alors par additivité $\exp(4iy) = 1$ et en particulier $\exp(4iy)$ est réel. Écrivons

$$\exp(4iy) = (a + ib)^4 = (a^4 - 6a^2b^2 + b^4) + 4iab(a^2 - b^2) \ .$$

Si $\exp(4iy)$ est réel, alors $a^2 - b^2 = 0$, d'où (avec $a^2 + b^2 = 1$) $a^2 = b^2 = 1/2$, ce qui donne $\exp(4iy) = -1$, une contradiction.

Montrons que ceci donne l'assertion d'unicité du (b). Soit y_1 et y_2 avec $0 \leq y_1, y_2 < 2\pi$. Alors effectivement

$$\exp(iy_1) \exp(iy_2)^{-1} = \exp(i(y_1 - y_2)) \neq 1$$

et donc $\exp(iy_1) \neq \exp(iy_2)$. Pour montrer l'affirmation d'existence, soit $z = a + ib$ un nombre complexe de module 1 (a et b réels). Supposons d'abord a et b positifs. Sur $[0, \pi/2]$ la fonction continue cos décroît de 1 à 0, donc par le Théorème des valeurs intermédiaires il existe y dans $[0, \pi/2]$ tel que $\cos(y) = a$. Vu que $\cos^2(y) + \sin^2(y) = 1$ et $\sin(y) \geq 0$ on a bien que $\exp(iy) = z$. Pour $a < 0$ et $b \geq 0$ on applique ce qui précède à $-iz$, d'où l'existence de y tel que $\exp(iy) = -iz$ et comme $i = \exp(\pi i/2)$ on a $z = \exp(i(y + \pi/2))$. De même si $b < 0$ en utilisant ce qui précède on voit que $-z = \exp(iy)$ pour y dans $(0, \pi)$ et donc $z = -\exp(iy) = \exp(i(y + \pi))$.

La longueur du cercle et π. Nous venons de voir que l'exponentielle définit une bijection de l'intervalle $J = [0, 2\pi)$ sur le cercle unité. Cette bijection est de plus dérivable. On peut montrer que la longueur des courbes du plan, définies par des fonctions dérivables $f : I \to \mathbf{C}$, avec I un intervalle, se calcule à l'aide d'une intégrale (définie). Dans le cas qui nous intéresse ici, il s'agit de l'intégrale

$$\int_0^{2\pi} 1 \ dt$$

qui vaut (heureusement) 2π (ici π est défini comme le plus petit réel positif annulant cos).

Chapitre 10

Intégrales : aires et primitives.

La théorie de l'intégration, ou de la mesure, même à notre niveau, permet de donner des réponses très générales à des problèmes importants tels que : calculer l'aire d'une portion du plan, ou déterminer les valeurs d'une fonction F à partir de son taux de variation f sur un intervalle (p. ex. f la vitesse d'un objet et F sa position). [1]

Dans le premier paragraphe nous allons nous inspirer de la méthode des approximations par des polygones, pour définir l'aire des figures planes ; l'aire est caractérisée comme une *fonction* à valeurs réelles non-négatives, ayant certaines propriétés attendues. De même, dans le deuxième paragraphe, nous associons un nombre à chacun des éléments d'une classe générale de fonctions définies sur un intervalle : l'intégrale définie de la fonction. Les deux paragraphes sont très liés. En effet l'intégrale peut s'interpréter en termes d'aires et cette interprétation permet de la voir comme une opération inverse de la dérivation. Réciproquement le calcul intégral permet de systématiser les calculs d'aire.

Un exemple simple permet de comprendre le lien entre les concept d'aire et de (anti-)dérivation.

Exemple. On considère un réservoir troué, qui contient un liquide, et on veut trouver la quantité de liquide qui s'est échappée depuis sa perforation. Voici une méthode pour résoudre ce problème : il s'agit de mesurer à distance régulière dans le temps–disons une heure–le débit (mesuré en litres/heures l/h). Disons que l'on obtient les valeurs suivantes (en mesurant combien de temps met *un* litre à s'échapper) :

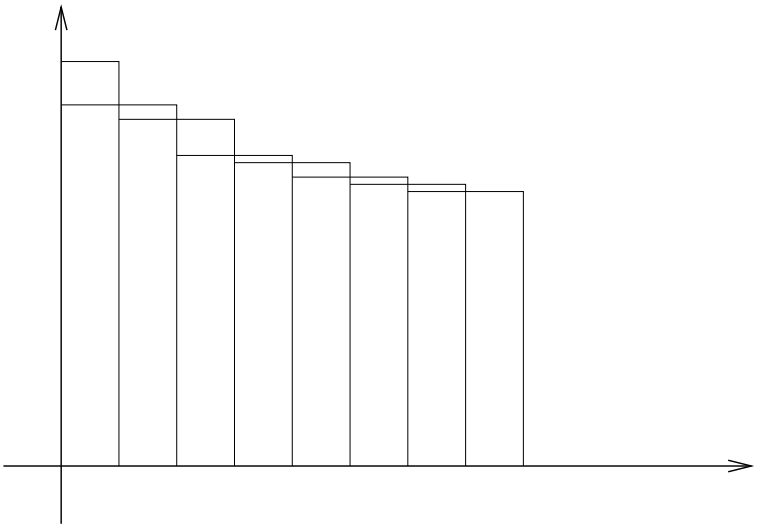
heures	0	1	2	3	4
l/h	35	30	26	23	21

La fonction à calculer, qui donne la quantité de liquide qui s'échappe, est décroissante. Ces données ne permettent pas de la calculer de façon exacte, mais en donnent des *approximations* : une majoration donnée par la somme $35 \cdot 1 + 30 \cdot 1 + 26 \cdot 1 + 23 \cdot 1 = 114$, et une minoration donnée par la somme $30 \cdot 1 + 26 \cdot 1 + 23 \cdot 1 + 21 \cdot 1 = 100$. On peut améliorer ces approximations en faisant plus de mesures. Disons que l'on obtient les données supplémentaires :

heures	1/2	3/2	5/2	7/2
l/h	33	27	24	22

(Évidemment on ne peut pas remonter le temps, il fallait y penser avant de commencer !) Avec ces mesures on obtient le majoration 110 et la minoration 103 (faire le calcul).
Si on devait représenter ces calculs sur un graphique on obtiendrait quelque chose comme sur la figure, où la majoration correspond à la somme des aires et des grands rectangles et la minoration à la somme des aires des petits rectangles. On voit donc qu'en augmentant indéfiniment les mesures on peut s'attendre

[1] Une telle fonction F sera dite *primitive* de f.



à ce que (théoriquement) : si $f(x)$ est la fonction cherchée ici, qui est positive, monotone et continue (pour des raisons physiques), et si $F(x)$ représente la fonction qui donne l'aire sous le graphe de f entre 0 et x, alors $F'(x) = f(x)$.

Nous n'allons pas donner une justification complète de cet énoncé dans le texte. Il nous importe de mettre en évidence quelques résultats sur lesquels ils se fonde et de le rendre plausible. Le premier tel énoncé est contenu dans l'exercice suivant.

Exercice. Soit E et F deux sous-ensembles de $\mathbf{R}$ tels que $\forall x \in E\ \forall y \in F\ :\ x \leq y$. Alors $\sup E$ et $\inf F$ existent et $\sup E \leq \inf F$. De plus on égalité si $\forall r > 0\ (\exists x \in E \wedge \exists y \in F)\ :\ y - x < r$.

10.1 L'aire des figures planes.

Il n'est pas facile de définir ce que l'on entend par l'aire d'*une* figure dans le plan. La notion d'aire est en quelque sorte primitive : elle vient avec celle de figure. On peut dire qu'il s'agit du contenu de la figure, qu'elle mesure sa taille, mais il est clair que ce ne sont pas là des définitions très utiles, bien qu'elles expliquent quelque chose de la notion. Euclide n'avait pas défini la notion d'aire, qu'il identifiait en fait à la figure elle-même. Cela ne l'a pas empêché pour autant de formuler des énoncés portant sur le *rapport* entre des aires. Ainsi la Proposition 2 du Livre XII, dit : *les cercles sont entre eux comme les carrés de leurs diamètres.* Il y a bien chez Euclide un début de calcul avec les aires, mais dans le contexte d'un calcul avec des grandeurs générales : il définit par exemple ce que sont des grandeurs en même raison[2]. Ce calcul n'est pas un calcul avec des nombres.

[2]Il s'agit de la Définition 5 du Livre X : *des grandeurs sont dites être en même raison, la première à la seconde, et la troisième à la quatrième, lorsque des équimultiples quelconques de la première et de la troisième, et d'autres équimultiples quelconques de la seconde et de la quatrième sont tels, que les premiers équimultiples surpassent, chacun à chacun, les seconds équimultiples, ou leur sont égaux à la fois, ou plus petits à la fois.* Ce qui signifie que l'on aura $A : B = C : D$ ("A et B en même raison que C et D") si et seulement si, quels que soient m et n entiers, se réalisent ensemble les relations : $mA > nB$ et $mC > nD$, ou $mA = nB$ et $mC = nD$, ou $mA < nB$ et $mC < nD$. On voit que la vérification de l'égalité des rapports dépend *a priori* d'une infinité d'opérations ! Cela dit, si l'on trouve un couple (m, n) de nombres tel que $mA = nB$ et $mC = nD$, alors les raisons $A : B$ et $C : D$ sont égales et pourraient être représentées par un nombre rationnel. Euclide donne une condition nécessaire et suffisante, fondée sur une construction géométrique style Théorème de Thalès, pour que deux couples de segments soient proportionnels (VI, 2). En ramenant le calcul des aires polygonales à

Sans recourir aux nombres, on pourrait par exemple développer une notion d'aire "relative" : on décrit l'aire des figures générales à partir de l'aire de figures plus simples. Par exemple en décomposant les figures polygonales en triangles. Cette approche mène au difficile problème de comparer les aires d'un disque et d'un carré... qu'elle ne permet pas de résoudre.

En fait il n'y a pas de manière élémentaire de définir la notion d'aire. Il faut pour cela faire intervenir un processus de passage à la limite. Dans ce paragraphe, nous allons caractériser l'aire des figures planes comme une *fonction*, qui à une figure plane associe un nombre. Les valeurs de cette fonction seront obtenues comme limites. Plus précisément, nous allons définir pour (certaines) figures planes F un nombre $s(F)$, qui représente l'aire/surface de F, et nous allons voir que :

(P) (positivité) $s(F)$ est un nombre réel *positif ou nul*;

(A) (additivité) si F' et F'' n'ont pas de points intérieur en commun, alors l'aire de leur réunion, notée $F' + F''$, est la somme des aires

$$s(F' + F'') = s(F') + s(F'') \ ;$$

(I) (invariance par transport parallèle) si F' est obtenue à partir de F par une translation, alors $s(F') = s(F)$; [3]

(N) (normalisation) le carré unité C a aire égale à 1, $s(C) = 1$.

Il s'agit là de propriétés élémentaires de l'aire. En fait il n'est pas clair que toute figure possède une aire, et à ce stade on pourrait même nous rétorquer de ne pas avoir défini ce que nous entendons par figure... Nous allons remédier à ces imprécisions : dans ce qui suit une *figure* (plane) est tout simplement un sous-ensemble du plan $\mathbf{R}^2$ et nous allons donner un sens précis aux mots "la figure F admet une aire $s(F)$". Ensuite nous allons indiquer comment montrer que la notion d'aire introduite est la seule vérifiant les quatre propriétés ci-dessus.

Figures mesurables.[4]

Même si nous ne savons pas (encore) comment définir la notion d'aire, nous avons une manière de calculer des valeurs approchées de l'aire d'une figure donnée (qui admet une aire)–comme nous l'avons fait pour le cercle ! On munit le plan d'un repère orthonormé et on le quadrille avec des carrés de côté 1. Appelons ce quadrillage le 0-ième quadrillage. Si la figure dont nous voulons calculer l'aire contient par exemple 3 carrés de ce quadrillage, alors son aire est au moins égale à 3. Si la figure est contenue dans la réunion de 5 carrés de ce quadrillage, alors son aire est inférieure ou égale à 5. Pour obtenir une meilleure approximation on passe à des quadrillages plus fins : on subdivise chaque carré du 0-ième quadrillage en 100 carrés de côtés 1/10 et on obtient le 1-er quadrillage, puis en subdivisant tour à tour les carrés de la même manière on obtient le *k-ième quadrillage*, dont les carrés sont de côtés $1/10^k$. La figure F contiendra a_k carrés du k-ième quadrillage et sera contenue dans la réunion de b_k carrés. Ainsi on obtient deux suites de nombres (a_k) et (b_k). Chaque carré du k-ième quadrillage contient 100 carrés du $(k+1)$-ième, et F pourrait contenir des carrés du $(k+1)$-ième quadrillage en plus des $100a_k$ carrés provenant du quadrillage précédent, ce qui donne les inégalités $a_k/10^{2k} \leq a_{k+1}/10^{2(k+1)}$. De même pour les b_k on a $b_{k+1}/10^{2(k+1)} \leq b_k/10^{2k}$, d'où

$$a_0 \leq \frac{a_1}{10^2} \leq \frac{a_2}{10^4} \leq \ldots \leq \frac{a_k}{10^{2k}} \leq \ldots \leq \frac{b_k}{10^{2k}} \leq \ldots \leq \frac{b_2}{10^4} \leq \frac{b_1}{10^2} \leq b_0 \ .$$

un calcul sur des segments, il fait de même pour les aires de ces figures simples (VI, 16 à 22). Dans les théories modernes, on a substitué la considération de multiples indéfiniment grands, par celle de sous-multiples infiniment petits (voir Chap. III de B. Levi, "En lisant Euclide", Agone, Paris, 2003).

[3] On s'attend évidemment que l'aire soit préservée par d'autres transformations–comme les rotations–, on verra plus loin, que c'est le cas.

[4] Nous suivons de près la présentation de V. G. Boltianskii, "Hilbert's third problem", John Wiley & sons, New York, 1978.

On voit donc que les limites

$$\underline{s}(F) := \lim_{k\to\infty} \frac{a_k}{10^{2k}} \quad \text{et} \quad \overline{s}(F) := \lim_{k\to\infty} \frac{b_k}{10^{2k}}$$

existent et satisfont $\underline{s}(F) \leq \overline{s}(F)$. D'après l'idée intuitive d'aire, on s'attend à ce que ces deux limites coïncident et donnent la valeur de l'aire. Nous allons renverser cette approche et décréter, que si ces limites coïncident, alors F est *mesurable* (ou quarrable) et on notera $s(F)$ la valeur commune $\underline{s}(F) = \overline{s}(F)$, que l'on appellera *l'aire de* F.[5]

Nous avons maintenant un candidat pour la notion d'aire d'une figure (mesurable). *A priori* il n'est pas clair que nous tenons là la bonne définition. Le défaut le plus évident de cette définition est qu'elle dépend d'un choix de quadrillages : avec un autre choix de quadrillages on obtiendrait des suites (a_k) et (b_k) toutes différentes et rien ne nous garantit, que les nombres définis par ces suites différentes donnent la même limite, si ce n'est...l'intuition. Il nous faut voir que cette définition résiste à tous les tests. Commençons par vérifier les propriétés (P)-(N).

Preuve de la positivité. Soit F une figure mesurable, qui contient a_0 carrés du 0-ième quadrillage. Vu que $0 \leq a_0$, on a que tous les $a_k/10^{2k}$ sont positifs ou nuls et par conséquent $s(F) = \underline{s}(F) \geq 0$.

Preuve de l'additivité. Soit F' et F'' des figures mesurables sans points intérieurs en commun et soit $F = F' + F''$ leur réunion. Nous allons montrer que F est mesurable et que $s(F) = s(F') + s(F'')$. Notons respectivement $a_k^{'}$, $a_k^{''}$, et a_k les (premiers) nombres associés à F', F'' et F. Par hypothèse aucun carré peut être contenu à la fois dans F' et dans F''. Par conséquent $a_k \geq a_k^{'} + a_k^{''}$, et en divisant par 10^{2k} et en passant à la limite on obtient

$$\underline{s}(F) \geq s(F') + s(F'') \ .$$

De même : $\overline{s}(F) \leq s(F') + s(F'')$. Mais comme on a toujours $\underline{s}(F) \leq \overline{s}(F)$ on a bien égalité, F est mesurable et l'additivité est prouvée.

Preuve de l'invariance (cas particulier) et de la normalisation. Le carré unité C est le carré ayant le repère orthonormé pour côtés. Nous allons commencer par montrer que C, ainsi que tout carré C' obtenu à partir de C par une translation *parallèle aux axes* a aire égale à 1. Soit q_0 le sommet du carré du premier quadrillage, qui se trouve le plus près du sommet, image de l'origine par la translation ("en bas à gauche"). Il est clair que q_0 se trouve à une distance inférieure à $1/10$ des deux côtés les plus proches de C'. On en déduit que C' contient un carré de côté $9/10$ formé de 81 carrés du premier quadrillage. De même, C' est contenu dans la réunion de 121 carrés du premier quadrillage. Plus généralement C' contient $(10^k - 1)^2$ carrés du k-ième quadrillage et est contenu dans $(10^k + 1)^2$ tels carrés. Ainsi les nombres a_k et b_k pour C' satisfont $a_k \geq (10^k - 1)^2$ et $b_k \leq (10^k + 1)^2$, d'où

$$\frac{a_k}{10^{2k}} \geq (1 - \frac{1}{10^k})^2 \quad \text{et} \quad \frac{b_k}{10^{2k}} \leq (1 + \frac{1}{10^k})^2 \ .$$

En prenant la limite on obtient $\overline{s}(C') \leq 1 \leq \underline{s}(C')$ et par conséquent, que C' est mesurable et d'aire 1.

On démontre de même que le translaté d'un carré du k-ième quadrillage est mesurable et a aire $1/10^{2k}$.

Si maintenant G est la somme de a carrés du k-ième quadrillage, disons $G = P_1 + \cdots + P_a$, alors par additivité $s(G)$ est la somme des $s(P_i)$, donc $s(G) = a/10^{2k}$. Donc, vu que l'aire des carrés du quadrillage est invariante par translation, il en est ainsi pour l'aire de G.

[5] Cette notion d'aire est attribuée à Jordan. Il faudrait que nous reprenions les propriétés (P)-(N) à la lumière de cette définition, et que nous ajoutions le mot "mesurable" après le mot "figure", pour toutes les figures dont on considère l'aire. Dans ce qui suit, nous n'allons considérer l'aire, que des figures mesurables...

Soit alors F une figure mesurable quelconque. Pour tout $\epsilon > 0$ on peut trouver k et une figure G composée de a_k carrés du k-ième quadrillage tels que $a_k/10^{2k} = s(G) > s(F) - \epsilon/2$. Les images G' et F' de G et F par une translation satisfont évidemment $G' \subset F'$, de plus nous savons que G' est mesurable. Par conséquent il existe ℓ et une figure G^* composée de a^*_ℓ carrés du ℓ-ième quadrillage avec

$$G^* \subset G' \subset F'$$

et

$$\underline{s}(F') \geq \frac{a^*_\ell}{10^{2\ell}} > s(G') - \frac{\epsilon}{2} = s(G) - \frac{\epsilon}{2} > s(F) - \epsilon \ .$$

Vu que ceci est vrai pour tout ϵ on obtient $\underline{s}(F') \geq s(F)$. Un argument similaire donne $\overline{s}(F') \leq s(F)$. D'où F' est mesurable et $s(F') = s(F)$.

Exercice. Réfléchir à une éventuelle "additivité infinie" de l'aire. Noter qu'un segment a aire 0, et un carré est somme d'une infinité (non-dénombrable) de segments.

L'aire des triangles et des polygones.

Pour l'instant nous n'avons calculé l'aire que de figures composées de carrés de quadrillages. Si on veut calculer l'aire d'une figure ayant des côtés "en biais" par rapport aux quadrillages–par exemple un triangle–, il faut changer la méthode d'approximation. Avant de calculer la valeur des aires des figures planes les plus simples, nous allons montrer que tous les polygones sont mesurables. En fait une fois que nous saurons que ceux-ci sont tous mesurables, nous pourrons utiliser la propriété d'additivité pour trouver les valeurs numériques à partir de l'aire des figures de base. Ainsi pour retrouver la formule pour l'aire d'un triangle quelconque on observera qu'elle se ramène au calcul de l'aire des triangles rectangles (tirer une hauteur du triangle) et l'aire d'un triangle rectangle est la moitié de l'aire d'un rectangle. Le calcul de l'aire d'un rectangle sera l'objet du prochain paragraphe[6].

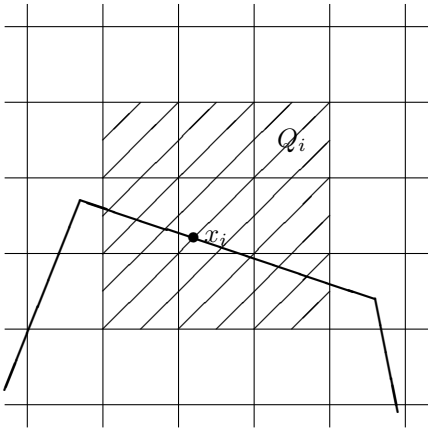
Un *polygone* est par définition une réunion finie de triangles, ou de manière équivalente c'est un sous-ensemble borné et fermé du plan ayant un bord qui est la réunion d'un nombre fini de segments de droite.

Proposition. *Tout polygone est mesurable.*

Preuve. Il faut que nous arrivions à estimer la différence entre les termes des suites (a_k) et (b_k), qui donnent les approximations par défaut et par excès de l'aire du polygone. La difficulté est donnée par le fait, que contrairement aux figures que nous avons considérées jusqu'à présent, les polygones peuvent avoir des côtés qui ne sont pas parallèles aux quadrillages. En effet, soit F un polygone et soit L la ligne polygonale, réunion des segments L_1, L_2, ..., L_m qui bordent F. Fixons un entier k. Tout carré du k-ième quadrillage ayant des points en commun avec F, mais qui ne fait pas partie des a_k carrés à l'intérieur de F doit avoir des points en commun avec L. Nous allons définir une une suite finie de points $x_1, x_2, \ldots, x_q$ sur la ligne L, qui aura la propriété, que tout point sur L est à une distance d'un des points x_i inférieure à $1/10^k$. Le nombre q de ces points sera borné en fonction du nombre et de la longueur des segments L_j. Cette borne nous donnera l'estimation de la différence $b_k - a_k$. Soit p_j la longueur du segment L_j et soit p la longueur de L, de manière que $p = p_1 + \cdots + p_m$. Soit x_1 une des extrémités de L_1. En partant de x_1 on subdivise les L_j en segments de longueur $1/10^k$. Les extrémités x_i de ces segments font l'affaire. Observons qu'avec cette construction nous avons obtenu au plus $p10^k + m$ points. En effet, sur chaque L_j nous avons au plus $p_j 10^k + 1$ points et $(p_1 + \cdots + p_m)10^k + (1 + \cdots 1) = p10^k + m$.

Pour chaque point x_i on considère alors la figure Q_i constituée des neuf carrés du k-ième quadrillage, qui l'entourent (voir dessin).

[6] Avant de lire ce qui suit essayez de déduire, par des moyens élémentaires, la valeur de l'aire d'un rectangle du fait que le carré unité a aire 1. Vous devriez arriver à le faire pour des rectangles ayant des côtés de longueur rationnelle.



Vu que tous les points à une distance inférieure à $1/10^k$ de x_i se trouvent dans Q_i, on déduit que les carrés du k-ième quadrillage qui rencontrent la ligne L sont dans la réunion $Q_1 \cup \ldots \cup Q_q$. Ainsi le nombre de ces carrés est inférieur à $9q \leq 9(p10^k + m)$. Par les remarques du début de la démonstration nous déduisons l'inégalité

$$0 \leq b_k - a_k \leq 9(p10^k + m) \ .$$

Ainsi, $0 \leq b_k/10^{2k} - a_k/10^{2k} \leq 9p/10^k + 9m/10^{2k}$, qui donne en passant à la limite sur k, que $\underline{s}(F)$ égale $\overline{s}(F)$ et donc que F est mesurable.

Exercice. On considère le parallélogramme défini par l'origine et par les points du plan de coordonnées (x_1, y_1) et (x_2, y_2). Relier l'aire du parallélogramme au nombre $x_1y_1 - x_2y_2$.

L'aire d'un rectangle.

Par le résultat du paragraphe précédent, nous savons que tout rectangle est mesurable. Ce que nous voulons montrer ici est le résultat attendu suivant.

Proposition. *Soit F un rectangle dont les côtés sont de longueur a et b, alors $s(F) = ab$.*

Preuve. Nous verrons que pour calculer l'aire d'un rectangle quelconque nous serons amenés à affiner notre étude de la fonction aire. Comme précédemment *on suppose d'abord que F a les côtés parallèles aux axes.* Alors vu la propriété d'invariance, que nous avons établi plus haut, nous pouvons supposer, que le sommet "en bas à gauche" de F coïncide avec un sommet du 0-ième quadrillage–disons q_0. Si p_0 dénote le sommet du k-ième quadrillage le plus proche du sommet de F "en haut à droite", alors F contient le rectangle G_k composé de a_k carrés ayant q_0 et p_0 comme sommets opposés. Si les côtés de G_k sont de longueur $\alpha_k/10^k$ et $\beta_k/10^k$, on a $a_k = \alpha_k\beta_k$ et

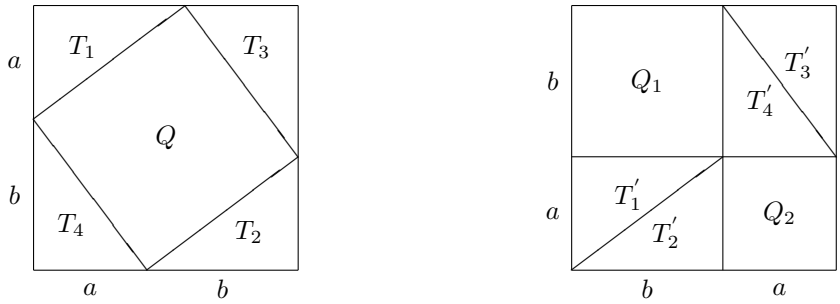
$$a10^k - 1 < \alpha_k \leq a10^k \quad \text{et} \quad b10^k - 1 < \beta_k \leq b10^k \ .$$

On tire de ceci les inégalités

$$(a10^k - 1)(b10^k - 1) < a_k \leq ab10^{2k} \ .$$

De même on peut circonscrire à F un rectangle H_k composé de $b_k = (\alpha_k + 1)(\beta_k + 1)$ carrés du k-ième quadrillage et donc

$$ab10^{2k} < b_k \leq (a10^k + 1)(b10^k + 1) \ .$$



Des inégalités précédentes on tire que

$$\frac{a_k}{10^{2k}} > (a - \frac{1}{10^k})(b - \frac{1}{10^k}) \quad \text{et} \quad \frac{b_k}{10^{2k}} > (a + \frac{1}{10^k})(b + \frac{1}{10^k}) \ ,$$

qui donne $\overline{s}(F) \leq ab \leq \underline{s}(F)$. D'où F est (mesurable et) d'aire ab.

Pour calculer l'aire d'autres rectangles nous observons, que ceux-ci sont obtenus d'un rectangle avec les côtés parallèles aux axes par une translation suivie éventuellement d'une rotation. Nous allons donc nous concentrer sur le comportement des aires par ces transformations. La remarque clé est basée sur le diagramme donnant la "démonstration indienne du théorème de Pythagore", à savoir que, tout carré Q peut être circonscrit par un plus petit carré P ayant les côtés parallèle aux axes (voir figure ci-dessous). Ceci va nous permettre de montrer que *tout carré congruent au carré unité a aire* 1. La différence entre P et Q est donnée par quatre triangles rectangles congruents. Soient a et b les longueurs des côtés les plus petits de ces triangles. Le carré P se décompose aussi en une somme de deux carrés Q_1 et Q_2 de côtés a et b respectivement et de deux rectangles formés par deux translatés des triangles précédents (voir la figure de droite, qui donne une démonstration géométrique de l'égalité $(a+b)^2 = a^2+2ab+b^2$).[7] Des deux décompositions de P, en utilisant l'invariance de l'aire par transport parallèle et l'additivité, on tire que $s(Q) = s(Q_1) + s(Q_2)$. Par conséquent si Q est congruent au carré unité il a aire 1 : en effet on a alors par le Théorème de Pythagore que $a^2 + b^2 = 1$ et vu que les carrés Q_1 et Q_2 ont les côtés parallèles aux axes, nous savons d'après le début de la démonstration, que $s(Q_1) = a^2$ et $s(Q_2) = b^2$.

On peut maintenant terminer la démonstration en observant que les résultats partiels obtenus nous permettent d'utiliser des quadrillages "tournés", ayant les côtés parallèles aux côtés du rectangle donné.

Unicité de la fonction aire.

Il n'est pas trop compliqué de montrer, que *la fonction aire que nous avons construit est la seule à vérifier les propriétés (P)-(N).* Autrement dit, la notion d'aire des figures planes est complètement caractérisée par ces quatre propriétés, et si une fonction associant un nombre à une figure satisfait à ces propriétés, alors il s'agit de la fonction aire. On montre d'abord que les propriétés déterminent les

[7] C'est le "théorème du gnomon", Proposition I.43 des Éléments, dont on tire une démonstration géométrique du Théorème de Pythagore, Proposition I.47. Pour ne pas tourner en rond, il faut ici utiliser une démonstration analytique de ce théorème.

valeurs de la fonction sur les polygones : en effet par la normalisation (N), par l'additivité (A) et par l'invariance (I), la valeur sur un carré du k-ième quadrillage doit être égale à $1/10^{2k}$; donc par additivité l'aire de toute réunion de carrés d'un quadrillage est déterminée ; ainsi, si F est un polygone et si G_k dénote la réunion des a_k carrés du k-ième quadrillage contenus dans F, on a $F = G_k + H_k$, où H_k est le polygone complémentaire de G_k dans F (on a $H_k = \overline{F \setminus G_k}$). Par la positivité (P) et par l'additivité, on a que la valeur sur F est supérieure ou égale à la valeur sur G_k. De même, en considérant la réunion $G_k^{'}$ des carrés du k-ième quadrillage contenant F on obtient, que la valeur sur F est inférieur ou égale à la valeur sur $G_k^{'}$; on termine en passant à la limite sur k (et en utilisant que F est mesurable).

Pour les figures quelconques on démontre d'abord l'énoncé suivant, qui permet de ce ramener au cas des polygones. *Une figure est mesurable si et seulement si pour tout réel positif ϵ il existe des polygones G et H tels que $G \subset F \subset H$ avec $s(H) - s(G) < \epsilon$.*

Invariance forte.

On peut utiliser le fait que la fonction aire sur les polygones est caractérisée par les propriétés (P)-(N), pour montrer qu'elle est invariante par des transformations du plan plus générales que les translations. Il s'agit des transformations qui préservent les distances. On peut montrer que celles-ci sont obtenues en composant des translations, des rotations et des réflexions. Elles sont aussi caractérisées par la donnée de six nombres (réels) a, b, c, d, p et q, tels que

$$a^2 + c^2 = 1 = b^2 + d^2 \quad \text{et} \quad ab + cd = 0 \ ,$$

ou encore tels que $(a, b, c, d) = (\cos\theta, \pm\sin\theta, \sin\theta, \mp\cos\theta)$ pour θ un nombre réel (avec $0 \leq \theta < 2\pi$). A un tel sextuplet on associe la transformation qui envoie le point du plan de coordonnées (x, y) sur le point (x', y') où :

$$\begin{aligned} x' &= ax + by + p \\ y' &= cx + dy + q \ . \end{aligned}$$

Exercice. Vérifier que la transformation donnée par le sextuplet $(0, -1, 0, 1, 0, 0)$ correspond à une rotation d'angle droit. Déterminer le sextuplet correspondant à une translation. Vérifier que les transformations ainsi définies préservent les longueurs et la relation orthogonalité.

Une telle transformation g transforme le carré unité en un carré, qui par ce que nous avons montré plus haut, est encore d'aire 1. Ainsi, si on pose $s^*(F) = s(g(F))$ on obtient une fonction, qui certainement satisfait les propriétés (P)-(N) pour les polygones. Par l'unicité d'une telle fonction on a donc $s^*(F) = s(F)$ pour les polygones. Ainsi $s(g(F)) = s(F)$ pour les polygones et en approchant les figures quelconques par des polygones, nous voyons que l'aire est invariante par toute transformation.

Indépendance des propriétés caractéristiques.

On peut aussi montrer que *les propriétés (P)-(N) sont indépendantes*, c'est-à-dire que l'on ne peut pas déduire l'une d'entre elles de l'ensemble des autres. Ainsi, ces propriétés donnent une caractérisation minimale de la notion d'aire (des figures mesurables).

Indépendance de la normalisation. ça c'est facile ! Il suffit de considérer la fonction qui a toute figure associe le nombre 0. Elle satisfait toutes les propriétés sauf celle qui donne la valeur 1 au carré unité.

Indépendance de l'additivité. Ce n'est pas beaucoup plus dur de montrer l'indépendance de l'additivité : il suffit de considérer la fonction qui a toute figure associe la valeur 1.

Indépendance de l'invariance. Pour construire une fonction qui ne prend que des valeurs positives, qui soit additive et qui donne au carré unité la valeur 1, sans être invariante, considérons une droite ℓ, qui subdivise le plan en deux demi-plans P_1 et P_2 tels que P_1 contienne le carré unité. Soit f la fonction, qui à la figure mesurable F associe le nombre $f(F) = s(F \cap P_1) + 2s(F \cap P_2)$. Il est clair que si Q dénote un carré obtenu en translatant le carré unité dans le demi-plan P_2 on aura $f(Q) = 2$ et donc f n'est pas invariante. Par contre, f satisfait aux autres propriétés.

Indépendance de la positivité. C'est le point le plus délicat, mais aussi le plus intéressant. A priori, lorsqu'on veut mettre en évidence les propriétés de l'aire, la positivité est une propriété, que l'on oublierait presque, tellement elle est évidente. Pourtant : *la formule pour l'aire d'un rectangle ne peut pas se démontrer avec les seules propriétés d'additivité, d'invariance (même généralisée) et de normalisation.* En effet, on peut construire une fonction g, qui satisfait toutes ces propriétés et qui–par exemple–pour un rectangle de côtés 1 et $\sqrt{2}$ prend la valeur -1 (et n'est donc pas positive). La construction d'une telle fonction est un joli exemple de comment on peut utiliser l'axiome du choix, que nous avons mentionné rapidement dans notre discussion de l'axiomatique de Zermelo-Frænkel pour la théorie des ensembles.

On procède comme suit.[8] On utilise l'axiome du choix pour montrer l'existence d'un sous-ensemble B^* de l'ensemble $\mathbf{R}$ des nombres réels tel que :

(CLR) (combinaison linéaire rationnelle) tout nombre réel x peut s'écrire comme une combinaison linéaire (finie)
$$x = q_1 b_1 + \cdots + q_m b_m \ ,$$
d'éléments b_i de B avec les q_i *rationnels* (m dépend de x) ;

(IR) (indépendance rationnelle) aucun élément de B n'est combinaison linéaire rationnelle d'autres éléments de B.

Soit M l'ensemble des sous-ensembles B de $\mathbf{R}$, qui satisfont la condition (IR) et qui contiennent 1 et $\sqrt{2}$. Cet ensemble est non-vide, il contient par exemple l'ensemble formé de 1 et $\sqrt{2}$. L'ensemble B^* sera défini comme un élément maximal dans M, où nous ordonnons les éléments B par inclusion : $B_1 < B_2$ signifie $B_1 \subset B_2$. L'axiome du choix sert à montrer, que M a bien un élément maximal ! La maximalité de B^* entraîne qu'il a aussi la propriété (CLR). Car, si x est un nombre réel, de deux choses l'une : ou bien x appartient à B^* et $x = x$ est la combinaison linéaire recherchée, ou alors en ajoutant x à B^* on obtient un sous-ensemble E de $\mathbf{R}$ (contenant 1 et $\sqrt{2}$), qui ne peut être élément de M par la maximalité de B^*, donc il doit exister une combinaison linéaire pour x en terme des éléments de B^*.

Si M a un élément maximal, alors tout sous-ensemble C d'éléments de M est forcément majoré (dans M), c'est-à-dire que il existe un élément q dans M (et pas forcément dans C) tel que pour tout élément a de C on a $a \leq q$. L'axiome du choix permet de montrer, que si cette propriété est satisfaite pour toute suite bien ordonnée C dans M, alors M possède un élément maximal. Plus précisément, on appelle *chaîne* dans M un sous-ensemble C de M, tel que pour toute paire d'éléments différents a et b de C on a soit $a < b$, soit $b < a$.

L'axiome du choix est équivalent au *Lemme de Zorn*, qui dit que *si toute chaîne C de l'ensemble ordonné (non-vide) M est majorée (dans M), alors M possède un élément maximal.*

Pour appliquer le Lemme de Zorn à notre cas, il faut vérifier que si C est une chaîne de M, alors elle est majorée. Définissons l'ensemble B_C comme la réunion de tous les éléments de C, qui–rappelons le–sont des sous-ensembles de $\mathbf{R}$. On obtient ainsi un sous-ensemble de $\mathbf{R}$. Il est clair que B_C donne une majoration de la chaîne C, mais il nous faut montrer que B_C est bien un élément de M, c'est-à-dire qu'il satisfait (IR). Or, s'il existait une relation de dépendance rationnelle entre des éléments de B_C, disons $x_0 = q_1 x_1 + \cdots + q_m x_m$, avec q_i rationnel et x_0, x_i dans B_C, alors on aurait une contradiction. En effet, chacun des x_i se trouve dans un B_i élément de C et, comme C est une chaîne, parmi les B_i il en existe un plus grand, disons B_m. Alors x_0 et les x_i sont dans B_m. Mais B_m est élément de M et donc satisfait (IR), d'où la contradiction. Le Lemme de Zorn nous garantit donc l'existence d'un élément maximal B^* dans M.

Retournons maintenant à la démonstration de l'indépendance de la propriété de positivité. Donné un ensemble B^* ayant les propriétés (CLR) et (IR), construisons une fonction $h : \mathbf{R} \to \mathbf{R}$ comme suit. On pose $h(1) = 1$ et $h(\sqrt{2}) = -1$. Puis on donne des valeurs arbitraires aux images par h des éléments de B^*, par exemple on décrète que h prend la valeur 1 sur tous les éléments de B^*. Ceci suffit pour déterminer toutes les valeurs de f. En effet si x est réel, par (CLR) il existe une

[8] Cette démonstration sera peut-être plus claire après s'être familiarisé avec les notions de base de l'algèbre linéaire. *A contrario* elle permet d'introduire de manière originale la notion de combinaison linéaire dans un cas non géométrique.

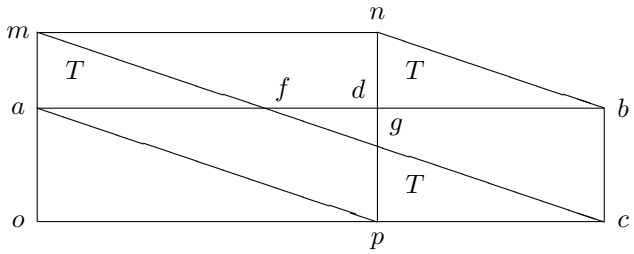


FIG. 10.1 – Rectangles équidécomposables.

expression de x comme combinaison linéaire $x = q_1 b_1 + \cdots + q_m b_m$ des éléments de B^*. On pose alors $h(x) = q_1 f(b_1) + \cdots + q_m f(b_m)$ et on vérifie que la fonction $s'(F) := h(s(F))$ satisfait toutes les propriétés, sauf la positivité : la positivité de s' est mise en défaut par le fait que pour le rectangle R de côtés 1 et $\sqrt{2}$ on a $s'(R) = h(s(R)) = h(\sqrt{2}) = -1$; la normalisation de s' est claire car $s'(C) = h(s(C)) = h(1) = 1$; l'additivité de s' suit de l'additivité de s et de ce que $h(x+y) = h(x)+h(y)$; l'invariance de s' suit directement de l'invariance de s.

Exercice. Montrer que la fonction h définie ci-dessus ne peut pas être continue.

On pourrait se dire, qu'en faisant appel à l'axiome du choix nous avons utilisé un marteau pour écraser une mouche. Ce n'est pas le cas. On peut démontrer que, *si l'on arrive à déduire la formule pour l'aire du rectangle à partir des seules propriétés d'additivité, d'invariance et de normalisation, alors c'est que la déduction repose sur la* négation *de l'axiome du choix.*[9]

Exercice. On dit que deux figures F et H sont *equidécomposables* si l'on peut décomposer la figure F en une somme de figures, qui ré-assemblées donnent la figure H.

a) Montrer que deux figures equidécomposables ont la même aire.
b) Montrer qu'un parallélogramme est équidécomposable avec le rectangle qui a la même base et la même hauteur.
c) Montrer qu'un triangle est équidécomposable avec le parallélogramme qui a la même base et hauteur égale à la moitié de la hauteur du triangle.
d) Dériver la formule usuelle pour l'aire d'un triangle.
e) Montrer que tout triangle est équidécomposable avec un rectangle. (Indications : considérer le rectangle construit en traçant la parallèle au côté le plus long du triangle et qui passe par le point au milieu de la hauteur perpendiculaire à ce côté.)
f) Montrer que deux rectangles ayant la même aire sont équidécomposables. (Indications : on peut disposer les deux rectangles de manière à ce qu'ils aient un angle droit en commun, comme sur la figure ; l'hypothèse que les rectangles *oabc* et *omnp* ont la même aire se traduit par le fait que les segments *ap*, *mc* et *nb* sont parallèles ; le cas où *mc* intersecte le rectangle *oadp* est alors clair (c'est le cas de la figure) ; quand *mc* n'intersecte pas on a que la longueur de *oc* est plus que le double de la longueur de *op* ; considérer le point *e* à mi-chemin entre *o* et *c* et le plus petit entier k tel que k fois le segment *op* recouvre *oe* ; découper le rectangle *omnp* en k parties congruentes à l'aide de segments parallèles à *op* ; en réarrangeant les k rectangles ainsi obtenus en disposant

[9]La démonstration suit par exemple des résultats de R. Solovay, dans "A model of set-theory in which every set of reals is Lebesgue mesurable", Annals of Mathematics, **92**(1970), 1–56. Car soit s'' une fonction satisfaisant toutes les propriétés sauf la positivité, qui est différente de la fonction donnant l'aire et soit k la fonction de la variable réelle obtenue en posant $k(x) = s''(R_x)$, où R_x dénote le rectangle de côtés 1 et x. Alors k est une fonction additive, qui ne peut pas être mesurable, et par les résultats de Solovay ceci implique l'axiome du choix.

le côté de longueur $\overline{op}$ sur oc on obtient un rectangle équidécomposable avec $omnp$, qui permet de se ramener au cas précédent.)

g) Montrer que deux polygones avec la même aire sont équidécomposables. [10] (Indications : décomposer en triangles pour montrer que tout polygone est équidécomposable avec une somme de rectangles ayant un côté commun et donc équidécomposable avec un rectangle unique.)

Exercice.

Un pâtissier distrait cuit par inadvertance un gâteau triangulaire qui a les trois côtés de longueur différente de manière à ce que le gâteau et la boîte prévue pour le contenir sont symétriques. Peut-on couper le gâteau de manière économique pour le faire rentrer dans la boîte sans en retourner aucune partie ? [11]

L'aire sous une parabole.

Nous avons maintenant défini la notion d'aire et nous avons calculé l'aire de quelques figures simples, essentiellement les polygones. Pour calculer l'aire de figures curvilignes il a longtemps fallu développer des méthodes *ad hoc*, jusqu'au moment où le calcul intégral a fourni des méthodes générales. Ici nous reproduisons les grandes lignes du calcul par Archimède de l'aire limitée par une parabole. [12] Il s'agit de construire une suite de triangles qui donnent une bonne approximation de la parabole (voir figure). Les triangles congruents ABD et ACE sont obtenus à partir des points D et E, qui sont les points

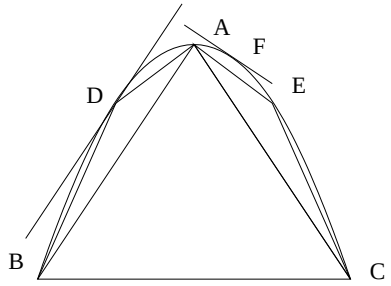


FIG. 10.2 – L'aire sous une parabole d'après Archimède.

sur la parabole par lesquels passe la tangente à la parabole parallèle à AB et à AC respectivement. On obtient deux triangles plus petits en considérant les tangentes parallèles aux côtés des triangles construits précédemment (voir le point F), et ainsi de suite. Archimède montre que si l'aire de ABC vaut T, alors l'aire de $ADBCE$ vaut $T + 1/4 \cdot T$. En insérant les quatre autres triangles comme AEF, on obtient une figure polygonale dont l'aire vaut $T + 1/4 \cdot T + 1/4^2 \cdot T$, *etc.* Ainsi, après la n-ième itération, l'aire du polygone approximant vaut $T\frac{1-a^n}{1-a}$, avec $a = 1/4$.[13] Ce qui donne (en passant à la limite !), que l'aire sous la parabole vaut $P = T/(1-a) = 4T/3$.

[10] Ce résultat a été démontré par Bolyai et Gerwein au 19ème siècle. Il dit en substance que pour les polygones il existe une théorie élémentaire de l'aire. Des résultats de même nature étaient connus des Grecs. On peut montrer qu'en dimension 3 il existe des polyèdres qui ont même volume, mais qui ne sont pas équidécomposables : c'est le Théorème de Dehn, qui résout le troisième d'une fameuse liste de problèmes posés par D. Hilbert lors du deuxième Congrès international des mathématiciens tenu à Paris en 1900. Des travaux autour du troisième problème de Hilbert ont encore récemment vu le jour et ont établi des connexions intéressantes entre des généralisations du problème et des questions de la théorie des nombres.

[11] La réponse à cette question permet de décider si l'on peut obtenir l'équidécomposabilité des polygones de même aire par des déplacements qui préservent l'orientation.

[12] Dans son ouvrage "La quadrature de la section orthogonale du cône".

[13] Archimède disait plutôt quelque chose comme "la somme d'un nombre fini d'aires en progression géométrique décroissante de raison 1/4, sommée à 1/3 de la dernière aire, donne 4/3 de la première".

Par le même type de calcul, Fermat a montré quelques 2000 ans plus tard que l'aire sous une courbe d'équation $y = x^a$, entre $x = 0$ et $x = B$, vaut $B^{a+1}/(a+1)$. Nous allons effectuer un calcul analogue plus loin pour illustrer le Théorème fondamental du calcul intégral.

10.2 Primitives.

Soit $f : I := [a, b] \to \mathbf{R}$ une fonction. Nous voulons définir la *mesure*

$$m(f) = \int_a^b f(x)\mathrm{dx} \ .$$

Considérons d'abord le cas d'une fonction f *étagée*, c'est-à-dire telle qu'il existe une partition $I = I_1 \cup \ldots \cup I_p$ en intervalles $I_k = (a_k, b_k)$ avec $I_i \cap I_j = \emptyset$ dès que $i \neq j$, ayant la propriété que f restreinte à chacun des I_k est constante : disons $f(x) = c_k$ pour tout x dans I_k. Alors on définit $m(I_k)$ comme étant la longueur de I_k, à savoir $m(I_k) = |b_k - a_k|$, et on pose

$$m(f) := \sum_k c_k \cdot m(I_k) \ .$$

Ensuite on utilise l'exercice du début du chapitre pour traiter le cas des fonctions f approchées par des fonctions étagées. On dit que f est *réglée* si pour tout $r > 0$, il existe une fonction étagée φ_r (définie sur I), telle que

$$\varphi_r(x) \leq f(x) \leq \varphi_r(x) + r \ .$$

Soit $\psi_r(x) := \varphi_r(x) + r$, alors $m(\psi_r) - m(\varphi_r) = (b-a)r$. Soit E l'ensemble des valeurs des $m(\varphi_r)$ et soit F l'ensemble des valeurs des $m(\psi_r)$. Alors nous savons qu'il existe un nombre $m(f)$, aussi noté $\int_a^b f(x)\mathrm{dx}$, [14] et appelé l'*intégrale* de f sur $[a, b]$, tel que pour tout r

$$m(\varphi_r) \leq m(f) = \int_a^b f(x)\mathrm{dx} \leq m(\psi_r) \ .$$

Plus généralement il suffirait d'avoir deux fonctions étagées φ_r et ψ_r indépendantes ayant la propriété que pour tout r on a $m(\psi_r) - m(\varphi_r) < r$.

Exercice. Une fonction continue en tout point d'un intervalle est-elle réglée ?

Exemple. Faisons le calcul pour $f(x) = 1/x$ sur un intervalle $[a, b]$ avec $a > 0$. [15] On fixe n et on pose $q = q_n = (b/a)^{1/n}$. On subdivise l'intervalle $[a, b]$ à l'aide des points $a = aq^0$, aq, aq^2, ..., $aq^n = b$. On appelle φ_n la fonction définie sur $[a, b]$ constante sur (aq^k, aq^{k+1}) de valeur $c_k = 1/aq^{k+1}$. C'est la valeur f "à droite de l'intervalle". On a $\varphi_n \leq f$ et

$$\begin{aligned} m(\varphi_n) &= \frac{1}{aq}(aq - a) + \frac{1}{aq^2}(aq^2 - aq) + \cdots + \frac{1}{aq^n}(aq^n - aq^{n-1}) \\ &= \frac{1}{q}(q-1)n = n\left((\frac{b}{a})^{1/n} - 1\right)\frac{1}{(b/a)^{1/n}} \ . \end{aligned}$$

[14] Le signe $\int$ est censé rappeler la lettre "S", pour "somme". Le dx, que l'on peut oublier, mais qui est utile pour désigner la variable d'intégration, rappelle les différences $|b_k - a_k|$ donnant la longueur des intervalles.

[15] Pour compléter le calcul nous aurons besoin d'une expression pour la fonction log comme limite, à savoir $\log x \overset{(*)}{=} \lim_{n\to\infty} n(x^{1/n} - 1)$. En utilisant le fait que $(\exp(t) - 1)/t$ tend vers 1 lorsque t tend vers 0, on voit que le membre de droite de (*) est bien l'inverse de exp en substituant x par $\exp x$, ce qui donne $\lim_{n\to\infty} n(\exp(x)^{1/n} - 1) = \lim_{n\to\infty} n(\exp(x/n) - 1) = \lim_{t\to 0} x(\exp(t) - 1)/t = x$ (poser $t = x/n$).

Soit de même ψ_n prenant sur le même intervalle la valeur "à gauche" : ψ_n vaut $1/aq^k$ sur (aq^k, aq^{k+1}). Alors $\psi_n = q\varphi_n \geq f$ et $m(\psi_n) = n\left((b/a)^{1/n} - 1\right)$. Posons $c = b/a$, alors on obtient

$$n(c^{1/n} - 1) \leq \int_a^b f(x)\mathrm{dx} \leq n(c^{1/n} - 1)\frac{1}{c^{1/n}} \ ,$$

qui d'après la note ci-dessus, et le fait que $c^{1/n}$ tend vers 1 avec n, donne, pour n tendant vers l'infini, la valeur

$$\int_a^b \frac{1}{x}\mathrm{dx} = \log b - \log a \ .$$

Si F a dérivée f, on dit que F est une *primitive* de f. On se souvient que la dérivée de log est $1/x$. Nous venons donc de voir que l'intégrale de $1/x$ entre a et b se calcule comme la différence des valeurs d'une de ses primitives.

Les résultats généraux qui nous intéressent sont le suivants :

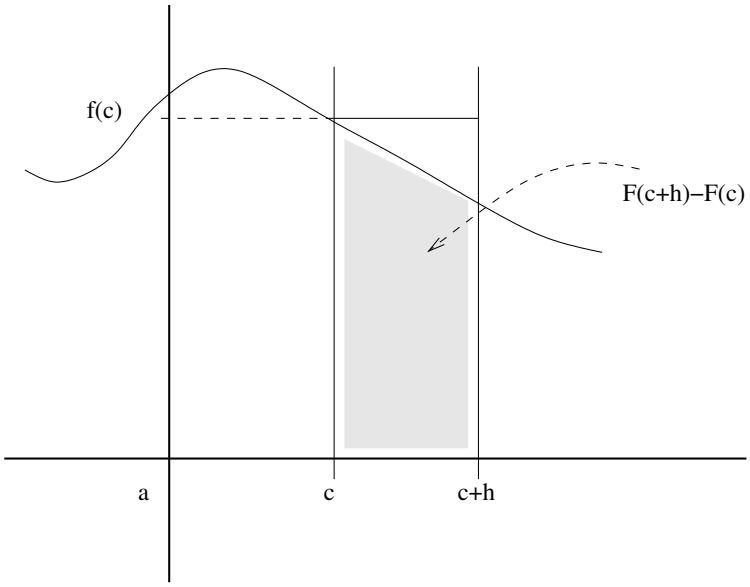
- *pour toute fonction $f : [a, b] \to \mathbf{R}$, continue en tout point de l'intervalle $[a, b]$, on peut définir* l'intégrale (définie)
$$m(f) = \int_a^b f(t)\,dt \ ;$$
Pour g une autre fonction continue définie sur $[a, b]$ et pour λ et μ réels, on la propriété de linéarité *:*
$$m(\alpha f + \beta g) = \alpha m(f) + \beta m(g) \ .$$
- *une fonction F, dont la dérivée est f, est appelée une* primitive *de f ; deux primitives F et G de f diffèrent par une constante : $F = G + c$;*
- *pour x dans $[a, b]$ on a la formule*
$$\int_a^x f(t)\,dt = F(x) - F(a) \ ;$$
- *la fonction F définie sur $[a, b]$ par*
$$F : x \mapsto \int_a^x f(t)\,dt$$
est une primitive de f : on a $F'(c) = f(c)$. C'est la primitive qui s'annule en $x = a$.

Arrêtons-nous pour justifier cette dernière affirmation, qui est aussi connue comme le *Théorème fondamental du calcul intégral*. Considérons la figure.
Si nous interprétons l'intégrale de f comme l'aire sous le graphe de f, nous voyons que l'aire hachurée vaut $F(c + h) - F(c)$. La dérivée de F en c (si elle existe) est la limite de $(F(c + h) - F(c))/h$ pour h tendant vers 0. Or pour h petit, l'aire hachurée est proche de l'aire du rectangle de côtés $f(c)$ et h (à condition que f soit suffisamment régulière), donc "on voit" que la limite est bien $f(c) = f(c)h/h$.

Le fait que l'intégrale donne une opération inverse à la dérivation, une anti-dérivation, permet d'établir des formules très utiles pour le calcul.

La formule de changement de variable. Soit $\phi : [\alpha, \beta] \to \mathbf{R}$ une fonction dérivable ayant une dérivée ϕ' continue. Soit $c = \inf \phi$, $d = \sup \phi$ et $f : [c, d] \to \mathbf{R}$ une fonction continue. Alors $f \circ \phi$ est bien définie et, en posant $a = \phi(\alpha)$, $b = \phi(\beta)$, on a

$$\int_a^b f = \int_\alpha^\beta (f \circ \phi)\phi' \ .$$



Pour la *démonstration* notons que, si on définit $F : [c,d] \to \mathbf{R}$ par $F(x) = \int_a^x f$, alors $F' = f$ (par le résultat justifié ci-dessus), et si $G = F \circ \phi : [\alpha, \beta] \to \mathbf{R}$, alors par la règle de dérivation d'une composition

$$G'(x) = F'(\phi(x))\phi(x) = (f \circ \phi)\phi'(x) \ .$$

Donc

$$\int_\alpha^\beta (f \circ \phi)\phi' = \int_\alpha^\beta G' = G(\beta) - G(\alpha) = F(b) - F(a) = \int_a^b f \ .$$

Exercice. Montrer que $\int_a^b f(u)\mathrm{du} = -\int_b^a f(u)\mathrm{du}$.

Intégration par parties. Si F et G sont des primitives des fonctions continues $f, g : [a,b] \to \mathbf{R}$, alors

$$\int_a^b fG + \int_a^b Fg = F(b)G(b) - F(a)G(a) \ .$$

Ceci suit directement de la formule pour la dérivation du produit FG.

10.3 Le nombre π est irrationnel.

Voici une démonstration élémentaire du fait que π est irrationnel. On va même obtenir un résultat plus précis.

Théorème. Le nombre π^2 n'est pas rationnel.

Démonstration. Pour tout entier $n \geq 1$, on définit la fonction f_n sur l'intervalle $[0,1]$ par

$$f_n(x) = \frac{x^n(1-x)^n}{n!}.$$

On note tout d'abord que

$$0 < f(x) < \frac{1}{n!} \text{ pour } x \in]0,1[, \tag{10.3.0}$$

et d'autre part on montre que toutes les valeurs des dérivées de f en 0 et 1, c'est-à-dire les nombres $f^{(k)}(0)$ et $f^{(k)}(1)$ sont des entiers.

Supposons que $\pi^2 = a/b$ avec $(a,b) = 1$, a et b des entiers positifs. Posons pour $n \geq 1$

$$G_n(x) = b^n \left\{ \pi^{2n} f(x) - \pi^{2n-2} f''(x) + \cdots + (-1)^n f^{(2n)}(x) \right\}.$$

Noter que $G_n(0)$ et $G_n(1)$ sont des entiers d'après ce que l'on a dit que les dérivées de f et le fait que $b^n\pi^2$ est un entier. On calcule la dérivée de $H_n(x) = G_n'(x)\sin\pi x - \pi G_n(x)\cos\pi x$, et on trouve facilement

$$H_n'(x) = b^n\pi^{2n+2}(\sin\pi x) f_n(x) = \pi^2 a^n (\sin\pi x) f_n(x),$$

donc en intégrant $\pi^{-1}H_n'$ entre 0 et 1 on trouve

$$\pi \int_0^1 a^n (\sin\pi x) f_n(x)dx = \left[\frac{G_n'(x)\sin\pi x}{\pi} - G_n(x)\cos\pi x \right]_0^1 = G_n(0) + G_n(1),$$

car $\cos 0 = 1$, $\sin 0 = 0$ et $\cos\pi = -1$, $\sin\pi = 0$ (ce sont donc ces propriétés caractéristiques de π qui sont utilisées). En particulier ceci est un entier. Cependant d'après l'encadrement (10.3.0) pour $f_n(x)$ on a

$$0 < \pi \int_0^1 a^n (\sin\pi x) f_n(x)dx < \frac{\pi a^n}{n!} < 1$$

si n est assez grand.

Troisième partie

Mathématiques et réel.

"La philosophie est écrite dans cet immense livre qui se tient ouvert devant nos yeux, je veux dire l'Univers, mais on ne peut le comprendre si l'on ne s'applique d'abord à en comprendre la langue et à connaître les caractères avec lesquels il est écrit. Il est écrit dans la langue mathématique et ses caractères sont les triangles, des cercles et autres figures géométriques, sans le moyen desquels il est humainement impossible d'en comprendre un mot."
(Galilée, *L'Essayeur*, trad. C. Chauviré, Paris, Les Belles Lettres, 1980, p. 141)

"Il faut avouer [...] que les géomètres abusent quelquefois de cette application de l'algèbre à la physique. Au défaut d'expériences propres à servir de base à leur calcul, ils se permettent des hypothèses, les plus commodes à la vérité qu'il leur est possible, mais souvent très éloignées de ce qui est réellement dans la nature."
(D'Alembert, *Discours préliminaire de L'Encyclopédie*)

[Les mathématiques constituent] une machine à broyer les esprits. La subversion mathématique est une des plus dangereuses. Elle tend à l'acceptation inconditionnelle d'un langage abstrait sans rapport avec le réel, qui prépare à merveille à la langue de bois et aux diktats de l'idéologie.
(Tiré d'un tract d'extrême droite, *Instruction nationale*.)

Le périmètre, la pratique et l'utilisation des mathématiques ont beaucoup varié à travers les siècles. De même pour la relation des mathématiques au monde. Le *Petit Robert* définit les mathématiques comme

> *"l'ensemble des sciences qui ont pour objet [...] l'étude des êtres abstraits [...], ainsi que les relations qui existent entre eux."*

Il semblerait donc y avoir une différence fondamentale entre mathématiques et réel qui, schématiquement, relèverait de l'opposition *abstrait/concret* : d'un côté une science de l'abstrait, une création de l'esprit et de l'autre la vie quotidienne, la vie de ce monde. D'une certaine manière, cette opposition est présente dès le début des mathématiques comme nous les connaissons, et il est très intéressant de voir comment l'équilibre entre les deux termes en jeu se déplace suivant les époques.

Voici comment Simplicius formule une pensée de Platon :

> *"Platon admet en principe que les corps célestes se meuvent d'un mouvement circulaire, uniforme et constamment régulier [c'est-à-dire de même sens] ; il pose alors aux mathématiciens ce problème : quels sont les mouvements circulaires, uniformes et parfaitement réguliers qu'il convient de prendre pour hypothèses, afin que l'on puisse sauver les apparences présentées par les planètes ?"*

(Simplicii, *In Aristotelis quatuor libros de Coelo commentaria*, Ed. Karsten, p. 219, col. a et p. 221, col. a.)

Il est demandé de développer une description mathématique, mieux géométrique, du mouvement des planètes (et plus généralement des astres). Il faut noter qu'à l'époque les phénomènes terrestres étaient l'objet d'une Physique (celle d'Aristote), qui était qualitative et qui ne faisait pas appel aux mathématiques. En fait on considérait que le monde des cieux était régi par des lois différentes de celles qui régissaient le monde sublunaire. De plus, avec l'Astronomie, on ne prétendait pas expliquer le mouvement des planètes : ce qui importait était de *sauver les apparences*, de faire en sorte de pouvoir décrire tous les phénomènes astronomiques observés le plus précisément possible. Au contraire la Physique était censée donner une explication de la nature des phénomènes terrestres. C'est ainsi que virent le jour plusieurs modèles géométriques décrivant le mouvement des planètes, qui étaient basés sur des principes différents.

Au début, les modèles posaient tous la Terre au centre (Ptolémée) et ne différaient que dans la manière d'organiser les mouvement circulaires autour de ce centre, seuls mouvements permis par principe (utilisation d'épicycles ou d'excentriques...). Puis, Copernic, en reprenant une idée attribuée par Archimède à Aristarque de Samos, met la Terre en mouvement dans son célèbre *Sur les révolutions des orbes célestes*, paru en 1543, année de sa mort. On pourrait se demander, comme A. Piccolomini, contemporain de Copernic, "par manière de digression, si les suppositions imaginées par les Astrologues [sic] pour sauver les apparences des planètes ont leur fondement en la vérité de la Nature". C'est-à-dire, est-ce que les uns pensaient que les planètes se trouvent vraiment accrochées à des sphères avec une réalité quelconque et les autres pensaient-ils vraiment que la Terre se meut d'un (triple) mouvement ? La réponse à cette question est difficile à donner en ce qui concerne Copernic lui-même, mais on peut dire, que la réponse pour la plupart des contemporains était certainement négative. Les arguments développés pour justifier cette réponse sont tout à fait compréhensibles. Des arguments logiques comme : vu que des descriptions différentes/concurrentes existaient, aucune ne pouvait prétendre à capturer l'essence des phénomènes sur la base du simple fait qu'elle sauve les apparences (vu que toutes le font). Des arguments physiques comme : la Terre ne peut pas avoir un triple mouvement, parce que "selon les philosophes/physiciens un corps simple unique a droit à un seul mouvement".

Copernic avait utilisé sa description pour calculer des tables astronomiques plus précises que celles qui existaient, et ce fait à lui seul suffisait, lui semblait-il, à justifier que l'on s'intéresse à ses hypothèses (ici à prendre au sens de fictions). D'ailleurs, les calculs basés sur le système copernicien permirent au Pape Grégoire XIII d'accomplir, en 1582, la réforme du calendrier. Une telle réforme était absolument nécessaire pour arriver à calculer correctement la date de Pâques et d'autres festivités. On sait que l'Écriture affirme que la Terre est immobile, le Pape ne pouvait donc faire autre chose que de considérer le mouvement de la Terre comme une fiction pratique.

La situation a changé de manière radicale avec Galilée. Pour faire simple, on peut dire que Galilée voulait attacher au modèle copernicien une plus grande réalité. Mais il ne s'agissait pas seulement d'affirmer que ce modèle, dans la version de Kepler, capturait la vraie nature des mouvements des planètes. Galilée a importé sur Terre, au sein de la Physique, les méthodes et les résultats de l'Astronomie ! [16]

[16] Nul ne doute que Galilée était un homme de génie, mais l'étude des notes manuscrites, qu'il nous a laissées montre que "C'est une erreur de penser qu'il fit dès le départ l'hypothèse que les mathématiques gouvernent la nature et que la physique doit s'y conformer ; en fait, les mathématiques se sont graduellement imposées à lui dans la question épineuse du changement littéralement continu" (S. Drake, *Galilée*, Actes Sud, 1987).

Avec l'énoncé du principe d'inertie on pouvait commencer à entrevoir que les mêmes lois régissent les phénomènes astronomiques et les phénomènes sublunaires. Du coup, ces lois aspiraient à remplacer les "vérités éternelles", contenues dans l'Écriture et dans les œuvres d'Aristote, chères aux Scolastiques. On connaît les déboires judiciaires, que son attitude effrontée a causé à Galilée, qui, en 1632, a dû se rétracter suite à une condamnation d'hérésie par le tribunal de l'Inquisition. Pour bien comprendre cette condamnation il faut se demander quel était l'enjeu (nous passons sur les questions de pouvoir). C'était au fond la question qui nous intéresse ici : la relation entre mathématiques et réel (au sens fort). *Est-ce que, si on sait décrire, alors on a vraiment compris ?*

Depuis le 17ème siècle la mathématisation des sciences physiques a beaucoup progressé et a permis la description d'une énorme quantité de phénomènes sur la base d'un nombre très restreint de principes. Sans faire attention, on aurait presque écrit "a permis la *compréhension* des phénomènes", tant il est vrai que l'efficacité des descriptions mathématiques nous mène à oublier qu'il ne s'agit que de descriptions, de modèles. Mais le problème demeure : comment se fait-il que l'on puisse même donner une description des phénomènes avec les mathématiques ? Pour revenir à la discussion ci-dessus, soulignons qu'au départ les mathématiques étaient seulement censées sauver les apparences célestes, mais que finalement elles ont été amenées à jouer un rôle primordial dans la description de tous les phénomènes physiques : l'Astronomie des grecs a vécu plus longtemps que leur Physique. En fait, une partie du mystère réside dans le fait que souvent les mathématiciens, en suivant la logique de développement propre à leur discipline établissent des résultats, qui se révèlent utiles pour la description des phénomènes. Ainsi, lorsque Kepler affinait le modèle copernicien en utilisant des ellipses pour décrire les orbites des planètes, il profitait de l'étude des coniques faite par Apollonius. De même, dans leurs réflexions, Maxwell et Heisenberg ont mis en évidence la pertinence de calculs algébriques abstraits (théorie des quaternions et des matrices), dont personne n'aurait prévu l'apparition au coeur des théories physiques. (Comme nous le rappelle la citation de D'Alembert au début de cette introduction, tous les résultats mathématiques n'ont évidemment pas cette fortune.)

Dans cette partie, nous allons commencer par montrer comment utiliser quelques notions de base de la géométrie dans l'espace pour, par exemple, représenter le globe terrestre sur une carte. Nous donnons aussi quelques exemples de "géométrie qualitative", sans coordonnées, centrée plutôt sur l'étude des formes et de l'invariance de certaines propriétés (longueurs, incidence, orientation, ...). Dans un deuxième chapitre nous allons préciser ce qu'on entend par modélisation mathématique, en fournissant plusieurs exemples pour mettre en évidence la richesse de la démarche et le grand nombre de problèmes encore ouverts dans ce domaine. Ce faisant nous serons amenés à passer en revue différents types d'équations différentielles, dont nous donnerons quelques méthodes de résolution.

La cartographie et la modélisation partagent une problématique essentielle, au coeur de la relation entre mathématiques et réel. En effet

> *"Toute représentation cartographique [comme la modélisation] suppose un compromis entre la précision et la lisibilité, et donc des sacrifices."*

(R. Brunet, *La carte, mode d'emploi*, Fayard/Reclus, Paris, 1982, Chap. 16 "Généraliser ou modéliser", p. 51)

Pour clore cette troisième partie nous allons élaborer sur ce qui précède et donner un aperçu de ce que peut être l'activité mathématique. On se propose de "montrer des mathématiques" à travers l'étude d'une classe d'objets mathématiques, les courbes planes, dont les coniques sont un exemple.

Chapitre 11

Droites et plans de $\mathbf{R}^2$ ou $\mathbf{R}^3$

Le premier objectif de ce chapitre est d'introduire de manière concrète les premiers rudiments de l'algèbre linéaire en petites dimensions (en l'occurrence 2 et 3). On y verra, en situation, diverses méthodes introduites dans d'autres parties du cours, en particulier la *méthode du pivot de Gauss* développée dans la Partie IV, Chapt. 14. Les propriétés de l'ensemble des nombres réels (qui sera tantôt notre alphabet pour repérer les points, tantôt notre règle graduée pour calculer les distances ou mesurer les angles et les surfaces) joueront un rôle déterminant.

Les espaces vectoriels qui seront notre terrain de jeu seront donc le plan $\mathbf{R}^2$ et l'espace $\mathbf{R}^3$ que l'on envisagera sous leurs divers aspects (vectoriel, affine, euclidien). Mais on apprendra aussi au fil de ce chapitre à s'affranchir des coordonnées et à penser les êtres géométriques de manière non plus cartésienne (les points étant repérés à partir du choix *a priori* d'un repère), mais cette fois topologique (ce sont les formes géométriques des êtres qui cette fois entrent en jeu). Le plan ou le globe terrestre sont des mondes géométriques sur lesquels on sait s'orienter, ce qui n'est pas le cas du ruban de Mœbius ou de la bouteille sans fond ; c'est ici la forme qui joue un rôle fondamental, feuilles de papier, colle et ciseaux remplaçant maintenant notre mètre "réel" gradué.

Nous verrons aussi, au travers de quelques exemples, comment les droites du plan interviennent dans des problèmes concrets issus des mathématiques appliquées : comment par exemple chercher la corrélation entre deux informations, comment décoder les images fournies par un dispositif de scanner, ou comment accéder à une information inconnue de manière itérative à l'aide d'algorithmes construits sur des idées pythagoriciennes.

11.1 Le plan et l'espace et leurs structures respectives d'espaces vectoriels

11.1.a Plan vectoriel, plan affine

L'ensemble $\mathbf{R}^2$ des couples (x, y) de nombres réels hérite naturellement d'une structure de $\mathbf{R}$-*espace vectoriel* [1] ; c'est le *plan vectoriel*.

En effet, on peut équiper cet ensemble d'une loi d'addition interne

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

[1] Pour la notion générale d'espace vectoriel voir l'Annexe.

et définir une action externe de $\mathbf{R}$ sur $\mathbf{R}^2$ par

$$(\lambda, (x, y)) \in \mathbf{R} \times \mathbf{R}^2 \longmapsto \lambda \cdot (x, y) := (\lambda x, \lambda y)$$

de manière à ce que 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 $\mathbf{R}^2$;
- 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 $\mathbf{R}^2$ muni de l'addition une structure de *groupe commutatif* ou *groupe abélien* (du nom du mathématicien norvégien Nils Henrik Abel, 1802-1829). La structure de $\mathbf{R}$-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.

Les deux vecteurs $\vec{i} := (1, 0)$ et $\vec{j} := (0, 1)$ constituent la *base canonique* de $\mathbf{R}^2$ et les deux nombres réels x et y constituent les *coordonnées cartésiennes* (ce qualificatif est emprunté au philosophe et mathématicien français René Descartes, 1596-1650). 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* $\mathbf{R}^2$ ou de *point de* $\mathbf{R}^2$; 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 $\mathbf{R}^2$ constitue le *plan vectoriel*, $\mathbf{R}$-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 $\mathbf{R}^2$ pensé comme ensemble de points est le *plan affine*. Travailler avec le point de vue consistant à considérer les couples (x, y) de $\mathbf{R}^2$ 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 $\mathbf{R}^2$ 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 $\mathbf{R}^2$ 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. Comme le couple (x, y) peut aussi être repéré par son *affixe*, à savoir le nombre complexe $x+iy$, le plan $\mathbf{R}^2$ s'identifie à $\mathbf{C}$ et, sous cet angle, on parle encore de *plan complexe* à propos de $\mathbf{R}^2$. Notons que le choix d'un système d'unités de longueur sur les axes conditionne la visualisation des points.

11.1.b Espace vectoriel $\mathbf{R}^3$, espace affine $\mathbf{R}^3$

Addition et multiplication externe sur $\mathbf{R}^3$ sont définies de manière analogue et l'ensemble $\mathbf{R}^3$ muni de ces deux opérations hérite d'une structure de $\mathbf{R}$-espace vectoriel ; c'est *l'espace vectoriel* $\mathbf{R}^3$, dont les points constituent *l'espace affine* $\mathbf{R}^3$. 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 $\mathbf{R}^3$ est la base constituée des trois vecteurs $\vec{i} = (1, 0, 0)$, $\vec{j} = (0, 1, 0)$ 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.

11.2 Formes linéaires dans le plan ou l'espace

11.2.a Le cas du plan $\mathbf{R}^2$

Une *forme linéaire* sur le plan vectoriel $\mathbf{R}^2$ est une application $L : \mathbf{R}^2 \longrightarrow \mathbf{R}$ telle que

$$L(\lambda\vec{V_1} + \mu\vec{V_2}) = \lambda L(\vec{V_1}) + \mu L(\vec{V_2})$$

pour tout choix de $\vec{V_1}, \vec{V_2}$ dans $\mathbf{R}^2$ et pour tout choix de nombres réels λ et μ. En utilisant la linéarité, on voit que, si L est une forme linéaire sur $\mathbf{R}^2$,

$$L((x, y)) = xL((1, 0)) + yL((0, 1)) = ax + by\,,$$

où $a := L((1, 0))$ et $b := L((0, 1))$. On notera pour abréger $L(x, y) = L((x, y))$.

Si $L : (x, y) \longmapsto ax + by$ est une forme linéaire sur $\mathbf{R}^2$, les vecteurs (x, y) dont le dispositif physique matérialisant L ne rend pas compte sont les vecteurs (x, y) tels que

$$ax + by = 0\,;$$

si L n'est pas l'application identiquement nulle (c'est-à-dire si a et b ne sont pas tous les deux nuls), le sous-ensemble du plan défini par

$$D_{a,b} := \{(x, y) \in \mathbf{R}^2 \,:\, ax + by = 0\}$$

est une *droite vectorielle* du plan dite *noyau* de L. Notons que si deux formes linéaires non nulles sur $\mathbf{R}^2$ sont proportionnelles (*i.e* $L_1 = \lambda L_2$ avec $\lambda \in \mathbf{R}^*$), elles ont même noyau et génèrent donc la même droite vectorielle. Si l'on utilise les nombres complexes et que l'on écrive $a + ib = re^{i\theta}$, cette droite vectorielle est définie par

$$x\cos\theta + y\sin\theta = 0\,.$$

On voit ainsi que l'ensemble des droites vectorielles du plan est en correspondance avec l'ensemble des nombres complexes de module 1.

Deux points distincts $M_1 = (x_1, y_1)$ et $M_2 = (x_2, y_2)$ de $\mathbf{R}^2$ déterminent de manière unique un sous-ensemble de $\mathbf{R}^2$, la *droite affine* qui les joint ; cette droite affine D_{M_1,M_2} est par définition l'ensemble des points (x, y) de la forme

$$D_{M_1,M_2} := \{(x, y) \in \mathbf{R}^2 \ : \ (x, y) = (x_1 + t(x_2 - x_1), y_1 + t(y_2 - y_1)) \text{ avec } \ t \in \mathbf{R}\} \, . \qquad (*)$$

Par deux points distincts du plan passe donc une et une seule droite affine. La description de D_{M_1,M_2} sous la forme $(*)$ est une *représentation sous forme paramétrique* de la droite affine D_{M_1,M_2}. Notons que cette représentation paramétrique dépend ici encore d'un seul paramètre (ici t) ; le nombre de degrés de liberté nécessaires à fixer pour définir un point de cette droite affine est donc égal à 1, ce qu'on exprime en disant que cette droite affine est *de dimension* 1.

Si L est l'unique forme linéaire annulant le vecteur $(x_2 - x_1, y_2 - y_1)$, on peut aussi représenter la droite D_{M_1,M_2} par

$$D_{M_1,M_2} = \{(x, y) \in \mathbf{R}^2 \, ; \, L(x, y) = L(x_1, y_1)\} \, ,$$

soit

$$D_{M_1,M_2} = \{(x, y) \in \mathbf{R}^2 \, ; \, (y_2 - y_1)(x - x_1) - (x_2 - x_1)(y - y_1) = 0\} \, .$$

Cette représentation est dite *représentation cartésienne* de la droite affine D_{M_1,M_2}.

Une droite affine D de $\mathbf{R}^2$ peut donc aussi être donnée par une représentation cartésienne

$$D := \{(x, y, z) \in \mathbf{R}^3 \, ; \, ax + by = c\} \, ,$$

où $(a, b) \in \mathbf{R}^2 \setminus \{(0, 0)\}$ et $c \in \mathbf{R}$; notons que (a_1, b_1, c_1) et (a_2, b_2, c_2) définissent la même droite affine du plan si et seulement si les deux vecteurs (a_1, b_1, c_1) et (a_2, b_2, c_2) sont colinéaires dans $\mathbf{R}^3$.

Nous reviendrons sur l'étude de l'ensemble des droites affines du plan $\mathbf{R}^2$ au paragraphe 11.4 de ce chapitre.

11.2.b Le cas de l'espace $\mathbf{R}^3$

Passons maintenant au cas de l'espace : tout ce que nous avons dit dans la section précédente se transcrit au cadre de l'espace $\mathbf{R}^3$. Se donner une forme linéaire sur $\mathbf{R}^3$ (c'est-à-dire une application L de $\mathbf{R}^3$ dans $\mathbf{R}$ telle que, si $\vec{V_1}$ et $\vec{V_2}$ sont deux vecteurs de $\mathbf{R}^3$, λ et μ deux nombres réels, on ait $L(\lambda\vec{V_1} + \mu\vec{V_2}) = \lambda L(\vec{V_1}) + \mu L(\vec{V_2})$) revient à se donner les images $a = L((1, 0, 0))$, $b = L((0, 0, 1))$, $c = L((0, 0, 1))$ des trois vecteurs $\vec{i}$, $\vec{j}$, $\vec{k}$ de la base canonique. Si L est une telle forme linéaire sur $\mathbf{R}^3$, dont l'action sur $\mathbf{R}^3$ est donc définie par

$$L(x, y, z) = ax + by + cz \, ,$$

les points (ou vecteurs) (x, y, z) de l'espace $\mathbf{R}^3$ dont le dispositif physique matérialisant L ne rend pas compte sont les points (x, y, z) tels que

$$ax + by + cz = 0 \, ;$$

si L n'est pas l'application nulle (c'est-à-dire si les trois nombres a, b, c ne sont pas tous les trois nuls), ce sous-ensemble de $\mathbf{R}^3$ est un *plan vectoriel*, dit *noyau* de la forme linéaire L. Notons encore que si deux formes linéaires non nulles sont proportionnelles (*i.e* $L_1 = \lambda L_2$ avec $\lambda \in \mathbf{R}^*$), elles ont même noyau et génèrent donc le même plan vectoriel.

Deux points distincts $M_1 = (x_1, y_1, z_1)$ et $M_2 = (x_2, y_2, z_2)$ de $\mathbf{R}^3$ déterminent de manière unique un sous-ensemble de $\mathbf{R}^3$, la *droite affine* qui les joint ; cette droite affine D_{M_1,M_2} est par définition l'ensemble

$$\{(x, y, z) \in \mathbf{R}^3 \ : \ (x, y, z) = (x_1 + t(x_2 - x_1), y_1 + t(y_2 - y_1)), z_1 + t(z_2 - z_1)) \text{ avec } \ t \in \mathbf{R}\}$$

Par deux points distincts de l'espace passe donc une et une seule droite affine. La description de D_{M_1,M_2} sous cette dernière forme est une *représentation sous forme paramétrique* de la droite affine D_{M_1,M_2}. Notons encore que cette représentation paramétrique dépend ici encore d'un seul paramètre (ici t) ; le nombre de degrés de liberté nécessaires à fixer pour définir un point de cette droite affine est donc égal à 1, ce qu'on exprime en disant que cette droite affine est *de dimension* 1.

Si $\vec{V_1} = (x_1, y_1, z_1)$ et $\vec{V_2} = (x_2, y_2, z_2)$ sont deux vecteurs non colinéaires de $\mathbf{R}^3$, il existe une unique forme linéaire L telle que $L(\vec{V_1}) = L(\vec{V_2}) = 0$. On vérifiera en exercice que cette forme linéaire est donnée par

$$L(x, y, z) = (y_1 z_2 - y_2 z_1)x + (z_1 x_2 - x_1 z_2)y + (x_1 y_2 - x_2 y_1)z\,;$$

le noyau de L est le plan vectoriel défini aussi comme

$$\{(x, y, z) \in \mathbf{R}^3 \ : \ (x, y, z) = t\vec{V_1} + sV_2 \text{ avec } (t, s) \in \mathbf{R}^2\}\,,$$

représentation que l'on appelle *représentation paramétrique* du plan vectoriel. Notons qu'il s'agit ici d'un sous-espace vectoriel de dimension 2 (il y a deux degrés de liberté, matérialisés ici par t et s).

En particulier, si $M_1 = (x_1, y_1, z_1), M_2 = (x_2, y_2, z_2), M_3 = (x_3, y_3, z_3)$ sont trois points du plan tels que les vecteurs

$$\vec{V_1} := (x_2 - x_1, y_2 - y_1, z_2 - z_1)\,,\ \vec{V_2} := (x_3 - x_1, y_3 - y_1, z_3 - z_1)$$

ne soient pas colinéaires (ce qui revient à dire que les trois points ne sont pas sur une même droite affine de l'espace), les vecteurs $\vec{V_1}$ et $\vec{V_2}$ définissent un plan vectoriel $\Pi(\vec{V_1}, \vec{V_2})$ et l'ensemble des points (x, y, z) de $\mathbf{R}^3$ tels que $(x - x_1, y - y_1, z - z_1)$ soit dans $\Pi(\vec{V_1}, \vec{V_2})$ est par définition le plan affine passant par M_1, M_2, M_3. Ce plan affine est donc défini sous forme paramétrique (avec cette fois deux paramètres) par

$$\Pi_{M_1,M_2,M_3} = \{(x, y, z) \in \mathbf{R}^3 \ : \ (x, y, z) = (x_1, y_1, z_1) + t\vec{V_1} + s\vec{V_2} \text{ avec } t, s \in \mathbf{R}^2\}\,.$$

Si L est l'unique forme linéaire annulant $\vec{V_1}$ et $\vec{V_2}$, on peut aussi représenter $\Pi_{M_1,M_2,M3}$ par

$$\Pi_{M_1,M_2,M3} = \{(x, y, z) \in \mathbf{R}^3 \ : \ L(x, y, z) = L(x_1, y_1, z_1)\}\,,$$

soit

$$\Pi_{M_1,M_2,M_3} = \\ \{(x, y, z)\,;\, A_{M_1,M_2,M_3}(x - x_1) + B_{M_1,M_2,M_3}(y - y_1) + C_{M_1,M_2,M_3}(z - z_1) = 0\}$$

avec

$$\begin{aligned} A_{M_1,M_2,M_3} &:= (y_2 - y_1)(z_3 - z_1) - (z_2 - z_1)(y_3 - y_1) \\ B_{M_1,M_2,M_3} &:= (z_2 - z_1)(x_3 - x_1) - (x_2 - x_1)(z_3 - z_1) \\ C_{M_1,M_2,M_3} &:= (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1) \end{aligned}$$

(on fera en exercice toutes les vérifications). Cette représentation est dite *représentation cartésienne* de l'unique plan affine noté Π_{M_1,M_2,M_3} passant par les trois points M_1, M_2, M_3.

Un plan affine de $\mathbf{R}^3$ peut donc aussi être donné par une représentation cartésienne

$$\Pi := \{(x, y, z) \in \mathbf{R}^3\,;\, ax + by + cz = d\}\,,$$

où $(a, b, c) \in \mathbf{R}^3 \setminus \{(0, 0, 0)\}$ et $d \in \mathbf{R}$; notons que (a_1, b_1, c_1, d_1) et (a_2, b_2, c_2, d_2) définissent le même plan affine de l'espace si et seulement si les deux vecteurs (a_1, b_1, c_1, d_1) et (a_2, b_2, c_2, d_2) sont colinéaires dans $\mathbf{R}^4$.

Nous reviendrons sur l'étude des droites et plans affines de l'espace $\mathbf{R}^3$ au paragraphe 11.7 de ce chapitre.

11.3 Comment cartographier la surface du globe terrestre ?

Cartographier le globe terrestre (qui n'est pas un univers plan) a posé problème depuis que l'on a réalisé la rotondité de la terre ; cette question a suscité nombre d'idées clef sous-tendant la géométrie moderne ; la notion d'*infini comme limite*, inhérente à l'analyse mathématique du réel, transparaît elle aussi dans ce type de problème.

Une des manières de cartographier la surface du globe terrestre est d'opérer ce que l'on appelle une *projection stéréographique* depuis un des pôles (par exemple le pôle Nord). La projection stéréographique depuis le pôle Nord est représentée sur la figure ci-dessous, où, pour fixer les idées, on a supposé le globe sous-ensemble de $\mathbf{R}^3$, avec pour centre le point $(0,0,0)$ et pour pôle Nord le point $(0,0,1)$. Un point M de l'hémisphère Sud est projeté sur le point m obtenu comme le point d'intersection du plan vectoriel $\{z=0\}$ avec la droite joignant M au pôle Nord ; un point M' de l'hémisphère Nord (distinct du pôle Nord) est projeté au point m' intersection du plan vectoriel $\{z=0\}$ avec la droite joignant le pôle Nord à M.

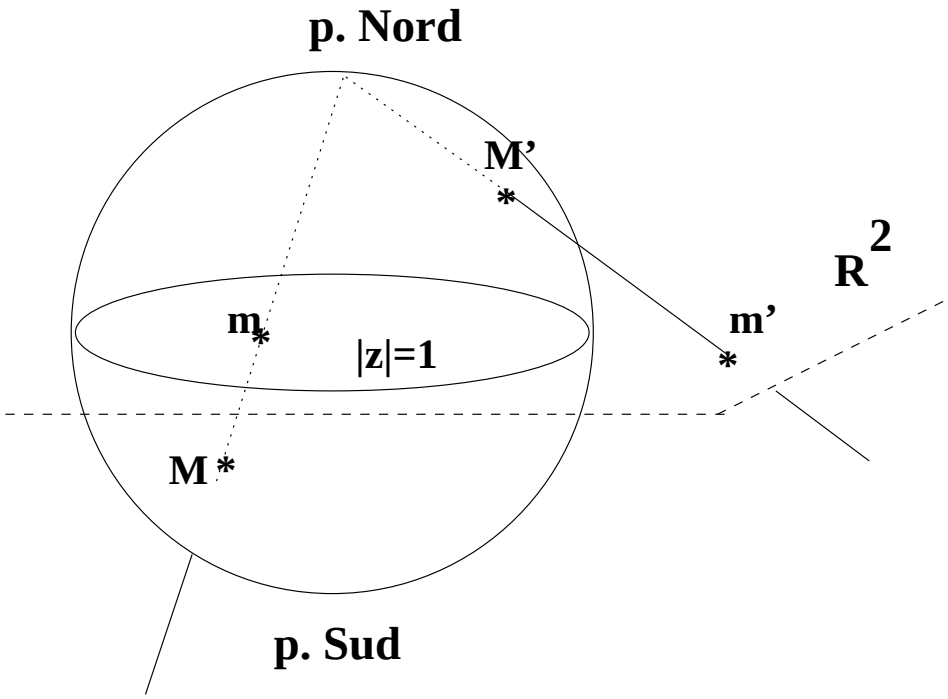


FIG. 11.1 – Le globe terrestre et la projection stéréographique depuis le pôle Nord

Exercice 11.3.1 *Vérifier (en utilisant par exemple le théorème de Thalès que l'on rappellera) que la projection stéréographique du point (X,Y,Z) de la surface du globe terrestre (le centre de la terre étant l'origine $(0,0,0)$, le pôle Nord le point $(0,0,1)$) est le point du plan ayant pour affixe le nombre complexe*

$$z = \frac{X+iY}{1-Z}\,.$$

Inversement, vérifier que l'antécédent du point (x,y) du plan via la projection stéréographique depuis

le pôle Nord est le point de la surface du globe terrestre de coordonnées

$$X = \frac{2x}{1+x^2+y^2}\,,\ Y = \frac{2y}{1+x^2+y^2}\,,\ Z = \frac{x^2+y^2-1}{x^2+y^2+1}\,.$$

La projection stéréographique est, on le sait bien, entachée de défauts : le Groëndland par exemple voit son territoire considérablement "élargi" après projection stéréographique depuis le pôle Nord, tandis que le continent antarctique se voit lui peu déformé. Nous quantifierons ceci dans la sous-section suivante.

La vision du plan "remonté" sur la surface du globe terrestre nous servira essentiellement d'une part à appréhender la notion d'infini dans un univers plan, d'autre part à comprendre pourquoi *cercles* et *droites* constituent une famille de sous-ensembles du plan qu'il s'avère intéressant d'unifier.

11.3.a Droites et cercles du plan

Coupons la surface du globe terrestre par un plan affine dont une équation cartésienne est

$$ax + by + cz = d\,.$$

L'intersection de la surface du globe terrestre avec un tel plan affine est, lorsqu'elle n'est pas vide (à propos, quand est-on sûr qu'elle n'est pas vide ?), un cercle tracé sur cette surface. Ce cercle peut ou non passer par le pôle Nord du globe terrestre : ceci se produit si et seulement si $c = d$.

L'image par la projection stéréographique du cercle tracé sur la surface du globe terrestre lorsqu'on l'intersecte par ce plan affine (après exclusion éventuelle du pôle Nord) est le sous-ensemble du plan $\mathbf{R}^2$ défini comme

$$\{(x,y) \in \mathbf{R}^2 \,:\, 2ax + 2by + c(x^2+y^2-1) = d(x^2+y^2+1)\}$$

on remarque que si $c = d$, ce sous-ensemble est la droite affine d'équation cartésienne

$$ax + by - c = 0\,;$$

en revanche, si $c - d \neq 0$, le sous-ensemble est défini comme

$$\{(x,y,z) \in \mathbf{R}^3 \,:\, (x^2+y^2)(c-d) + 2ax + 2by = c+d\}\,.$$

Exercice 11.3.2 *Quel est cet ensemble ? En utilisant la formule du trinôme, vérifier que c'est soit l'ensemble vide (dans quel cas ?), soit un cercle dont on précisera le centre et le rayon. En utilisant des disques tracés sur la surface du globe terrestre et proches du pôle Nord (sans le contenir), expliquez, en regardant ce que devient l'image par la projection stéréographique (c'est, d'après ce que l'on a vu, un disque du plan) pourquoi la surface de territoires tels le Groëndland se trouve exagérément "gonflée" après projection stéréographique (ce que l'on observe sur une planisphère quand bien même on a affaire là à des versions corrigées plus subtiles que la simple projection stéréographique, par exemple la projection de Mercator).*

Les cercles ou droites du plan sont donc simplement les images par projection stéréographique depuis le pôle Nord des cercles tracés sur la surface du globe terrestre. Les droites du plan correspondent aux cercles tracés sur la surface du globe et passant par le pôle Nord, les cercles du plan correspondent aux cercles tracés sur le globe terrestre et ne passant pas par le pôle Nord. Ainsi, une géodésique entre deux points de la surface du globe (c'est à dire le chemin le plus court tracé sur la surface du globe et

joignant ces deux points, qui est en l'occurrence ici un arc de cercle) devient soit un arc de cercle, soit un segment de droite dans le plan.

Cercles et droites du plan sont donc à considérer *dans une même famille*, bien qu'il y ait une distinction significative du point de vue algébrique entre les deux : la représentation cartésienne d'une droite affine est donnée, on l'a vu, par un polynôme de degré total 1 en x, y, celle d'un cercle par un polynôme de degré total 2 en (x, y) ; on peut toutefois contourner cette ambigüité en remarquant que $(ax+by-c)^2 = 0$ est aussi une représentation cartésienne pour la droite affine $\{(x, y)\,;\, ax + by = c\}$!

11.3.b Utilisation des nombres complexes : translations, similitudes, homographies

Une *translation* du plan $\mathbf{R}^2$ se traduit, si le plan est pensé comme le plan complexe, comme une application du type

$$z \mapsto z + b\,,$$

où b est un nombre complexe. Si $b = 0$, il s'agit d'une application linéaire de $\mathbf{R}^2$ dans $\mathbf{R}^2$ particulière, l'identité.

Une *similitude directe* du plan est une application du type

$$z \mapsto az + b$$

où a et b sont deux nombres complexes ; si $b = 0$, il s'agit d'une *application affine* de $\mathbf{R}^2$ dans $\mathbf{R}^2$, c'est-à-dire une application f de $\mathbf{R}^2$ dans lui-même couplée avec une unique application linéaire L (ici $z \mapsto az$) telle que $f(M) = M + L(\overrightarrow{OM})$ pour tout point M du plan ; le fait qu'une application affine non nulle de $\mathbf{R}^2$ dans $\mathbf{R}^2$ soit une similitude directe du type $z \mapsto az + b$ équivaut d'ailleurs au fait que pareille application affine respecte les angles orientés des figures : c'est en effet la composée d'une homothétie de centre $(0, 0)$ et de rapport r, d'une rotation d'angle θ si $a = re^{i\theta}$, puis d'une translation par b ; comme ces opérations conservent les angles orientés des figures ne modifie pas les figures (rotations et translation ne font que déplacer les objets, l'homothétie les dilate ou les contracte), une similitude a le mérite de respecter les angles orientés des figures et donc de ne pas les déformer (autrement que par dilatation, contraction ou rotation).

Les transformations envoyant un sous-ensemble du plan complexe dans le plan complexe et respectant les angles orientés des figures se rencontrent fréquemment car la nature a souvent tendance à gérer les phénomènes physiques "à l'économie", entre autres en ne déformant pas les figures (en en préservant par exemple les angles) : c'est un avatar du *principe de moindre action*.

Il existe de telles transformations qui ne soient pas des transformations affines. Il est ainsi une autre transformation importante du plan complexe (privée de l'origine) dans lui-même qui préserve elle aussi les angles orientés des figures, même si ce n'est plus une application affine : c'est la transformation I qui à $z \in \mathbf{C}^*$ associe $1/z$. Celle-ci transforme un cercle passant par l'origine en une droite, un cercle ne passant pas par l'origine en un cercle (on fera l'exercice). C'est l'*inversion* de pôle $(0, 0)$ et de puissance 1. Elle permet de ramener l'étude de ce qui se passe dans le plan au voisinage de l'infini (pour un phénomène physique) à l'étude d'un nouveau phénomène, mais cette fois près de l'origine, donc dans l'univers fini.

Exercice 11.3.3 *Représenter l'image par inversion I de pôle $(0, 0)$ et de puissance 1 du triangle isocèle de sommets $(1, 0), (2, 2), (0, 1)$. Vérifier la conservation attendue des angles orientés.*

Une *homographie* du plan complexe est une application du type

$$z \mapsto \frac{az + b}{cz + d}\,,$$

avec c et d non tous les deux nuls (éventuellement non définie pour $z = -d/c$ si $c \neq 0$). Comme une homographie se décompose en la composition d'une similitude directe, d'une inversion, puis d'une similitude directe, elle préserve globalement l'importante famille des cercles-droites introduite dans la sous-section précédente. Ici encore, il y a conservation des angles orientés des figures.

Exercice 11.3.4 *Décomposer explicitement l'homographie*

$$z \mapsto \frac{az+b}{cz+d}$$

sous la forme de la composée d'une similitude directe, de l'inversion, puis d'une similitude directe.

11.3.c Une première vision de l'infini de $\mathbf{R}^2$: un point

Si l'on reprend l'image du plan en correspondance avec le globe terrestre privé du pôle Nord par projection stéréographique, on voit que ce pôle Nord joue le rôle de l'unique point qui manque à la surface du globe terrestre (donc au plan $\mathbf{R}^2$ qui lui correspond par projection stéréographique) pour en faire un univers qu'un mathématicien qualifie de *compact* (ou de "serré"), au sens où toute limite d'une suite de points de l'univers est encore dans l'univers et où celui-ci peut toujours être recouvert par un nombre fini de sous-ensembles de diamètre arbitrairement petit, ce qui le cas de la surface du globe terrestre, pôle Nord inclus.

On peut donc considérer le pôle Nord comme le *point à l'infini* du plan (que l'on note $\{\infty\}$). Une droite affine du plan se lit donc comme un cercle qui "se fermerait" au point à l'infini.

Si a, b, c, d sont des nombres complexes avec $c \neq 0$, l'homographie $z \mapsto \dfrac{az+b}{cz+d}$ peut être pensée comme une application de $\mathbf{R}^2 \cup \{\infty\}$ dans lui-même, envoyant le point ∞ en a/c et le point $-d/c$ en ∞. C'est même ainsi une application bijective.

Cette première vision de l'infini éclaire le fait que droites et cercles doivent être considérés "en famille" et qu'une homographie se doit d'être pensée comme une bijection du plan complexe avec son point à l'infini dans lui-même.

La droite réelle $\mathbf{R}$ peut aussi être pensée de cette manière comme image par projection stéréographique du cercle unité du plan privé du point $(1, 0)$, auquel cas, il y a un point à l'infini (correspondant au point $(0, 1)$) ; cependant, il est naturel aussi de considérer en analyse dans le contexte Mathématiques et Réel la *droite numérique achevée* $\mathbf{R} \cup \{-\infty, +\infty\}$ (il y a dans ce cas deux points à l'infini). La distinction cruciale entre $\mathbf{R}$ et $\mathbf{R}^2$ est que si l'on retire à $\mathbf{R}^2$ un disque fermé, on conserve un domaine "d'un seul tenant" (le mathématicien dira *connexe*) tandis que si l'on retire à $\mathbf{R}$ un intervalle $[\alpha, \beta]$, on obtient un ensemble constitué de deux parties d'un seul tenant ! Ceci explique pourquoi on peut se soucier d'ajouter deux (et non un) point à l'infini à la droite réelle, tandis qu'on ajoute sans équivoque un point à l'infini au plan réel. On verra plus loin (Sect. 11.4.b) comment dans le plan on peut aussi réaliser, comme les peintres de la Renaissance, une notion d'infini spécifique à chaque direction de ligne de fuite.

11.4 L'ensemble des droites affines du plan

11.4.a Un repérage "cartésien" : $ax + by + c = 0$

Une droite affine de $\mathbf{R}^2$ est donnée par son équation cartésienne

$$ax + by + c = 0\,,$$

où a, b sont deux nombres réels non tous les deux nuls et c est un nombre réel.

De plus, si (a_1, b_1, c_1) et (a_2, b_2, c_2) sont colinéaires, ces deux triplets induisent la même droite affine.

Deux droites affines d'équations cartésiennes respectives

$$a_1x + b_1y + c_1 = 0\,,\ a_2x + b_2y + c_2 = 0$$

sont *parallèles* si (a_1, b_1) et (a_2, b_2) sont deux vecteurs colinéaires non nuls du plan et si les vecteurs (a_1, b_1, c_1) et (a_2, b_2, c_2) ne sont pas colinéaires (si les vecteurs (a_1, b_1, c_1) et (a_2, b_2, c_2) sont colinéaires, c'est-à-dire s'il existe $\lambda \in \mathbf{R}^*$ tel que $(a_2, b_2, c_2) = \lambda(a_1, b_1, c_1)$, les deux droites en question sont dites *confondues*). Deux droites parallèles ne se coupent pas dans $\mathbf{R}^2$ (on s'en assurera en faisant l'exercice). Deux droites confondues ont tous leurs points en commun.

Exercice 11.4.1 *Comment deux droites parallèles du plan se remontent-elles par projection stéréographique inverse sur la surface du globe ? Où s'intersectent les deux images inverses ? Les deux images inverses se croisent-elles franchement ou ont-elles une tangente commune en leur unique point d'intersection ? Que se passe-t'il au niveau des intersections des images inverses si l'on perturbe les deux droites en deux droites voisines mais non parallèles ?*

En revanche, deux droites affines non parallèles ni confondues se coupent en un unique point du plan ; on dit qu'elles sont *sécantes*. On fera l'exercice en cherchant explicitement les coordonnées du point d'intersection, ce qui nous oblige à résoudre un système de deux équations à deux inconnues, ce que la méthode du pivot par exemple nous apprend à faire (c'est ici très simple), et l'on vérifiera que ce point d'intersection est le point de coordonnées

$$x = \frac{b_1c_2 - b_2c_1}{a_1b_2 - a_2b_1} \qquad y = \frac{a_2c_1 - a_1c_2}{a_1b_2 - a_2b_1}$$

Exercice 11.4.2 *Comment deux droites sécantes du plan se remontent-elles par projection stéréographique inverse sur la surface du globe ? En combien de points s'intersectent les deux images inverses ?*

11.4.b Une seconde vision de l'infini de $\mathbf{R}^2$: une "droite" à l'infini

Ce paragraphe est à prendre en compte comme un thème de lecture et de réflexion sur la notion d'infini et ses relations avec l'art pictural.

Considérons l'espace $\mathbf{R}^3$ privé de l'origine, dans lequel nous identifions dans une même classe (nous considérons comme "équivalents") tous les vecteurs colinéaires à un vecteur donné. Appelons $\mathbf{P}^2(\mathbf{R})$ cet ensemble de classes. La classe de $(0, 0, 1)$ est un point particulier Q de l'ensemble $\mathbf{P}^2(\mathbf{R})$; puisque l'ensemble des points de $\mathbf{P}^2(\mathbf{R})$ différents de Q est en correspondance avec l'ensemble des droites affines de $\mathbf{R}^2$ (on associe à une droite affine D la classe de (a, b, c), où $ax+by+c = 0$ est une équation cartésienne de D), il est naturel d'appeler ce point Q la *droite à l'infini* du plan $\mathbf{R}^2$. Ainsi, l'ensemble de toutes les droites affines de $\mathbf{R}^2$, droite à l'infini comprise, est-il cette fois en correspondance bijective avec notre ensemble $\mathbf{P}^2(\mathbf{R})$ que nous appellerons *plan projectif*.

Changeons notre fusil d'épaule. Si $(x, y) \in \mathbf{R}^2$, la classe de $(x, y, 1)$ est un point du plan projectif. On obtient d'ailleurs ainsi tous les points du plan projectif sauf ceux qui correspondent aux classes des points $(x, y, 0)$, avec $(x, y) \in \mathbf{R}^2 \setminus \{(0, 0)\}$. Ainsi le plan projectif contient-il le plan $\mathbf{R}^2$. Ce que nous avons appelé la droite à l'infini de $\mathbf{R}^2$ est précisément l'ensemble des points du plan projectif qui n'appartiennent pas à $\mathbf{R}^2$, c'est-à-dire les classes des points $(x, y, 0)$, avec $(x, y) \in \mathbf{R}^2 \setminus \{(0, 0)\}$.

On s'exercera à comprendre la construction du plan projectif en examinant la figure ci-dessous : l'univers plan $\mathbf{R}^2$ est artificiellement "élevé" à l'altitude $z = 1$ dans l'espace $\mathbf{R}^3$. On observe une correspondance entre les points du plan projectif et les lignes de fuite de $\mathbf{R}^3$ issues de l'origine, auquel cas les points de l'univers plan $\mathbf{R}^2$ correspondent aux lignes de fuite de $\mathbf{R}^3$ issues de l'origine et ne se trouvant pas dans le plan $z = 0$ (comme celle matérialisée par D sur la figure) ; en revanche chaque droite du plan $\{z = 0\}$ correspond, elle, à un point du plan projectif qui n'est pas un point de l'univers plan réel. Si $(x, y) \neq (0, 0)$, la ligne de fuite $L_{x,y}$ de $\mathbf{R}^3$ paramétrée par $t \mapsto (tx, ty, 0)$ peut s'interpréter comme le point à l'infini de $\mathbf{R}^2$ le long de la ligne de fuite $l_{x,y}$ (de $\mathbf{R}^2$ cette fois) dirigée par le vecteur (x, y) et représentée à l'altitude $z = 1$. L'infini du plan n'est plus un point unique comme celui que nous avions introduit par le biais de la projection stéréographique depuis le globe terrestre (en l'occurrence le pôle Nord depuis lequel s'opérait la projection stéréographique) ; il y a ainsi un point à l'infini pour chaque ligne de fuite l du plan (représenté à l'altitude $z = 1$) issue de l'origine.

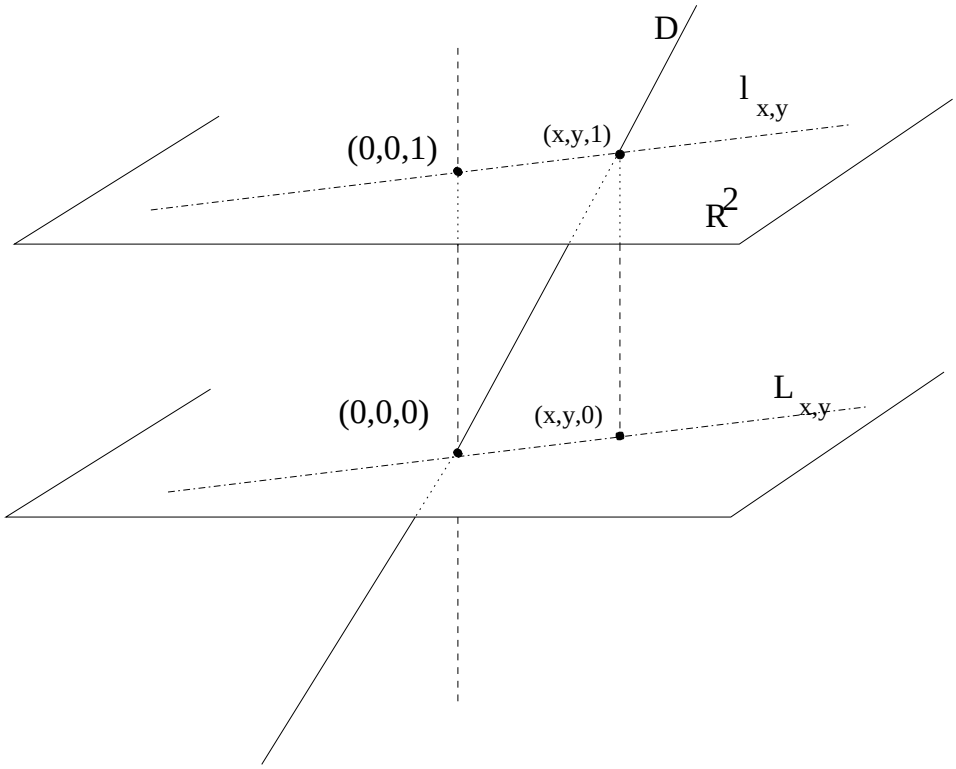


FIG. 11.2 – Le plan projectif : un point à l'infini pour chaque ligne de fuite

Dans ce contexte, deux droites affines parallèles du plan d'équations cartésiennes $ax + by + c_1 = 0$ et $ax + by + c_2 = 0$ (avec $c_1 \neq c_2$) se coupent au point correspondant précisément à la classe de $(-b, a, 0)$, classe qui est donc un point de la droite à l'infini de $\mathbf{R}^2$ (c'est un point du plan projectif, mais ce n'est plus un point du plan). Ainsi la droite à l'infini de $\mathbf{R}^2$ apparaît comme l'ensemble des points du plan projectif où se coupent les droites parallèles du plan $\mathbf{R}^2$, ce qui renvoie à l'art de la perspective et à la notion de *point de fuite*.

Ces idées mathématiques se sont élaborées (sous l'impulsion de Gérard Desargues, 1591-1661) après que

les artistes depuis la Renaissance (comme ci-dessous Canaletto dans ce tableau de la place Saint-Marc) aient développé de manière systématique l'usage de la perspective : les points du plan projectif peuvent aussi être considérés comme les points à l'infini des "lignes de fuite" issues de l'origine dans l'espace $\mathbf{R}^3$; ceux de la droite à l'infini de $\mathbf{R}^2$ peuvent eux, être considérés comme les points à l'infini des "lignes de fuite" issues de l'origine dans le plan $\mathbf{R}^2$.

L'imagerie informatique et le graphisme $3D$ ont aujourd'hui redonné un souffle nouveau à l'utilisation du concept de perspective et donc aux notions de plan ou d'espace projectif que nous avons tenté d'introduire ici à travers la matérialisation de l'idée d'infini dans le plan.

FIG. 11.3 – Les lignes de fuite chez Canaletto

11.4.c Droites du plan et trinômes $aX^2 + bX + c$: une correspondance inattendue

Le but de cette section (à lire à tête reposée) est de voir comment on peut penser différemment la géométrie en ne raisonnant plus de manière cartésienne avec les coordonnées comme vous aviez jusque là l'habitude de faire. On y voit aussi la magie des correspondances entre un monde algébrique (la classification des trinômes) et un univers géométrique (la visualisation du plan projectif).

Comme vous le savez depuis le lycée, les trinômes du second degré au plus du type $aX^2 + bX + c$ avec $(a, b, c) \in \mathbf{R}^3 \setminus \{(0, 0, 0)\}$ se rangent en plusieurs catégories suivant les valeurs du triplet (a, b, c). Ce paragraphe nous fournit l'occasion de nous remémorer cette importante classification.

– Lorsque $b^2 - 4ac < 0$, le trinôme $aX^2 + bX + c$ (qui est dans ce cas un vrai trinôme car a ne peut être nul, sinon on aurait $b^2 < 0$, ce qui est absurde) admet deux racines complexes conjuguées, une et une seule de ces racines se trouvant dans le demi-plan $\{z \in \mathbf{C}\,;\, \mathrm{Im}\, z > 0\}$.
– Lorsque $b^2 - 4ac = 0$, deux sous-cas peuvent se produire :
  • soit $a = b = 0$, auquel cas le trinôme se réduit au terme constant $c \neq 0$ et n'a pas de racine dans $\mathbf{C}$; en fait, on peut considérer qu'il y a une racine double, racine qui s'est échappé à l'infini de la droite réelle ;
  • soit $a \neq 0$, auquel cas le trinôme est un vrai trinôme ayant une racine double réelle ; notons que cette racine double est à la frontière du demi-plan $\{z \in \mathbf{C}\,;\, \mathrm{Im}\, z > 0\}$.
– Lorsque $b^2 - 4ac > 0$, deux sous-cas sont encore à envisager :
  • si $a \neq 0$, le trinôme (qui est un vrai trinôme) admet deux racines réelles distinctes formant une paire (on ne sait les identifier indépendamment sans utiliser de notion d'ordre et l'on doit les considérer en paire) ;
  • si $a = 0$ et $b \neq 0$, il n'y a plus qu'une racine réelle, l'autre s'étant "échappé" à l'infini (comme on le voit en perturbant légèrement a).

Parmi les homographies (transformations du plan complexe dans lui-même que nous avons introduit dans la Sect. 11.3.b), il en est une intéressante, l'application

$$z \mapsto \frac{z - i}{z + i}\,.$$

On s'exercera à montrer que cette application échange le demi-plan $\{z \in \mathbf{C}\,;\, \mathrm{Im}\, z \geq 0\}$ (auquel on a adjoint le point à l'infini de l'axe réel) et le disque $\{z \in \mathbf{C}\,;\, |z| \leq 1\}$.

D'autre part, deux trinômes $a_1X^2 + b_1X + c_1$ et $a_2X^2 + b_2X + c_2$ ont même ensemble de racines si et seulement si (a_1, b_1, c_1) et (a_2, b_2, c_2) sont des vecteurs non nuls de $\mathbf{R}^3$ colinéaires (on expliquera pourquoi). Puisque nous avons classifié les trinômes suivant leurs ensembles de racines, nous avons ainsi classifié en fait les points du plan projectif. On voit toute la force des correspondances.

Ainsi, il y a correspondance (on expliquera comment en réfléchissant sur la classification rappelée plus haut) entre les points (a, b, c) du plan projectif correspondant aux triplets (a, b, c) tels que $b^2 - 4ac \leq 0$ et les points du disque $\{z \in \mathbf{C}\,;\, |z| \leq 1\}$. Quant aux autres points, il sont en correspondance avec l'ensemble E des couples de points distincts du cercle unité, les couples (α, β) et (β, α) étant identifiés.

Comment peut-on visualiser ce dernier ensemble E (identification de points prises en compte) ? On va le faire avec du papier et une paire de ciseaux. Réfléchissons à ce qu'est le produit de deux cercles, comme nous y invite la figure ci-dessous. On le voit, c'est une chambre à air, paramétrée comme l'indique la figure suivante avec les deux angles θ et φ !

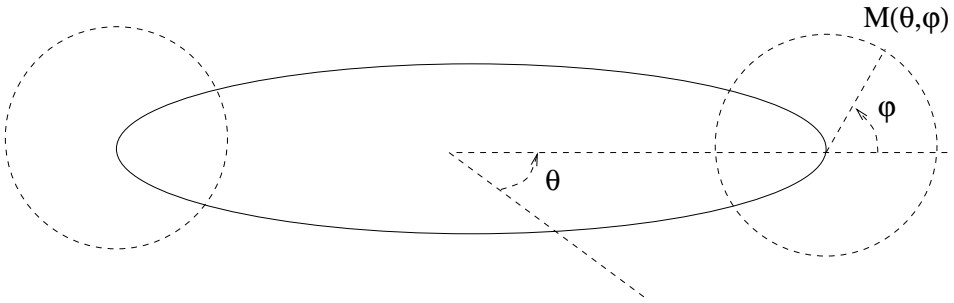


Fig. 11.4 – Une chambre à air paramétrée

Prenez maintenant une feuille de papier sur laquelle vous avez marqué la diagonale, roulez-la pour en faire un tube, fermez votre tube pour réaliser une chambre à air ; découpez ensuite la diagonale ; le modèle que vous cherchez à réaliser est le modèle obtenu à partir de votre demi-feuille de papier en identifiant les points m et M comme indiqué sur la figure suivante. On réalise, en découpant suivant le pointillé, puis en recollant en tenant compte des contraintes, ce que l'on appelle un *ruban de Mœbius* (du nom du géomètre allemand August Mœbius, 1790-1868, à qui l'on doit l'introduction de cette surface troublante sur laquelle on ne sait plus comment s'orienter !).

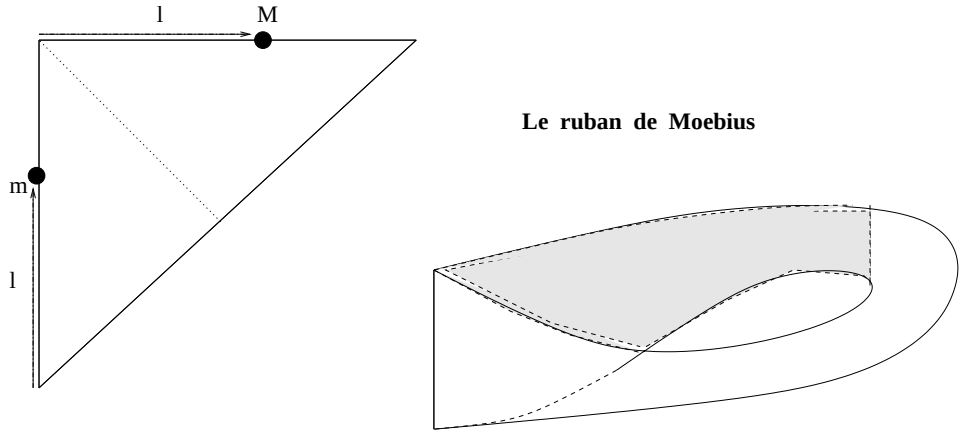


FIG. 11.5 – Réaliser un ruban de Mœbius

Nous sommes ainsi en mesure d'imaginer le plan projectif comme un disque fermé (correspondant aux trinômes tels que $b^2 - 4ac \leq 0$) auquel on aurait collé bord-à-bord un ruban de Mœbius (correspondant aux trinômes tels que $b^2 - 4ac > 0$) ; on verra au paragraphe suivant qu'autant l'on sait s'orienter dans le plan, autant cette manière de concevoir le plan projectif explique que l'on ne puisse s'y orienter !

11.4.d Qu'est-ce que savoir s'orienter dans le plan ?

Les bases $(\vec{V_1}, \vec{V_2})$ du plan se rangent en deux classes : celles pour lesquelles la quantité $x_1y_2 - x_2y_1$ (ou $\vec{V_1} := (x_1, y_1)$ et $\vec{V_2} := (x_2, y_2)$) est strictement positive, celles pour lesquelles cette même quantité est strictement négative (elle ne peut être nulle puisque $\vec{V_1}$ et $\vec{V_2}$ forment une base). Chacune de ces classes est ce que l'on appelle une *orientation* et il y a donc dans l'univers plan deux orientations possibles.

Supposons qu'on ait privilégié une orientation, c'est-à-dire choisi une de ces deux classes ; les bases de la classe choisie sont appelées *directes*, les autres *indirectes*. Lorsque l'on se déplace dans le plan, tel un point $M(t) = (x(t), y(t))$, t désignant le temps, le vecteur vitesse $\vec{V}(t) := d(\overrightarrow{OM})/dt = (x'(t), y'(t))$ (que l'on suppose non nul) nous montre le chemin à suivre à l'instant t ; tous les vecteurs $\vec{U}$ non colinéaires à $\vec{V}(t)$ et tels que $\{\vec{V}(t), \vec{U}\}$ soit une base directe (c'est-à-dire soient dans la classe que l'on a privilégié) pointent à notre gauche, les autres pointent vers notre droite. Nous savons ainsi à tout instant où est notre gauche, où est notre droite lorsque nous nous déplaçons dans le plan (pourvu que l'on ait choisi *a priori* une orientation). En ce sens, le plan est un univers où l'on sait s'orienter.

On essaiera de se convaincre que la surface du globe est aussi un univers sur lequel on peut s'orienter : pensez à le faire en repérant votre gauche et votre droite en fonction de de votre vecteur vitesse et du vecteur normal au globe au point où vous vous trouvez, cela vous rappellera certainement la *règle du bonhomme d'Ampère* que vous avez peut-être déjà manié en physique). Au contraire le ruban

de Mœbius (introduit dans la sous-section précédente) est un univers sur lequel on en est incapable (expliquez pourquoi en exercice).

11.5 Pythagore dans le plan

11.5.a Produit scalaire, angles et surfaces

Dans le plan $\mathbf{R}^2$, 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

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

La *distance euclidienne* entre deux points $M_1 := (x_1, y_1)$ et $M_2 := (x_2, y_2)$ du plan est par définition

$$d(M_1, M_2) := \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = \sqrt{\langle x_2 - x_1\,,\, y_2 - y_1 \rangle}\,.$$

Cette distance obéit aux quatre impératifs exigés d'une distance :
- c'est une fonction positive sur $\mathbf{R}^2 \times \mathbf{R}^2$;
- 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 existe d'autres distances dans le plan (c'est-à-dire des fonctions de l'ensemble $\mathbf{R}^2 \times \mathbf{R}^2$ à valeurs dans $[0, +\infty[$ se pliant à ces quatre exigences). En voici par exemple une : on pourrait par exemple appeler distance entre M_1 et M_2 la distance (dans $\mathbf{R}^3$) de leurs images réciproques par projection stéréographique (c'est la *distance cordale* que l'on calculera en exercice).

Mais la distance euclidiennne se trouve liée au produit scalaire par l'importante *formule de Pythagore* : 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

$$d^2(M_1, M_3) = d^2(M_1, M_2) + d^2(M_2, M_3) + 2\Big\langle (x_2 - x_1, y_2 - y_1)\,,\, (x_3 - x_2, y_3 - y_2) \Big\rangle\,; \qquad (\dagger)$$

si en particulier les deux vecteurs $\overrightarrow{M_1 M_2}$ et $\overrightarrow{M_2 M_3}$ 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

$$(x_1 y_2 - x_2 y_1)^2 + (x_1 x_2 + y_1 y_2)^2 = (x_1^2 + x_2^2)(y_1^2 + y_2^2)\,,$$

ou encore

$$\left(\frac{\langle (x_1, y_1)\,,\, (x_2, y_2) \rangle}{\sqrt{x_1^2 + y_1^2}\,\sqrt{x_2^2 + y_2^2}} \right)^2 + \left(\frac{x_1 y_2 - x_2 y_1}{\sqrt{x_1^2 + y_1^2}\,\sqrt{x_2^2 + y_2^2}} \right)^2 = 1\,.$$

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

$$\begin{aligned} \cos\theta &= \frac{\langle (x_1, y_1)\,,\, (x_2, y_2) \rangle}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}} \\ \sin\theta &= \frac{x_1 y_2 - x_2 y_1}{\sqrt{x_1^2 + y_1^2}\sqrt{x_2^2 + y_2^2}}\,. \end{aligned}$$

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 $\mathbf{R}^2$ dans $\mathbf{R}^3$ 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

$$\vec{V_1} \wedge \vec{V_2} := (0, 0, x_1y_2 - x_2y_1)\,.$$

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 :

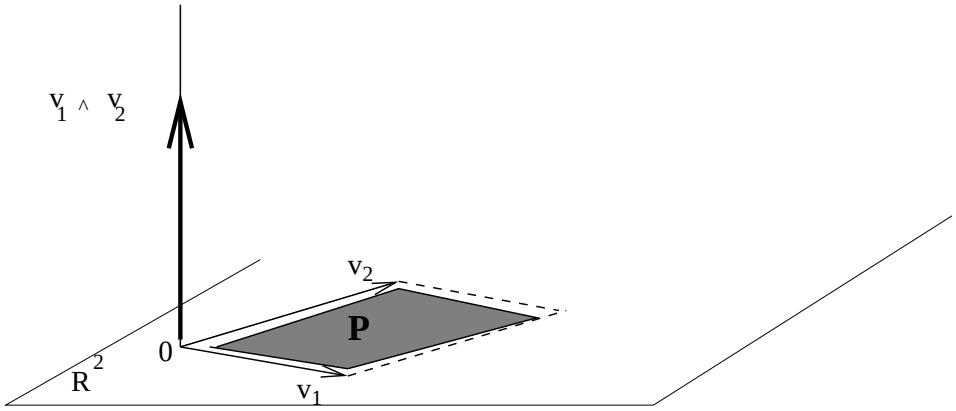


FIG. 11.6 – Le produit extérieur et l'aire d'un parallélogramme

Revenons à la formule (†) de Pythagore. Dans le cas où les vecteurs $\overrightarrow{M_2M_1}$ et $\overrightarrow{M_2M_3}$ sont non nuls et font un angle orienté $\widehat{M_1M_2M_3}$, la formule de Pythagore devient

$$d^2(M_1, M_3) = d^2(M_1, M_2) + d^2(M_2, M_3) - 2d(M_1, M_2)d(M_2, M_3)\cos(\widehat{M_1M_2M_3}) \qquad (\dagger\dagger)$$

(voir la figure ci-dessous). La distance $d(M_1, M_3)$ est supérieure ou égale à la racine carrée de $d^2(M_1, M_2)+ d^2(M_2, M_3)$ si et seulement si l'angle orienté $\widehat{M_1M_2M_3}$ est supérieur òu égal à $\pi/2$; la distance $d(M_1, M_3)$ est strictement inférieure à la racine carrée de

$$d^2(M_1, M_2) + d^2(M_2, M_3)$$

si et seulement si l'angle orienté $\widehat{M_1M_2M_3}$ est strictement inférieur à $\pi/2$ (comme c'est le cas sur la figure).

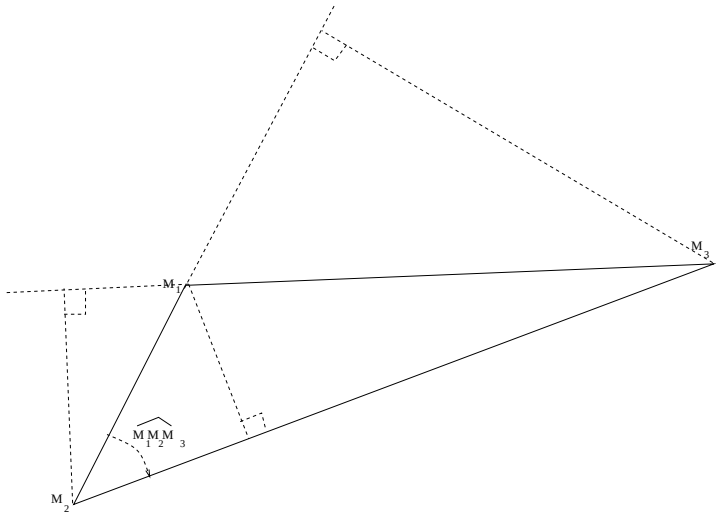


FIG. 11.7 – Figure devant servir de trame pour prouver Pythagore en exercice

Exercice 11.5.1 *Prouver Pythagore par un argument géométrique à partir de la figure ci-dessus ; on fera dans un premier temps la preuve dans le cas où l'angle de vecteurs* $\widehat{M_1M_2M_3}$ *vaut* $\pi/2$ *ou* $3\pi/2$*, puis on passera au cas général.*

Si D_1 et D_2 sont deux droites affines sécantes du plan, de directions dirigées respectivement par les vecteurs $\vec{V}_1, \vec{V}_2$, le cosinus de l'angle orienté $(\vec{V}_1, \vec{V}_2)$ ne dépend que des droites D_1 et D_2 et non des vecteurs choisis pour les diriger (ce n'est pas le cas du sinus). Ce cosinus est aussi noté $\cos(D_1, D_2)$; il est nul si et seulement si les droites D_1 et D_2 sont orthogonales.

11.5.b Projection orthogonale sur une droite, distance d'un point du plan à une droite

Soit D une droite affine du plan, d'équation cartésienne $ax + by + c = 0$; le vecteur non nul (a, b) est un vecteur orthogonal au vecteur $\vec{V} = (-b, a)$ dirigeant la droite D.

De plus, si $M = (x, y)$ est un point de $\mathbf{R}^2$, la fonction

$$f_D \ : m \in D_1 \mapsto d^2(M, m)$$

admet un minimum sur D_1 ; on peut pour voir cela choisir une représentation paramétrique de D sous la forme

$$t \mapsto M(t) := (x_0 - tb, y_0 + ta)$$

(où $(x_0, y_0) \in D$) et étudier la fonction polynomiale du second degré

$$d^2(M, M(t)) = (x - x_0 + tb)^2 + (y - y_0 - ta)^2 \, .$$

On vérifiera en exercice que cette fonction admet un minimum (qu'elle atteint) et que la valeur de ce minimum est

$$\min_{t \in \mathbf{R}} d^2(M, M(t)) = \frac{(a(x - x_0) + b(y - y_0))^2}{a^2 + b^2} \, .$$

Ainsi, il existe un unique point m de D tel que la distance de M à m soit minimale ; c'est le pied sur D de la perpendiculaire menée de M ; ce point est appelé *projection orthogonale de M sur la droite affine D* ; la distance de M à D est par définition la distance de M à ce point et vaut donc, si $M = (x, y)$,

$$d(M, D) = \frac{|ax + by + c|}{\sqrt{a^2 + b^2}}$$

si $ax + by + c = 0$ est une équation affine de D (on fera ce calcul directement en utilisant la figure ci-dessous).

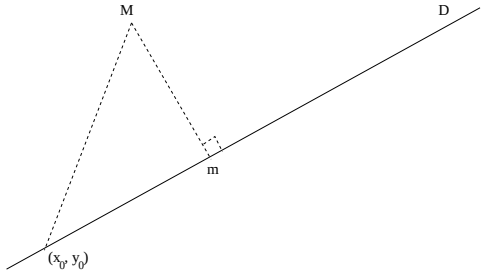


FIG. 11.8 – Projection orthogonale sur une droite affine

11.5.c *Projections itérées : deux algorithmes "pythagoriciens".*

Exercice 11.5.2 *Soient $D_1, ..., D_M$ M droites affines distinctes du plan ($M \geq 2$) toutes sécantes en un même point A (comme sur la figure ci-dessous où nous avons pris $M = 4$ pour fixer les idées). Soit M un point arbitraire du plan. Considérons l'algorithme qui consiste à projeter M orthogonalement sur D_1 (en un point m_1), puis m_1 sur D_2 (en un point m_2), etc. ; une fois toutes la liste de droites $D_1, ..., D_M$ épuisée et l'algorithme nous ayant conduit au point m_N, on le fait tourner à nouveau à partir de m_N cette fois. Que se passe-t-il intuitivement si l'on continue de la sorte (on s'aidera de la figure) ? Montrer rigoureusement qu'en fait, cet algorithme nous même vers le point A.*

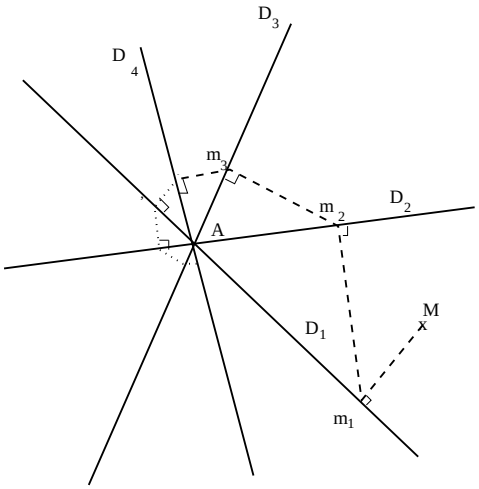


FIG. 11.9 – Les projections itérées (1)

Nous venons d'esquisser ici (dans le cas du plan euclidien) un algorithme "pythagoricien" très utile en mathématiques appliquées pour reconstituer un objet inconnu à partir d'un certain nombre d'observations. Nous en verrons une illustration plus loin avec le dispositif de CAT-Scanner.

Voici un second exemple d'un algorithme du même type que l'on développera aussi en exercice :

Exercice 11.5.3 *Soient deux droites affines sécantes distinctes D_1 et D_2, M un point inconnu de D_1, dont on suppose connue la projection orthogonale m sur D_2. Montrer que le processus algorithmique consistant à partir de m, à le projeter sur D_1 en un point m_1, puis à projeter m_1 sur la droite affine D orthogonale à D_2 en m pour obtenir un point m_2, puis à recommencer ces deux opérations à partir de m_2, et ainsi de suite, nous conduit de manière itérative vers le point inconnu M. Ici encore, on s'inspirera intuitivement de la figure ci-dessous pour construire un raisonnement mathématique propre justifiant notre assertion.*

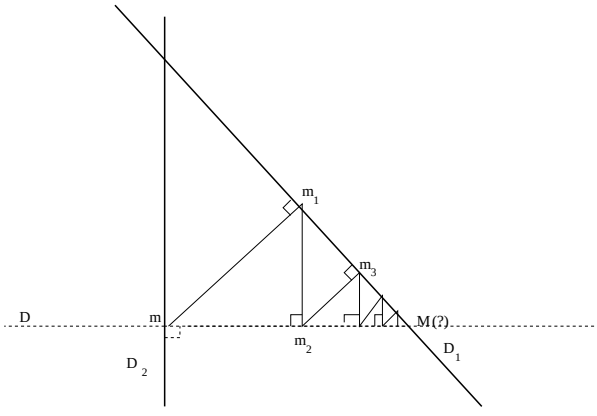


FIG. 11.10 – Les projections itérées (2)

11.5.d Nuages de points dans le plan ; droite de régression

Voici une autre application importante des idées "pythagoriciennes" dans le plan euclidien ; considérons, comme sur la figure ci-dessous, un "nuage de points", $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ (il y en a six sur notre figure) correspondant par exemple aux mesures simultanées de deux phénomènes physiques dont on veut savoir quelle est le meilleur compromis possible pour les supposer linéairement dépendants (ce qu'ils ne sont bien sûr à première vue rigoureusement pas, comme la figure nous le confirme, sinon tous les points seraient alignés).

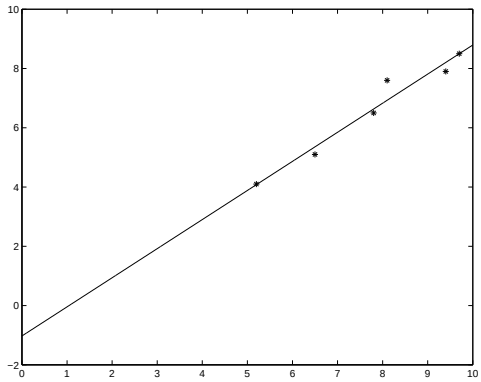


FIG. 11.11 – Tracé d'une droite de régression linéaire

Une droite est intéressante à rechercher : c'est la droite d'équation affine $y - ax - b = 0$ telle que la quantité

$$F(a, b) := (y_1 - ax_1 - b)^2 + \cdots + (y_N - ax_N - b)^2$$

soit minimale (si elle existe). Introduisons les moyennes m_x et m_y des valeurs respectives des x_j et y_j, soit

$$m_x := \frac{x_1 + \cdots + x_N}{N}$$

$$m_y := \frac{y_1 + \cdots + y_N}{N}$$

et posons $x'_j = x_j - m_x$ et $y'_j = y_j - m_y$ pour $j = 1, ..., N$. On a $\sum_{j=1}^{N} x'_j = \sum_{j=1}^{N} y'_j = 0$. Si nous trouvons a' et b' minimisant la fonction

$$G(a', b') := (y'_1 - a'x'_1 - b')^2 + \cdots + (y'_N - a'x'_N - b')^2 ,$$

on en déduira que les valeurs de a et b minimisant $F(a, b)$ sont

$$a = a' , \quad b = b' + m_y - a'm_x .$$

Un calcul simple montre que

$$\begin{aligned} G(a',b') &= a'^2\sum_{j=1}^{N} x_j'^2 - 2a\sum_{j=1}^{N} x_j'y_j' + Nb'^2 + \sum_{j=1}^{N} y_j'^2 \\ &= \Big(a'\sqrt{\sum_{j=1}^{N} x_j'^2} - \frac{\sum\limits_{j=1}^{N} x_j'y_j'}{\sqrt{\sum\limits_{j=1}^{N} x_j'^2}}\Big)^2 + Nb'^2 + \sum_{j=1}^{N} y_j'^2 - \frac{\Big(\sum\limits_{j=1}^{N} x_j'y_j'\Big)^2}{\sum\limits_{j=1}^{N} x_j'^2}\,; \end{aligned}$$

le minimum de cette fonction est donc atteint pour

$$a' = \frac{\sum\limits_{j=1}^{N} x_j'y_j'}{\sum\limits_{j=1}^{N} x_j'^2}, \quad b' = 0$$

et vaut

$$\min G(a',b') = \min F(a,b) = \sum_{j=1}^{N} y_j'^2 - \frac{\Big(\sum\limits_{j=1}^{N} x_j'y_j'\Big)^2}{\sum\limits_{j=1}^{N} x_j'^2}$$

(qui, remarquons-le, est forcément une quantité positive ou nulle, ce qui donne une inégalité très importante dans toutes les mathématiques, dite *inégalité de Cauchy-Schwarz*).

La droite affine réalisant le meilleur compromis concernant la dépendance linéaire de l'information y à partir de l'information x au sein du nuage de points est donc la droite affine d'équation

$$y - m_y = a'(x - m_x)\,;$$

cette droite importante est dite *droite de régression linéaire* et le nombre

$$\rho := \frac{\sum\limits_{j=1}^{n} x_j'y_j'}{\sqrt{\sum\limits_{j=1}^{N} x_j'^2}\sqrt{\sum\limits_{j=1}^{N} y_j'^2}}$$

est dit *coefficient de corrélation entre les x_j et les y_j* au sein du nuage. Ainsi la droite de régression linéaire a-t-elle pour équation cartésienne

$$\frac{y - m_y}{\sqrt{\sum\limits_{j=1}^{N} y_j'^2}} = \rho\frac{x - m_x}{\sqrt{\sum\limits_{j=1}^{N} x_j'^2}}\,.$$

Ces notions jouent un rôle très important en théorie des probabilités et plus particulièrement dans l'étude des modèles statistiques (les enquêtes d'opinion par exemple).

11.6 Les droites du plan terrain de modélisation numérique

11.6.a Le repérage $x\cos\theta + y\sin\theta = p$

Il existe d'autres manières d'organiser les droites affines du plan ; nous en mentionnerons une ici, intéressante car en relation avec les techniques basées sur le principe du CAT-Scanner (CAT pour *Computer Aid Tomography*) permettant l'inversion de la transformation dite *aux rayons X*. Cette transformation faisant intervenir la famille des droites du plan convenablement indexée illustrera la mise en équation d'un problème mathématique au service de préoccupations pratiques dont l'importance s'est faite capitale aujourd'hui. Nous y reviendrons plus loin.

Comme sur la figure ci-dessous, on peut repérer une droite du plan en se donnant un angle $\theta \in [0, 2\pi[$ (ou, ce qui revient au même, un point $e^{i\theta}$ du cercle unité du plan complexe) et un nombre p, la droite affine $D_{\theta,p}$ étant la droite affine ayant la représentation paramétrique suivante :

$$t \in \mathbf{R} \mapsto (p\cos\theta - t\sin\theta, p\sin\theta + t\cos\theta)$$

ou l'équation cartésienne

$$x\cos\theta + y\sin\theta - p = 0\,.$$

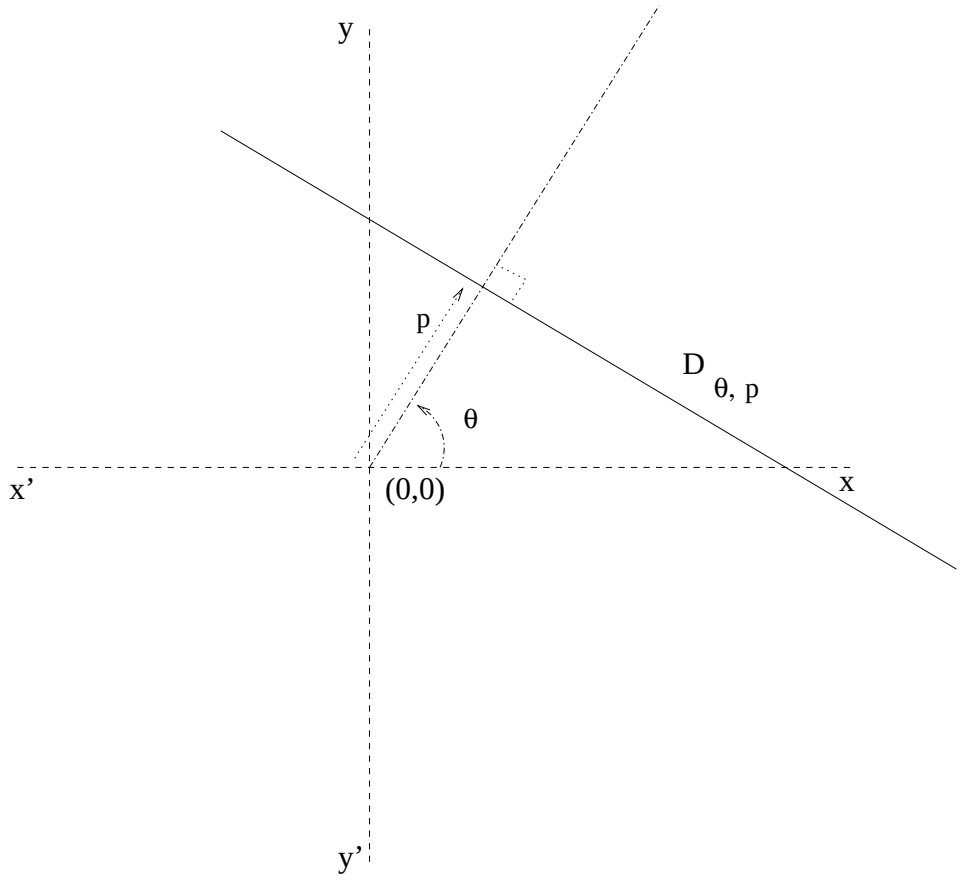


FIG. 11.12 – Le repérage d'une droite par θ et p

Comme on le remarque facilement, les couples (θ, p) et $(-\theta, -p)$ repèrent de fait la même droite affine ; il y a donc correspondance entre l'ensemble des droites affines du plan d'une part et l'ensemble des couples (z, p) où z est un nombre complexe de module 1, les couples $(-z, -p)$ et (z, p) étant identifiés. Ceci nous fournit une représentation de l'ensemble des droites du plan différente de celle introduite dans la Sect. 11.4.a.

11.6.b Droites du plan et rayonnement gamma : le principe du scanner

Les deux sections qui suivent sont là comme thème de réflexion pour montrer comment les idées mathématiques élémentaires que nous avons introduit en étudiant la famille des droites du plan peuvent très rapidement devenir des outils fondamentaux au service de problèmes pratiques très concrets.

Nous avons choisi d'illustrer comment intervient la famille des droites du plan dans un problème devenu depuis les années 1970 (avec l'attribution du prix Nobel de Médecine en 1979 à Cormack et Hounsfield) tout à fait actuel : celui de la gammagraphie et du CAT-Scanner.

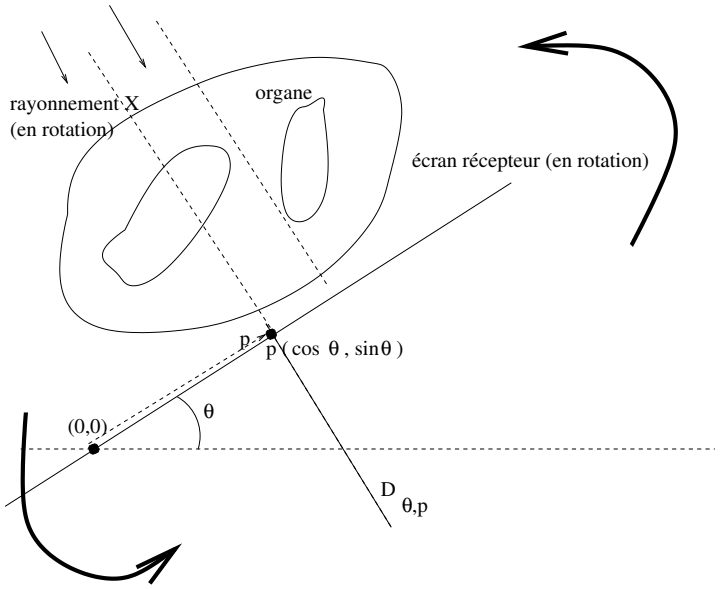


FIG. 11.13 – Le principe du CAT-Scanner en 2D

Supposons qu'un organe (supposé ici plan et contenu dans le disque de centre 0 et de rayon R) sur lequel vit une certaine fonction $(x, y) \to f(x, y)$ (mesurant par exemple la densité du tissu organique) soit soumis à un rayonnement gamma émis par une caméra orientée vers l'organe tout en pivotant autour de lui ; couplé avec la caméra, se trouve un écran récepteur qui enregistre la trace du rayonnement une fois sortie de l'organe (comme sur la figure ci-dessus) ; la densité du tissu ayant tendance à atténuer le rayonnement, on dispose d'une liste de clichés radio-graphiques (après balayage de la caméra autour de l'organe) nous permettant de connaître (en fait en négatif, comme sur toute radiographie) pour chaque valeur de θ entre 0 et 2π la fonction du paramètre réel p définie par

$$F(p, \theta) := \int_{-\infty}^{\infty} f(p\cos\theta - t\sin\theta, p\sin\theta + t\cos\theta)\, dt\,.$$

Au lieu de la fonction inconnue f, de deux variables, on ne dispose donc en fait que de la fonction $F(p,\theta)$ définie sur $[0,2\pi]\times[0,R]$. Cette fonction F n'a *a priori* rien à voir avec f ; c'est le *sinogramme* de f.

Construire le sinogramme d'une image circulaire limitée par le cercle de centre $(0,0)$ et de rayon R (l'intensité lumineuse étant f et jouant le rôle de la densité du tissu organique) n'est pas chose tout à fait immédiate ; c'est déjà un problème de modélisation consistant en la conception d'une "caméra digitale" qui, pour chaque valeur de l'angle θ entre 0 et 2π et pour chaque valeur de p entre 0 et R, stocke les intensités lumineuses $f(p\cos\theta - t\sin\theta, p\sin\theta + t\cos\theta)$ pour toutes les valeurs de t telles que le point

$$(p\cos\theta - t\sin\theta, p\sin\theta + t\cos\theta)$$

soit dans le cadre de l'image, c'est-à-dire pour les valeurs de t entre $-\sqrt{R^2-p^2}$ et $+\sqrt{R^2-p^2}$. Sur la figure ci-dessous, nous avons affiché dans le cadre supérieur gauche l'image originelle telle qu'elle apparaît dans les codes de couleur de l'ordinateur.

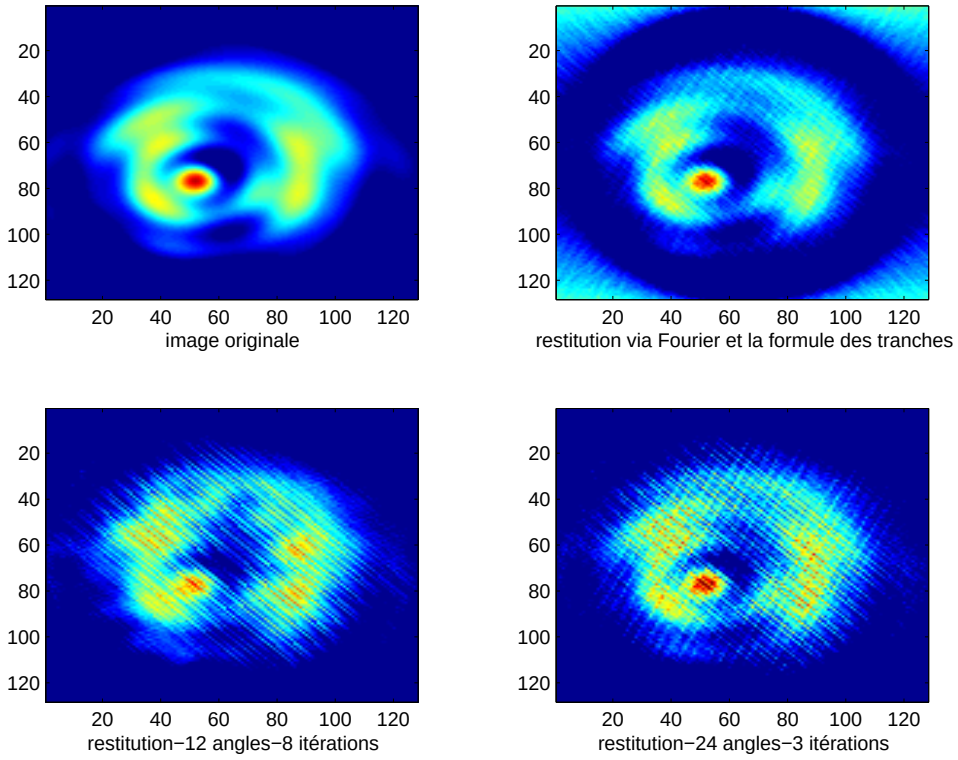


FIG. 11.14 – Une image originelle et diverses tentatives de restitution

Nous reviendrons à cette figure ultérieurement.

Sur la figure ci-dessous, nous avons affiché l'image digitale correspondant au sinogramme de l'image originelle, le paramètre p variant entre 0 et R figurant en abscisse et le paramètre θ variant lui entre 0 et 2π figurant en ordonnée.

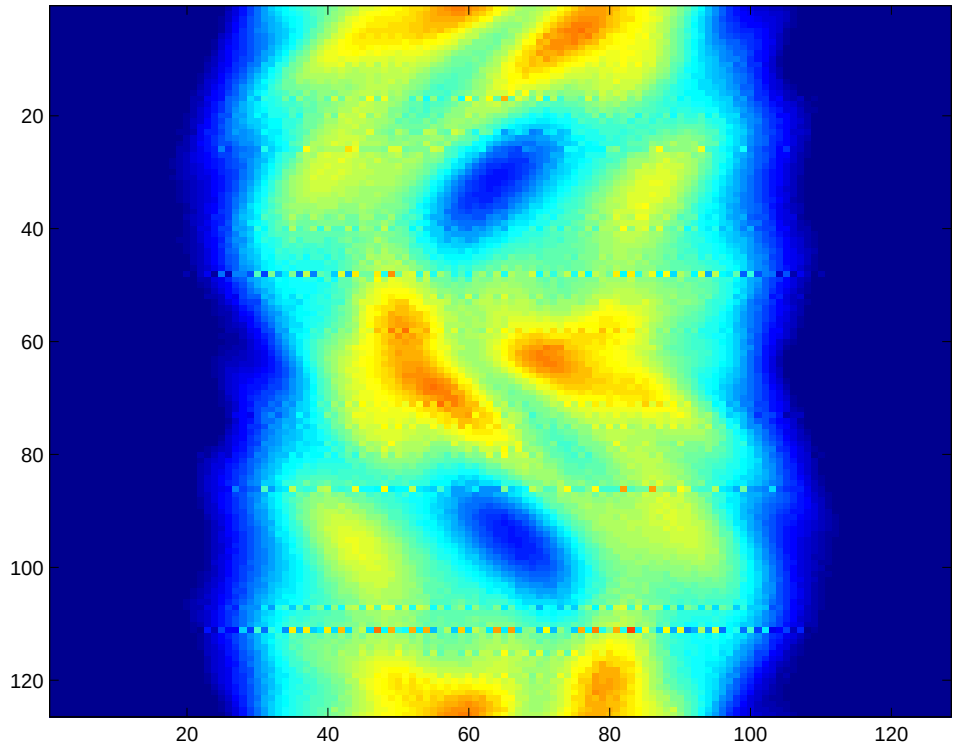


FIG. 11.15 – Un exemple de sinogramme

On se convaincra aisément que deviner l'image originelle que ce sinogramme dissimule n'est pas une opération immédiate.

11.6.c Retrouver une image à partir de son sinogramme (thème d'exercice à illustrer numériquement)

Il n'est pas question ici d'expliquer rigoureusement comment retrouver l'image originelle à partir de son sinogramme. Cependant, nous allons juste suggérer comment les idées *pythagoriciennes* que nous avons présenté dans la Sect. 11.5 dans le cadre du plan, transposées dans un tout autre cadre (où elles fonctionnent de la même manière), fournissent une méthode d'approche pour attaquer ce problème.

L'ensemble des images digitales f définies dans le disque de centre $(0,0)$ (matérialisé par une liste de pixels $A_1, ..., A_n$ régulièrement distribués dans ce disque) est un espace vectoriel s'identifiant à $\mathbf{R}^n$ (le nombre de degrés de liberté n est le nombre de pixels). On peut définir le produit scalaire de deux images digitales f et g par

$$\langle f, g \rangle := \sum_{j=1}^{n} f(A_j) g(A_j) \, .$$

Dans cet espace d'images (qui est cette fois de dimension n, le nombre de pixels et non plus 2 ou 3 comme dans le cas du plan ou de l'espace), le sous-ensemble des images pour lesquelles le dispositif caméra/écran récepteur ne voit rien lorsque le rayonnement se fait dans une direction θ donnée est un sous-espace vectoriel E_θ car il reste globalement stable par prise de combinaisons linéaires.

La connaissance du sinogramme de f permet, pour une image g arbitraire fixée et pour tout angle θ, de déterminer la projection orthogonale de notre image inconnue f sur le sous-espace affine $f + E_\theta$. Dès lors, le scénario de projections orthogonales itérées que nous avons décrit dans le cadre très simple du plan dans l'Ex. 11.5.2 peut se mettre en place : on choisit une liste d'angles $\theta_1, ..., \theta_N$ bien répartis dans l'intervalle $[0, 2\pi]$, puis, partant de l'image nulle $g \equiv 0$, on la projette sur $f + E_{\theta_1}$, puis on projette l'image obtenue sur $f + E_{\theta_2}$, et ainsi de suite, avec réutilisation de la liste des angles lorsque celle ci est épuisée. C'est ce que nous avons fait numériquement pour obtenir les deux images figurant dans le cadre inférieur de la première figure présentée dans la sous-section précédente.

11.7 Plans et droites de l'espace affine $\mathbf{R}^3$

Pour terminer ce chapitre nous exprimons en coordonnées la position relative de droites et plans de $\mathbf{R}^3$.

11.7.a Intersection de deux plans

Soient Π_1 et Π_2 deux plans affines de l'espace $\mathbf{R}^3$, d'équations cartésiennes respectives

$$\begin{aligned} (\Pi_1) \qquad & a_1x + b_1y + c_1z + d_1 = 0 \\ (\Pi_2) \qquad & a_2x + b_2y + c_2z + d_2 = 0\,. \end{aligned}$$

avec (a_1, b_1, c_1) et (a_2, b_2, c_2) dans $\mathbf{R}^3 \backslash \{(0, 0, 0)\}$. Plusieurs cas sont à envisager concernant l'intersection éventuelle de Π_1 et Π_2.

- Si les deux vecteurs (a_1, b_1, c_1, d_1) et (a_2, b_2, c_2, d_2) sont colinéaires dans $\mathbf{R}^4$, les deux plans affines sont confondus.
- Si les deux vecteurs (a_1, b_1, c_1) et (a_2, b_2, c_2) sont colinéaires dans $\mathbf{R}^3$, mais les vecteurs (a_1, b_1, c_1, d_1) et (a_2, b_2, c_2, d_2) ne le sont pas dans $\mathbf{R}^4$, alors les plans Π_1 et Π_2 ne s'intersectent pas dans l'espace affine $\mathbf{R}^3$; on dit qu'ils sont *parallèles*.
- Si les deux vecteurs (a_1, b_1, c_1) et (a_2, b_2, c_2) ne sont pas colinéaires dans $\mathbf{R}^3$, l'un des trois nombres $b_1c_2 - b_2c_1, c_1a_2 - a_1c_2, a_1b_2 - a_2b_1$ est non nul. Supposons pour fixer les idées que $a_1b_2 - a_2b_1 \neq 0$. Dans ce cas, la méthode du pivot nous permet de résoudre, pour toute valeur de z arbitraire, le système linéaire
$$\left\{ \begin{array}{rcl} a_1x + b_1y & = & -c_1z - d_1 \\ a_2x + b_2y & = & -c_2z - d_2\,; \end{array} \right.$$
on trouve
$$\left\{ \begin{array}{rcl} x & = & \frac{(b_1c_2 - b_2c_1)z + b_1d_2 - b_2d_1}{a_1b_2 - a_2b_1} \\ y & = & \frac{(a_2c_1 - a_1c_2)z + a_2d_1 - a_1d_2}{a_1b_2 - a_2b_1}\,; \end{array} \right.$$
on obtient ainsi la représentation paramétrique d'une droite affine, le paramètre étant z.

Ainsi l'intersection de deux plans affines non parallèles et non confondus est toujours une droite affine.

Inversement, pour se donner la représentation cartésienne d'une droite affine, on doit donner les équations d'exactement deux plans non confondus qui la contiennent.

11.7.b Intersection d'un plan et d'une droite

Considérons un plan affine Π d'équation cartésienne $ax + by + cz + d = 0$ et une droite paramétrée

$$\left\{ \begin{array}{rcl} x(t) & = & x_0 + tu \\ y(t) & = & y_0 + tv \\ z(t) & = & z_0 + tw \end{array} \right.$$

Pour chercher les éventuels points d'intersection du plan affine Π et de la droite affine ainsi paramétrée, on doit chercher à résoudre l'équation en t suivante :

$$t(au + bv + cw) + (ax_0 + by_0 + cz_0) + d = 0\,.$$

Ici encore, plusieurs cas sont à envisager :

- Les nombres $au + bv + cw$ et $ax_0 + by_0 + cz_0 + d$ sont nuls ; ceci signifie que la droite est incluse dans le plan Π.
- On a $au+bv+cw = 0$ mais $ax_0+by_0+cz_0+d \neq 0$; dans ce cas, il n'y a pas de point d'intersection, la droite est dite *parallèle* au plan Π.
- On a $au + bv + cw \neq 0$; dans ce cas, il y a un unique point d'intersection correspondant à la valeur du paramètre
$$t = -\frac{ax_0 + by_0 + cz_0 + d}{au + bv + cw}\,.$$

Ainsi l'intersection d'un plan affine et d'une droite non parallèle au plan ou non incluse dedans est toujours un point.

11.7.c Distance euclidienne, angles, aires et volumes dans l'espace $\mathbf{R}^3$

L'espace vectoriel $\mathbf{R}^3$ hérite aussi d'un produit scalaire défini cette fois par

$$\langle (x_1, y_1, z_1)\,,\, (x_2, y_2, z_2)\rangle := x_1x_2 + y_1y_2 + z_1z_2\,.$$

On peut définir l'angle (non plus cette fois orienté) de deux vecteurs $\vec{V_1}$ et $\vec{V_2}$ non nuls de coordonnées respectives (x_1, y_1, z_1) et (x_2, y_2, z_2). Pour cela, on introduit le vecteur $\vec{V_1} \wedge \vec{V_2}$ (dit *produit extérieur* de $\vec{V_1}$ et $\vec{V_2}$) défini par linéarité d'une part, avec les règles de calcul

$$\begin{aligned}
\vec{i} \wedge \vec{i} &= \vec{j} \wedge \vec{j} = \vec{k} \wedge \vec{k} = \vec{0}\\
\vec{i} \wedge \vec{j} &= -\vec{j} \wedge \vec{i} = \vec{k}\\
\vec{j} \wedge \vec{k} &= -\vec{k} \wedge \vec{j} = \vec{i}\\
\vec{k} \wedge \vec{i} &= -\vec{i} \wedge \vec{k} = \vec{j}\,,
\end{aligned}$$

d'autre part. Ainsi

$$\vec{V_1} \wedge \vec{V_2} = (y_1z_2 - y_2z_1, x_2z_1 - x_1z_2, x_1y_2 - x_2y_1)\,.$$

On a toujours, comme dans le cas du plan, la formule

$$\|\vec{V_1} \wedge \vec{V_2}\|^2 + (\langle \vec{V_1}\,,\, \vec{V_2}\rangle)^2 = \|\vec{V_1}\|^2\, \|\vec{V_2}\|^2$$

si

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

Ceci implique l'existence d'un unique réel θ dans $[0, \pi[$ (notons la différence avec le cas du plan où l'angle se trouvait être un angle orienté entre 0 et 2π, le signe de $x_1y_2 - x_2y_1$ étant pris en compte, ce qu'il n'est plus possible de faire ici) tel que

$$\begin{aligned}
\cos\theta &:= \frac{\langle \vec{V_1}\,,\, \vec{V_2}\rangle}{\|\vec{V_1}\|\, \|\vec{V_2}\|}\\
\sin\theta &:= \frac{\|\vec{V_1} \wedge \vec{V_2}\|}{\|\vec{V_1}\|\, \|\vec{V_2}\|}\,.
\end{aligned}$$

Comme dans le cas où les $\vec{V}_j$ sont des vecteurs du plan, la quantité $\|\vec{V}_1 \wedge \vec{V}_2\|$ représente la surface du parallélogramme plan construit dans l'espace à partir des vecteurs $\vec{V}_1$ et $\vec{V}_2$. Si $\vec{V}_3$ est un troisième vecteur tel que $(\vec{V}_1, \vec{V}_2, \vec{V}_3)$ forment une base de $\mathbf{R}^3$ (c'est-à-dire un système générateur maximal), le nombre réel positif

$$\left|\left\langle \vec{V}_1 \wedge \vec{V}_2 \,,\, \vec{V}_3 \right\rangle\right|$$

représente le volume (dans $\mathbf{R}^3$) du parallélépipède Q construit à partir de $\vec{V}_1, \vec{V}_2, \vec{V}_3$ comme sur la figure ci-dessous. Le nombre réel $\left\langle \vec{V}_1 \wedge \vec{V}_2 \,,\, \vec{V}_3 \right\rangle$ s'appelle le *produit mixte* des trois vecteurs $\vec{V}_1, \vec{V}_2, \vec{V}_3$ (l'ordre est important ici).

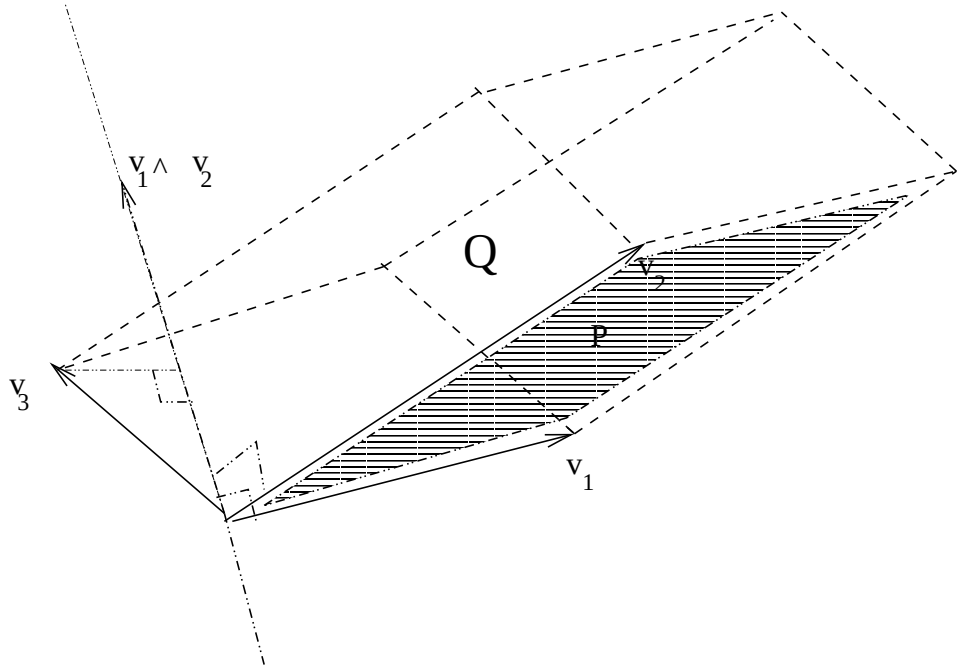


FIG. 11.16 – Volume d'un parallélépipède

(on s'inspirera de la figure pour vérifier ce résultat en exercice).

La formule de Pythagore (†) de la section 11.5.a pour trois points quelconques M_1, M_2, M_3 de l'espace et sa relecture (††) (lorsqu'en plus les trois points sont tels que les vecteurs $\overrightarrow{M_2M_1}$ et $\overrightarrow{M_2M_3}$ soient non nuls et définissent un angle de cosinus $\cos(M_1\widehat{M_2}M_3)$) restent encore valable dans ce nouveau cadre.

11.7.d Distance d'un point à un plan affine

Si Π un plan affine de l'espace $\mathbf{R}^3$ d'équation cartésienne

$$ax + by + cz + d = 0\,,$$

le vecteur (a, b, c) est orthogonal à tous les vecteurs $\overrightarrow{M_1M_2}$, où M_1 et M_2 sont deux points quelconques de Π.

Soit M un point quelconque de l'espace, D la droite passant par M et dirigée par ce vecteur (a, b, c) ; cette droite (qui ne peut être ni parallèle ni contenue dans Π) coupe le plan Π en un unique point m dit *projection orthogonale de M sur le plan affine* Π (voir la figure ci-dessous).

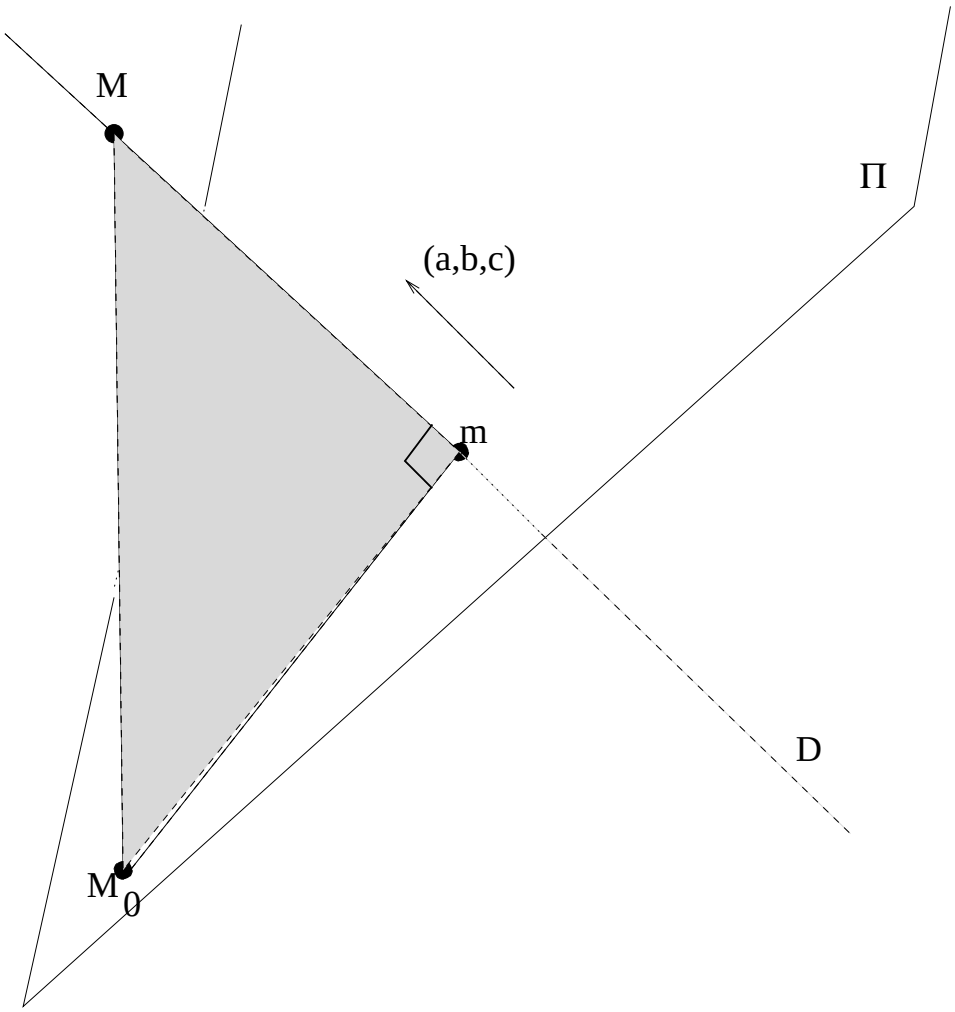


FIG. 11.17 – Projection orthogonale sur un plan affine

On voit en se plaçant dans tous les plans MmM_0 lorsque $M_0 \in \Pi$ (comme le plan hachuré de la figure) que le point m réalise le minimum de la distance de M à tous les points du plan Π. La valeur de ce minimum est la *distance du point M au plan* Π et vaut

$$d(M, \Pi) = \frac{|ax + by + cz + d|}{\sqrt{a_2 + b_2 + c^2}}\,.$$

11.7.e Projection sur une droite affine

Soit D une droite affine, définie sous forme paramétrique

$$\begin{aligned} x(t) &= x_0 + tu \\ y(t) &= y_0 + tv \\ z(t) &= z_0 + tw\,. \end{aligned}$$

Si $M = (x, y, z)$ est un point de l'espace, le carré de la distance du point M au point courant $M(t) = (x(t), y(t), z(t))$ de D vaut

$$\begin{aligned} d^2(M, M(t)) &= (x - x_0 - tu)^2 + (y - y_0 - tv)^2 + (z - z_0 - tw)^2 \\ &= t^2(u^2 + v^2 + w^2) - 2t((x - x_0)u + (y - y_0)v + (z - z_0)w) \\ &\quad +(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2. \end{aligned}$$

Ce trinôme en t prend sa valeur minimale pour

$$t = \frac{(x - x_0)u + (y - y_0)v + (z - z_0)w}{u^2 + v^2 + w^2}$$

et le point $m(t)$ où cette distance est atteinte est l'unique point m de la droite D tel que le vecteur $\overrightarrow{mM}$ soit orthogonal au vecteur directeur (u, v, w) de D. Ce point m est dit *projection orthogonale de M sur la droite affine D.*

Si la droite D est donnée sous forme cartésienne par le jeu d'équations

$$\begin{aligned} a_1x + b_1y + c_1z + d_1 &= 0 \\ a_2x + b_2y + c_2z + d_2 &= 0\,, \end{aligned}$$

où les deux vecteurs non nuls $\vec{V}_1 = (a_1, b_1, c_1)$ et $\vec{V}_2 = (a_2, b_2, c_2)$ ne sont pas colinéaires, la distance de M à D est aussi donnée si M_0 est un point arbitraire de D (on fera ici encore l'exercice) par la formule :

$$d(M, D) = \frac{\|\overrightarrow{M_0M} \wedge (\vec{V}_1 \wedge \vec{V}_2)\|}{\|\vec{V}_1 \wedge \vec{V}_2\|}\,.$$

Exercice 11.7.1 *Démontrer, tout d'abord par le calcul, puis ensuite géométriquement, la formule ci-dessus.*

11.7.f Distance entre deux droites affines de l'espace

Soient D_1 et D_2 deux droites affines de l'espace paramétrées respectivement par

$$M_1 + t\,\vec{V}_1 \quad \text{et} \quad M_2 + t\,\vec{V}_2$$

Si les vecteurs directeurs $\vec{V}_1$ et $\vec{V}_2$ sont colinéaires, les droites sont soit confondues (si $M_1 = M_2$), auquel cas l'intersection des deux droites est D_1, soit sont deux droites parallèles du plan défini par les trois points $M_1, M_2, M_1 + \vec{V}_1$, auquel cas l'intersection des deux droites est vide. Le cas intéressant restant est celui où les vecteurs $\vec{V}_1$ et $\vec{V}_2$ sont linéairement indépendants.

On suppose donc à partir de maintenant $\vec{V}_1$ et $\vec{V}_2$ linéairement indépendants. Soit Π le plan affine paramétré par

$$m(t, s) = M_1 + t\vec{V}_1 + s\vec{V}_2\,,\ t, s \in \mathbf{R}\,;$$

ce plan Π contient la droite D_1 ; concernant la droite D_2, la discussion faite dans la section 11.7.b montre que soit elle est incluse dans le plan Π, soit elle est parallèle à ce plan. Examinons la figure proposée ci-dessous :

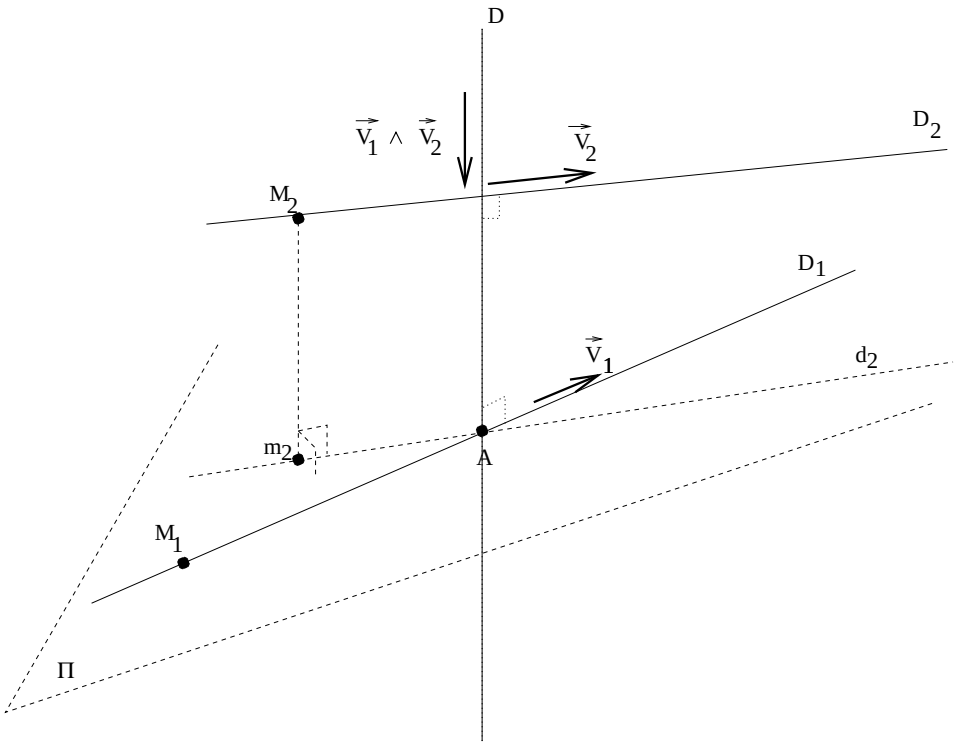


FIG. 11.18 – Distance de deux droites gauches et perpendiculaire commune

La droite D_2 est une droite parallèle au plan Π (ou incluse dedans) et la distance de D_2 à Π est aussi la distance du point M_2 au plan Π (qui est nulle si D_2 est aussi incluse dans Π). Notons m_2 la projection orthogonale de M_2 sur le plan Π. Si M_1 a pour coordonnées (x_1, y_1, z_1), le plan Π a pour équation cartésienne

$$\left\langle (x - x_1, y - y_1, z - z_1)\,,\, \vec{V}_1 \wedge \vec{V}_2 \right\rangle = 0\,.$$

La distance du point M_2 au plan Π, donc la distance entre les deux droites D_1 et D_2 vaut donc (d'après le résultat établi dans la section 11.7.d)

$$d(D_1, D_2) = \frac{|\langle \overrightarrow{M_1M_2}\,,\, \vec{V}_1 \wedge \vec{V}_2 \rangle|}{\|\vec{V}_1 \wedge \vec{V}_2\|}\,;$$

notons que le numérateur de cette expression est aussi le volume du parallélépipède de l'espace construit sur les trois vecteurs linéairement indépendants $\vec{V}_1, \vec{V}_2, \overrightarrow{M_1M_2}$ (voir la section 11.7.c).

Si d_2 est la parallèle à D_2 passant par le point m_2, les droites D_1 et d_2 sont sécantes en un point A dans le plan Π. La droite D, perpendiculaire à Π au point A rencontre les deux droites D_1 et D_2 et leur est perpendiculaire ; on appelle l'unique droite ayant ces propriétés la *perpendiculaire commune aux deux droites D_1 et D_2*.

Chapitre 12

Modélisation et équations différentielles

12.1 Introduction

Nous évoluons dans un espace-temps où l'espace est plus ou moins réduit à notre planète et le temps se déroule en dehors de notre volonté et conduit à un processus d'évolution. La terre se façonne au cours du temps et flore, faune et techniques naissent, perdurent ou disparaissent en fonction d'un certain nombre de paramètres. Cette évolution brève ou longue peut être mise en équations et résolue par un procédé analytique ou approchée par un procédé numérique. C'est là qu'interviennent mathématique et mathématiciens. La mathématique permet d'écrire les "bonnes" équations ou les "bons" systèmes, c'est le domaine de la modélisation qui consiste à représenter un phénomène par des équations et de lui associer des conditions initiales au départ de l'évolution considérée pour pouvoir la décrire avec pertinence. Le rôle des mathématiciens est de conduire cette modélisation et de développer les outils continus ou discrets pour prédire de façon exacte ou approchée l'évolution du phénomène. Cela signifie que le mathématicien essaie de prendre en considération le maximum de paramètres pour écrire un modèle réaliste, puis étudie le modèle d'un point de vue analytique et si nécessaire (dans le cas où l'on ne sait pas calculer de solution explicite) développe des outils d'approximation pour calculer une solution approchée sur un ensemble discret de points.

Les applications sont nombreuses et sans forcément nous en rendre compte nous utilisons tous les jours les résultats de cette démarche. Les prédictions météorologiques sont faites à partir de la résolution d'un ou plusieurs modèles qui représentent l'évolution du climat (on dit aussi temps, bizarrerie de la langue française !) sur une partie plus ou moins grande de la surface de la terre. Si on laisse évoluer le modèle (qui en l'occurrence est imparfait et de plus ne peut être résolu exactement) en temps on obtient une prédiction sans aucun lien avec la réalité. Il faut donc recaler le modèle, c'est à dire lui redonner une condition initiale pertinente en fonction des observations de toutes natures effectuées à des fréquences régulières sur l'ensemble du globe (mesures, sondes, satellites, etc.). D'après Lorenz le mouvement d'une mouette entraîne après un temps fini un écart qui est de l'ordre de grandeur de l'espace exploré. On comprend mieux la difficulté de faire des prévisions sur le long terme et on commence à toucher la notion de chaos. Il a été découvert récemment qu'il est impossible de décrire exactement le mouvement des planètes du système solaire sur de très longs intervalles de temps car notre terre par exemple a un mouvement chaotique sur des millions d'années (ce qui est très peu au regard de l'âge du système solaire).

12.2 La démarche de modélisation

Bien des domaines de notre vie de tous les jours peuvent être "mis en équations" et donc prédits dans une certaine mesure ; phénomènes physiques, réactions chimiques bien sûr, mais aussi environnement, finances ou biologie. Pour fixer les idées prenons l'*évolution des espèces*. Si une espèce vivante évolue dans un environnement favorable, elle peut avoir une croissance exponentielle de la forme e^{at} où a est le taux de croissance de l'espèce en dehors de toute influence extérieure et t la variable de temps. Ceci peut se traduire sous la forme d'une équation différentielle linéaire très simple :

$$x'(t) = a\, x(t)$$

où $x(t)$ représente l'espèce considérée. Dans cette équation le coefficient a est connu et positif ; donc la population va croître de façon exponentielle. Mais pour avoir des données quantitatives exactes, il faut rajouter une condition initiale au départ de l'intervalle de temps [0,T] sur lequel on observe l'évolution de l'espèce. C'est-à-dire $x(0) = x_0$ le nombre d'individus à l'instant initial. Ici nous avons fait l'hypothèse que les caractéristiques du milieu sont constantes au cours du temps. Sans quoi il faudrait considérer un coefficient $a(t)$ et alors la croissance ferait intervenir la primitive de cette fonction du temps qui s'annule en 0. Sous les mêmes hypothèses, une autre espèce y dans des conditions défavorables va décroître comme $y_0 e^{-bt}$. C'est le cas si une population de carnivores ne peut se nourrir en suffisance.

Un cas plus intéressant est de mettre ces deux populations, lièvres et lynx par exemple ensemble en supposant que les espèces ne sont sensibles qu'à leur action réciproque. Lorsque les deux espèces cohabitent, on suppose que le coefficient d'accroissement de la population x diminue proportionnellement à y lorsque celle-ci augmente. En exprimant par c la pression de prédation et par d l'accessibilité des proies, on obtient un système de deux équations différentielles :

$$\left\{ \begin{array}{rcl} x'(t) & = & (a - c\, y(t))\, x(t) \\ y'(t) & = & (-b + d\, x(t))\, y(t)\ . \end{array} \right.$$

Ce système non linéaire a été introduit il y a prés d'un siècle par Lotka et Volterra et est connu sous le nom de *système proie-prédateur*. On observe avec ce modèle des variations périodiques souvent sinusoïdales des populations de lièvres et de lynx au cours du temps. Quand les prédateurs ont suffisamment de nourriture leur population croît et ils consomment de plus en plus de proies qui diminuent en nombre et deviennent plus rares, entraînant le déclin des prédateurs. Les proies peuvent alors se multiplier en étant peu chassées et ainsi de suite. Ici encore on pourrait faire dépendre les coefficients du temps pour représenter les saisons ou encore prendre en compte d'autres facteurs comme la contamination par des maladies.

Cet exemple nous permet d'entrevoir la mise en équations du monde réel. Ce qui est fascinant c'est qu'avec le développement des outils mathématiques et des ordinateurs, on a l'impression de pouvoir prédire à terme l'évolution du monde. Quant à le maîtriser ... c'est une autre histoire !

12.3 Le problème de Cauchy

Définition 12.3.1 *On appelle problème de Cauchy le problème constitué d'une ou plusieurs équations différentielles à laquelle ou auxquelles est ou sont associée(s) une ou plusieurs conditions initiales.*

Le principal résultat est le suivant :

Théorème 12.3.2 *Soient I un intervalle ouvert de $\mathbf{R}$, a une fonction numérique continue définie sur I, $t_0 \in I$ et $x_0 \in \mathbf{R}$; le problème de Cauchy associé à l'équation linéaire*

$$\begin{aligned} x'(t) &= a(t)x(t) \ \ \forall t \in I \\ x(t_0) &= x_0 \end{aligned}$$

admet une solution unique

$$x(t) = x_0 e^{\int_{t_0}^{t} a(s)ds}$$

Remarque 12.3.3 *Dans ce théorème, l'équation différentielle est linéaire et homogène. Le résultat découle directement du théorème fondamental du calcul différentiel qui stipule que si a est une fonction continue, la fonction $A(t) = \int_{t_0}^{t} a(s)ds$ est la primitive de a qui s'annule en t_0.*

Ici il suffit de faire le changement de variable $y(t) = x(t)e^{-A(t)}$ pour avoir $x(t) = y(t)e^{A(t)}$ et

$$x'(t) = y'(t)e^{A(t)} + y(t)a(t)e^{A(t)} \ .$$

Et en reportant ces expressions dans l'équation on obtient

$$y'(t)e^{A(t)} + y(t)a(t)e^{A(t)} = a(t)y(t)e^{A(t)} \ .$$

Ce qui montre que $y'(t) = 0 \ \ \forall t \in I$ et donc que y est une fonction constante égale à x_0.

Remarque 12.3.4 *Si a est un nombre réel on obtient que la solution du problème de Cauchy est $x(t) = x_0 e^{a(t-t_0)}$ et l'on retrouve bien l'expression de la croissance exponentielle évoquée plus haut.*

Théorème 12.3.5 *Soient I un intervalle ouvert de $\mathbf{R}$, a et b deux fonctions numériques continues définies sur I, $t_0 \in I$ et $x_0 \in \mathbf{R}$; le problème de Cauchy*

$$\begin{aligned} x'(t) &= a(t)x(t) + b(t) \ \ \forall t \in I \\ x(t_0) &= x_0 \end{aligned}$$

admet une solution unique obtenue à partir d'une solution particulière $\tilde{x}$ de l'équation homogène ($b = 0$) sous la forme

$$x(t) = C(t)\tilde{x}(t) \ \ \text{où } C \text{ satisfait } \ C'(t)\tilde{x}(t) = b(t)$$

Prenons par exemple

$$\tilde{x}(t) = e^{\tilde{A(t)}} \ \ \text{où} \ \ \tilde{A}(t) = \int_{t_1}^{t} a(s)ds \ \ \text{est une primitive quelconque de} \ \ a.$$

Alors $x(t) = C(t)\tilde{x}(t)$ est solution de l'équation car

$$x'(t) = C'(t)\tilde{x}(t) + C(t)\tilde{x}'(t) \ \ \text{entraîne } C'(t)\tilde{x}(t) + C(t)\tilde{x}'(t) = a(t)C(t)\tilde{x}(t) + b(t) \ .$$

Ce qui est équivalent à $C'(t)\tilde{x}(t) = b(t)$ car $\tilde{x}$ est solution de l'équation homogène. Ce qui montre que la condition du théorème sur C équivaut à x solution. Il reste à déterminer C. Soit $\tilde{B}$ une primitive de $b/\tilde{x}$ il vient en notant cte une constante quelconque

$$\begin{aligned} x(t) &= (\tilde{B}(t) + cte) \ \tilde{x}(t) \\ &= (\int_{t_2}^{t} b(s)e^{-\tilde{A}(s)}ds + cte) \ e^{\tilde{A(t)}} \\ &= (\int_{t_2}^{t} b(s)e^{-\int_{t_1}^{s} a(u)du}ds + cte) \ e^{\int_{t_1}^{t} a(s)ds} \ . \end{aligned}$$

Ce qui donne pour $t_1 = t_2 = t_0$ l'expression de la solution du problème de Cauchy

$$x(t) = (\int_{t_0}^{t} b(s)e^{-\int_{t_0}^{s} a(u)du} ds + x_0)\, e^{\int_{t_0}^{t} a(s)ds} .$$

Remarque 12.3.6 *Le procédé ci-dessus qui ne concerne que les équations linéaires sert à résoudre l'*équation non linéaire de Bernoulli

$$x'(t) = a(t)x(t) + b(t)x^{\alpha}(t) \quad \textit{où } \alpha \textit{ est un nombre réel différent de } 0 \textit{ et } 1.$$

En effet en divisant l'équation par $x^{\alpha}(t)$ et en utilisant le changement de variable astucieux $u(t) = x^{1-\alpha}(t)$, on obtient, car $u'(t) = (1-\alpha)x'(t)/x^{\alpha}(t)$, l'équation linéaire

$$u'(t) = (1-\alpha)a(t)u(t) + (1-\alpha)b(t)$$

qui a pour solution $u(t)$ donnée par le théorème. On en déduit alors $x(t) = u^{1/(1-\alpha)}(t)$.

12.4 Les outils graphiques de résolution

Nous venons de voir qu'il y a quelques équations différentielles ordinaires même non linéaires que l'on sait résoudre analytiquement mais dans la plupart des cas il est impossible de trouver l'expression de la solution. Or nous avons vu que dès que l'on prend en compte plusieurs paramètres, la modélisation conduit à des systèmes non linéaires. Dans le cas d'une équation non linéaire autonome $x'(t) = f(x(t))$, on peut avoir une idée du comportement des solutions en traçant la trajectoire de la solution qui est la projection sur $\mathbf{R}$ du graphe de la fonction $t \mapsto x(t)$. Si $f(x(t)) = x(t)$, pour une donnée de Cauchy $x_0 = 0$ la derivée est nulle et donc la solution est stationnaire et reste nulle au cours du temps. La trajectoire est limitée à un point, l'origine. Pour une donnée de Cauchy strictement positive, la dérivée est positive et la trajectoire est la portion de droite issue de x_0 qui part vers $+\infty$. Inversement si elle est négative, on obtient la portion de droite qui part vers $-\infty$. Donc si on veut maintenant tracer les solutions (ou courbes intégrales) en fonction du temps, on obtient selon la donnée initiale x_0 l'axe réel si $x_0 = 0$, une portion d'exponentielle e^t si $x_0 > 0$ et une portion de $-e^t$ dans le cas contraire. Dans le cas où $f(x(t))$ est non linéaire, on peut chercher les zéros de f pour obtenir les points stationnaires puis regarder vers où évoluent les points intermédiaires de chaque intervalle. Si les trajectoires convergent vers un point stationnaire, on a un point stable qui "attire" les solutions. C'est à dire que si la donnée initiale est voisine de ce point, la solution va converger vers le point. Sinon le point est instable et la solution s'en écarte.

Dans le cas d'un système de deux équations, on pourra faire une représentation des trajectoires dans le plan $\mathbf{R} \times \mathbf{R}$. Si l'on considère le système linéaire simple :

$$\left\{ \begin{array}{rcl} x'(t) & = & y(t) \\ y'(t) & = & -x(t) \end{array} \right. ,$$

en partant d'une donnée initiale $(x_0, y_0) = (0, 0)$ on obtient encore des dérivées nulles et l'origine O est un point stationnaire. Si on part d'un autre point M du plan, les dérivées déterminent un vecteur perpendiculaire au vecteur $\overrightarrow{OM}$ et donc on va décrire un cercle centré à l'origine. Les trajectoires sont dans ce cas des cercles concentriques. Certaines perturbations non linéaires du système pourront donner des trajectoires spirales, qui se rapprocheront ou s'écarteront de l'origine selon la stabilité de ce point stationnaire. Prenons pour exemple le système non linéaire

$$\left\{ \begin{array}{rcl} x'(t) & = & y(t) + x(t)(1 - x(t)^2 - y(t)^2) \\ y'(t) & = & -x(t) + y(t)(1 - x(t)^2 - y(t)^2) \end{array} \right. ,$$

que l'on peut considérer comme une perturbation du système linéaire précédent. Quand x et y sont nuls les dérivées sont nulles et donc l'origine O est un point stationnaire. Quand x et y sont sur le cercle unité d'équation $x^2 + y^2 = 1$, la perturbation s'annule et l'on retrouve le système linéaire et donc le cercle unité comme trajectoire. La question est de savoir quel est le comportement de la solution pour d'autres conditions initiales. Ceci peut être facilement visualisé en traçant le plan de phase (ensemble des trajectoires) dans une fenêtre autour de O (Figure 12.1). On s'aperçoit alors que l'origine est instable alors que le cercle unité est stable car il attire les trajectoires issues de données de Cauchy à l'extérieur comme à l'intérieur du cercle en dehors de O.

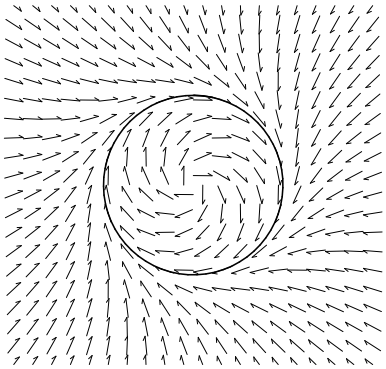


FIG. 12.1 – Visualisation des trajectoires au voisinage du cercle unité.

Un autre système non linéaire peut s'écrire

$$\begin{cases} x'(t) &= y(t) \\ y'(t) &= -2x(t)(2x(t)^2 - 1) \ , \end{cases}$$

pour lequel on a seulement les trois points stationnaires $(-\frac{1}{\sqrt{2}}, 0), (0, 0)$ et $(\frac{1}{\sqrt{2}}, 0)$. Cette fois ci les trajectoires ne sont plus des spirales mais des courbes fermées dont la forme dépend de la position de la donnée initiale par rapport aux points stationnaires (Figure 12.2).

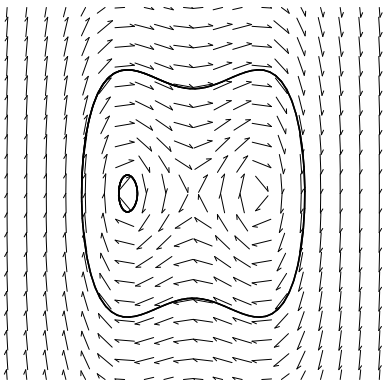


FIG. 12.2 – Visualisation des trajectoires au voisinage des points stationnaires.

Un autre outil d'investigation peut être de représenter les déformées au cours du temps par le même système d'un ensemble de données initiales. Dans ce dernier exemple on va prendre l'ensemble constitué du cercle de rayon 0.2 centré en $(-1, 0)$. L'évolution de cet ensemble au cours du temps est représentée sur la Figure 12.3. Il est à noter que pour ce système l'aire des déformés successifs reste constante

malgré de très grandes déformations. En effet le point $(-0.8, 0)$ est au voisinage du point stationnaire $(-\frac{1}{\sqrt{2}}, 0)$ et sa trajectoire est une courbe de forme elliptique autour de ce point. En revanche le point $(-1.2, 0)$ a une trajectoire qui sort de l'attraction de ce point comme on le voit avec les deux courbes en continu tracées sur la figure 12.2.

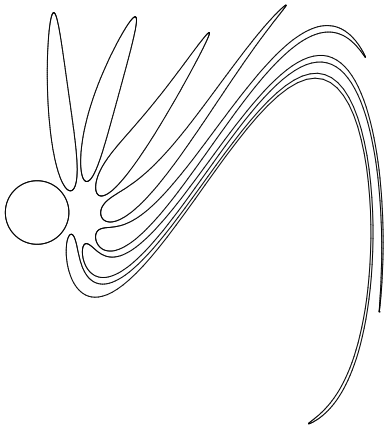


FIG. 12.3 – Cercle initial et ses déformés aux temps t=0.4, t=0.8, t=1.2, t=1.6, t=2, t=2.4 et t=2.8.

12.5 Les outils numériques de résolution

Pour un simple système de deux équations

$$\begin{cases} x'(t) &= f(x(t), y(t)) \\ y'(t) &= g(x(t), y(t)), \end{cases}$$

les comportements possibles sont infinis selon la définition des fonctions f et g. Il sera parfois très difficile de représenter correctement les trajectoires et d'imaginer le comportement des solutions et donc l'évolution du phénomène que l'on modélise. Cependant, on pourra toujours avoir recours à un calcul approché de la solution. Le cas le plus simple est de représenter la fonction continue $x(t)$ par une fonction discrète en calculant des valeurs approchées en des points choisis a priori et calculés au fur et à mesure. Dans le cas d'une seule équation $x'(t) = f(x(t))$ cela revient à construire une suite de valeurs que l'on notera x^n en partant de $x^0 = x_0$. Il faut d'une part approcher la dérivée $x'(t)$ et d'autre part faire en sorte que la solution approchée ne dérive pas progressivement vers une valeur approchée sans lien avec la solution exacte. Voilà l'objectif que l'on se fixe pour construire le schéma numérique. Le plus classique est le schéma explicite introduit par Euler. Il revient à construire la suite x^n tout simplement par l'algorithme suivant :

$$\begin{aligned} x^0 &= x_0 \\ x^{n+1} &= x^n + hf(x^n). \end{aligned}$$

A la première itération cela revient à écrire $(x^1 - x^0) \,/\, h = f(x^0)$ et donc la seule approximation est sur la dérivée qui est approchée de façon usuelle par le taux d'accroissement $(x(t_0 + h) - x(t_0)) \,/\, h$ qui tend bien vers la dérivée quand h tend vers zéro. Cette méthode est aussi connue sous le nom d'algorithme de la tangente car on construit les itérés successifs en prenant la droite qui a pour pente la tangente à la courbe solution.

Sous des hypothèses de régularité sur la fonction f, on peut montrer que la méthode converge et est du premier ordre. C'est-à-dire que l'on a :

$$\mid x(t_0 + nh) - x^n \mid = \mathcal{O}(h).$$

Remarque 12.5.1 *Cette estimation ne tient pas compte des erreurs dues aux arrondis des opérations en machine. On pourrait croire que le mieux est de prendre un pas de temps h tout petit pour obtenir de bonnes solutions approchées mais plus le pas est petit, plus il faut faire d'itérations pour atteindre le temps final T. Et donc on commet plus d'erreurs de calcul et on peut arriver à un stade où les erreurs d'arrondis sont supérieures aux erreurs de la méthode !*

12.6 Les équations différentielles linéaires du second ordre

Il existe une autre classe d'équations différentielles dont on sait déterminer la solution réelle analytiquement.

Définition 12.6.1 *On appelle* équation différentielle linéaire du second ordre à coefficients constants *une équation de la forme $ax''(t) + bx'(t) + cx(t) = 0$ où a, b et c sont des nombres réels (a non nul).*

Remarque 12.6.2 *Dans ce cas le problème de Cauchy s'écrit*

$$\begin{aligned} ax''(t) + bx'(t) + cx(t) = 0 \ \ \forall t \in I \\ x(t_0) = x_0 \ \ et \ \ x'(t_0) = x_1 \end{aligned}$$

où x_1 est une deuxième donnée initiale.

Définition 12.6.3 *On appelle* équation caractéristique *l'équation du trinôme du second degré*

$$a\lambda^2 + b\lambda + c = 0 \ .$$

Théorème 12.6.4 *Le problème de Cauchy ci-dessus admet une solution unique que l'on calcule à partir des racines du trinôme caractéristique.*
i) Si λ_1 et λ_2 sont deux racines réelles distinctes, la solution est de la forme $x(t) = \alpha_1 e^{\lambda_1 t} + \alpha_2 e^{\lambda_2 t}$;
ii) Si λ est une racine double, la solution est de la forme $x(t) = (\alpha_1 + \alpha_2 t)\, e^{\lambda t}$;
iii) Si $\lambda + i\mu$ et $\lambda - i\mu$ sont deux racines complexes conjuguées, la solution est de la forme $x(t) = (\alpha_1 \cos(\mu t) + \alpha_2 \sin(\mu t))\, e^{\lambda t}$.
Dans les trois cas les coefficients α_1 et α_2 sont déterminés de façon unique par résolution d'un système de 2 équations à 2 inconnues dont le second membre est constitué des données initiales.

Noter, que dans les autres cas, la solution $x(t)$ est complexe.
Dans les trois cas les deux fonctions utilisées pour définir $x(t)$ sont linéairement indépendantes [1]. Il suffit de vérifier que ce sont des solutions générales de l'équation et que les données initiales permettent de déterminer $x(t)$ de façon unique.

i) Pour $x(t) = \alpha_1 e^{\lambda_1 t} + \alpha_2 e^{\lambda_2 t}$ il vient

$$\begin{aligned} ax''(t) + bx'(t) + cx(t) &= \alpha_1 \left(a\lambda_1^2 + b\lambda_1 + c\right) e^{\lambda_1 t} \\ &+ \alpha_2 \left(a\lambda_2^2 + b\lambda_2 + c\right) e^{\lambda_2 t} \\ &= 0. \end{aligned}$$

[1] Voir l'Appendice pour cette notion.

Alors les coefficients α_1 et α_2 sont déterminés de façon unique par le système

$$\begin{aligned} \alpha_1 e^{\lambda_1 t_0} + \alpha_2 e^{\lambda_2 t_0} &= x_0 \\ \alpha_1 \lambda_1 e^{\lambda_1 t_0} + \alpha_2 \lambda_2 e^{\lambda_2 t_0} &= x_1 \end{aligned}$$

de déterminant principal $(\lambda_2 - \lambda_1) e^{\lambda_1 t_0} e^{\lambda_2 t_0}$ non nul.

ii) Pour $x(t) = (\alpha_1 + \alpha_2 t)\, e^{\lambda t}$ il vient

$$\begin{aligned} ax''(t) + bx'(t) + cx(t) &= (a\lambda^2 + b\lambda + c)\, (\alpha_1 + \alpha_2 t)\, e^{\lambda t} \\ &+ \alpha_2\, (2a\lambda + b)\, e^{\lambda t} \\ &= 0 \end{aligned}$$

car $\lambda = -b/2a$ est racine double. Alors les coefficients α_1 et α_2 sont déterminés de façon unique par le système

$$\begin{aligned} \alpha_1 e^{\lambda t_0} + \alpha_2 t_0 e^{\lambda t_0} &= x_0 \\ \alpha_1 \lambda e^{\lambda t_0} + \alpha_2\, (1 + \lambda t_0)\, e^{\lambda t_0} &= x_1 \end{aligned}$$

de déterminant principal $e^{2\lambda t_0}$ non nul.

iii) Pour $x(t) = (\alpha_1 \cos(\mu t) + \alpha_2 \sin(\mu t))\, e^{\lambda t}$ il vient

$$\begin{aligned} ax''(t) + bx'(t) + cx(t) &= (a\lambda^2 - a\mu^2 + b\lambda + c)\, (\alpha_1 \cos(\mu t) + \alpha_2 \sin(\mu t))\, e^{\lambda t} \\ &+ (2a\lambda\mu + b\mu)\, (-\alpha_1 \sin(\mu t) + \alpha_2 \cos(\mu t))\, e^{\lambda t} \\ &= 0 \end{aligned}$$

car si $\lambda + i\mu$ est racine du trinôme on a

$$a\, (\lambda + i\mu)^2 + b\, (\lambda + i\mu)\, + c = 0.$$

Ce qui est équivalent à

$$\begin{aligned} a\lambda^2 - a\mu^2 + b\lambda + c &= 0 \ \text{(partie réelle)} \\ 2a\lambda\mu + b\mu &= 0 \ \text{(partie imaginaire)}. \end{aligned}$$

Alors les coefficients α_1 et α_2 sont déterminés de façon unique par le système

$$\begin{aligned} \alpha_1 \cos(\mu t_0) e^{\lambda t_0} + \alpha_2 \sin(\mu t_0) e^{\lambda t_0} &= x_0 \\ \alpha_1\, (-\mu \sin(\mu t_0) + \lambda \cos(\mu t_0))\, e^{\lambda t_0} + \alpha_2\, (\mu \cos(\mu t_0) + \lambda \sin(\mu t_0))\, e^{\lambda t_0} &= x_1 \end{aligned}$$

de déterminant principal $\mu e^{2\lambda t_0}$ non nul.

Comme dans le cas du premier ordre on peut calculer la solution de l'équation avec second membre.

Théorème 12.6.5 *Soit f une fonction continue, le problème de Cauchy*

$$\begin{aligned} ax''(t) + bx'(t) + cx(t) = f(t) \ \ \forall t \in I \\ x(t_0) = x_0 \ \ et \ \ x'(t_0) = x_1 \end{aligned}$$

admet une solution unique, obtenue à partir de deux solutions particulières indépendantes $(\tilde{x}_1, \tilde{x}_2)$ *de l'équation homogène, sous la forme*

$$x(t) = \alpha_1(t)\tilde{x}_1(t) + \alpha_2(t)\tilde{x}_2(t)$$

où α_1 *et* α_2 *satisfont*

$$\begin{aligned} \alpha_1'(t)\tilde{x}_1(t) + \alpha_2'(t)\tilde{x}_2(t) &= 0 \\ \alpha_1'(t)\tilde{x}_1'(t) + \alpha_2'(t)\tilde{x}_2'(t) &= \frac{1}{a}f(t). \end{aligned}$$

Avec l'expression de $x(t)$ dans le théorème on obtient

$$x'(t) = \alpha_1'(t)\tilde{x}_1(t) + \alpha_1(t)\tilde{x}_1'(t) + \alpha_2'(t)\tilde{x}_2(t) + \alpha_2(t)\tilde{x}_2'(t)$$

et

$$\begin{aligned} x''(t) &= \alpha_1''(t)\tilde{x}_1(t) + 2\alpha_1'(t)\tilde{x}_1'(t) + \alpha_1(t)\tilde{x}_1''(t) \\ &+ \alpha_2''(t)\tilde{x}_2(t) + 2\alpha_2'(t)\tilde{x}_2'(t) + \alpha_2(t)\tilde{x}_2''(t). \end{aligned}$$

En reportant ces expressions dans l'équation il vient

$$\begin{aligned} & \alpha_1(t)\ (a\tilde{x}_1''(t) + b\tilde{x}_1'(t) + c\tilde{x}_1(t))\ + \alpha_2(t)\ (a\tilde{x}_2''(t) + b\tilde{x}_2'(t) + c\tilde{x}_2(t)) \\ + \ & a\ (\alpha_1''(t)\tilde{x}_1(t) + \alpha_1'(t)\tilde{x}_1'(t) + \alpha_2''(t)\tilde{x}_2(t) + \alpha_2'(t)\tilde{x}_2'(t)) \\ + \ & b\ (\alpha_1'(t)\tilde{x}_1(t) + \alpha_2'(t)\tilde{x}_2(t))\ + a\ (\alpha_1'(t)\tilde{x}_1'(t) + \alpha_2'(t)\tilde{x}_2'(t))\ = f(t). \end{aligned}$$

Ce qui montre que les conditions du théorème sur les fonctions α_1 et α_2 sont équivalentes à x solution (les 2 premiers termes sont nuls car $\tilde{x}_1$ et $\tilde{x}_2$ sont des solutions de l'équation homogène et le troisième terme est la dérivée de la première condition). Il reste à s'assurer que ces fonctions sont bien déterminées. Or le déterminant principal du système qu'elles vérifient est non nul car les deux solutions sont linéairement indépendantes. Ainsi ces fonctions α_1 et α_2 sont déterminées à une constante près que l'on déduit des conditions initiales.

12.7 Conclusion

L'étude et l'approximation des systèmes d'équations différentielles est un vaste champ qui suscite encore de nombreuses recherches. Nous n'en voyons dans ce cours qu'un tout petit aperçu. La présentation ci-dessus a pour objet de faire sentir la complexité du domaine et son intérêt. L'étude même locale de ces modèles permet de mieux comprendre des phénomènes complexes, chaotiques ou turbulents. Mais la route est longue et le domaine occupera encore de nombreuses générations de mathématiciens.

Chapitre 13

Introduction aux courbes planes

Un des objectifs de ce chapitre est d'essayer de donner un aperçu, une idée, de ce que peut être l'activité (de recherche) en mathématiques, quelles sont les questions qui peuvent intéresser un mathématicien. Pour cela, nous allons suivre un cheminement du "concret " à l'" abstrait", en partant de situations quotidiennes, qui feront apparaître une classe d'"êtres abstraits", d'objets mathématiques : les *courbes planes* (cf. §1 et §3), que nous essayerons alors d'étudier d'un point de vue (purement) mathématique (cf. §2 et §3). L'étude suivra deux axes *a priori* distincts, l'un plus relié à l'algèbre ou à la théorie des nombres (cf. §2), l'autre s'inscrivant plutôt dans une approche géométrique (cf. §3). À travers ces orientations nous espérons convaincre que les mathématiques ne sont pas aussi compartimentées que l'on peut penser et que les trois disciplines que sont l'*analyse*, l'*algèbre* et la *géométrie* ont des champs d'application communs.

La vocation de ce chapitre est donc de "montrer des mathématiques" à travers des questions diverses liées aux courbes planes, plutôt que de partir d'une définition arbitraire et de faire une étude exhaustive de leurs propriétés. En particulier, nous mettons en fin de paragraphe des " *sujets d'étude*" (cf. §2 et §3) qui présentent (sans démonstrations qui dépasseraient le cadre de ce chapitre) des énoncés de théorèmes importants dans l'étude des courbes, mais qui ne sont pas essentiels à la compréhension du reste des paragraphes en question.

13.1 Une approche concrète de la notion de courbe plane

Nous allons dans ce paragraphe étudier successivement quelques exemples concrets, afin de faire apparaître une première définition de la notion de *courbe plane* (que nous affinerons au paragraphe 3). Les courbes planes (telles que nous les envisageons dans ce chapitre) vont se définir comme des sous-ensembles du plan euclidien que nous identifierons à $\mathbf{R}^2$ muni de sa base canonique. Un point M du plan s'identifie alors à la donnée de son couple de coordonnées cartésiennes (x, y) (*i.e.* dans la base canonique). Dans les exemples qui suivent, les ensembles que nous considérerons vont apparaître dans des situations que l'on pourrait rencontrer dans la vie quotidienne. Nous donnerons une définition explicite de la notion de courbe (algébrique) plane à la fin de ce chapitre.

*

• Ce premier groupe d'exemples va prouver que les objets que nous allons étudier " existent " dans notre monde.

Exemple 13.1.1 *"Comprendre" ce que peuvent être les mathématiques c'est, sans doute, apprécier ce que cette science apporte aux autres, comment ces dernières peuvent (souvent) en dépendre pour assurer*

leur propre développement. Une très grande quantité d'exemples justifient cette thèse. Nous allons en choisir un très ancien et classique : cosmologie ou mécanique céleste *contre* étude des coniques. *En outre, cet exemple représente un modèle assez courant où mathématiques et physique se développent parallèlement avant de se rencontrer pour se nourrir l'une l'autre.*

◦ *Dans l'introduction du chapitre, un développement a déjà été consacré à la* cosmologie *et l'interaction avec les mathématiques. Rappelons donc qu'à l'époque antique (Platon, Aristote) jusqu'au Moyen-âge (en occident) les scientifiques adoptent le modèle de l'univers suivant : la Terre se trouve immobile au centre de l'univers (qui est fini et borné par les étoiles). Tous les astres connus alors (c'est-à-dire observables) sont des sphères qui tournent avec une trajectoire circulaire autour de la Terre.*

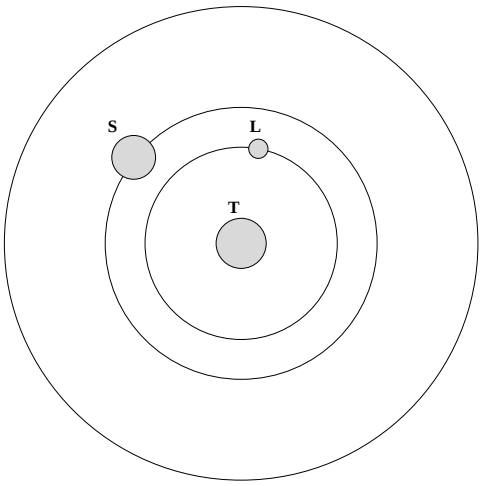


FIG. 13.1 – Le modèle aristotélicien de l'Univers

On saurait aujourd'hui, depuis le lycée au moins, décrire ce modèle dans le plan grâce à l'algèbre et aux équations de cercle. Toutefois l'histoire de la physique nous apprend que ce modèle ne correspond pas au réel. Sans parler des différentes étapes de construction (Ptolémée, Copernic, Galilée..., cf. introduction), c'est avec J. Kepler (astronome autrichien, 1571-1630) et ses textittrois lois que l'on arrive à un modèle cohérent de l'univers. Né de ses observations (et de celles de son maître, T. Brahé), Kepler énonce le principe suivant, que l'on appelle aujourd'hui première loi de Kepler *:*

"les orbites des planètes sont des ellipses dont le soleil est l'un des foyers."

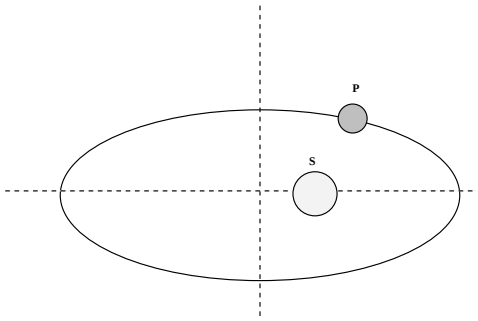


FIG. 13.2 – Le modèle keplérien de l'Univers

L'ellipse est une sorte de cercle étiré ou aplati. Mais c'est aussi un exemple de conique.

◦ *Parallèlement au développement de la cosmologie, l'*étude des coniques *(cf. §13.2.a) a participé à celui des mathématiques. Les coniques sont des ensembles de points du plan. Ils tirent leur nom du fait que l'on peut les obtenir comme des sections dans l'espace d'un cône par un plan.*

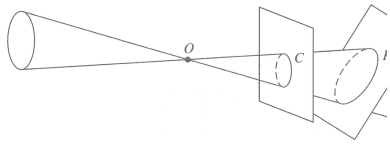


FIG. 13.3 – L'ellipse K et le cercle C

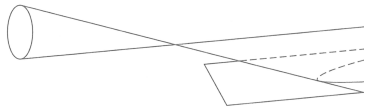


FIG. 13.4 – La parabole

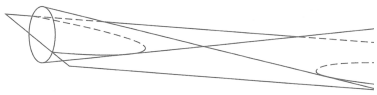


FIG. 13.5 – L'hyperbole

L'étude de ces ensembles débute probablement avec les Grecs dans l'antiquité (Apollonius, par exemple) qu'ils définissent alors par des propriétés purement géométriques. Par exemple, un cercle peut être défini comme le lieu des points du plan situé à une même distance d'un point que l'on appelle le centre du cercle. Il faut attendre le 17e siècle (et des mathématiciens comme Fermat, Pascal, Wallis...) pour décrire ces ensembles à partir d'équations cartésiennes. Par exemple, les ellipses ont des équations de la forme

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - 1 = 0,\ a, b \in \mathbf{R}^*,$$

et les hyperboles de la forme

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} - 1 = 0,\ a, b \in \mathbf{R}^*,$$

Il est important de noter qu'à l'origine, alors que l'on se représentait l'univers grâce au modèle d'Aristote, les mathématiciens commençaient déjà à s'intéresser aux ellipses qui n'entreront dans la cosmologie qu'avec Kepler (soit une vingtaine de siècles après).

Exemple 13.1.2 *Lorsqu'un canon propulse un obus, le point d'impact est déterminé par la donnée de sa trajectoire. Un problème naturel est donc de déterminer cette trajectoire. Les lois de la dynamique newtonienne nous permettent d'expliciter le chemin que va parcourir l'obus par le biais des*

mathématiques, en en représentant un modèle dans le plan euclidien ou l'espace. Précisément, si, pour simplifier, on ne tient pas compte de la résistance de l'air, on sait que la position de l'obus à l'instant t donné pourra s'exprimer dans le plan euclidien par :

$$\begin{cases} x(t) = (v_0 \times \cos(\alpha))t \\ \\ y(t) = -gt^2 + (v_0 \times \sin(\alpha))t \end{cases}$$

(v_0 étant la vitesse initiale, α l'angle du canon par rapport au sol, g la constante de pesanteur). Le temps apparaît comme un paramètre des équations. Une telle modélisation répond sans doute à la question. Mais il est intéressant de voir que, via t, les valeurs de l'abscisse et de l'ordonnée de l'obus ne sont pas indépendantes l'une de l'autre. En remarquant que l'on peut encore écrire

$$t = \frac{x(t)}{v_0 \times \cos(\alpha)}$$

et en substituant cette expression de t dans celle de $y(t)$, on trouve une relation cartésienne liant l'abscisse de l'obus à son ordonnée :

$$y = -g\frac{x^2}{v_0^2 \cos^2(\alpha)} + x\tan(\alpha)$$

Le temps n'apparaît plus explicitement dans cette expression, mais il est présent implicitement. Sous cette forme, la trajectoire se décrit donc comme le graphe d'une fonction et plus précisément comme une parabole (à v_0 et α fixé).

• Les exemples précédents nous ont montré que l'on peut côtoyer des objets ou des phénomènes qui ont un modèle dans le plan euclidien. Dans ce second groupe d'exemples nous allons voir qu'il peut être intéressant d'en étudier certaines propriétés, afin d'en tirer des informations qui peuvent s'avérer utiles dans la compréhension du "monde" qui nous entoure.

Exemple 13.1.3 *Le développement qui suit constitue un exemple concret où modèle mathématique et étude mathématique de ce modèle interviennent directement dans la vie quotidienne. Imaginons la situation suivante : vous décidez de construire une cloison perpendiculaire à un mur, mais vous ne disposez pas d'équerre (de chantier). Si vous demandez à un maçon comment faire face à ce genre de situation, il vous dira sûrement "30, 40, 50" ou "60, 80, 100" (ces triplets sont appelés triplets pythagoriciens, cf. exercice 13.2.15). Cette réponse s'explique grâce au théorème de Pythagore. Une équerre est par définition un triangle rectangle. Jusqu'ici, on a simplement reformulé le problème avec des mots issus du vocabulaire mathématique. Cela dit, on sait (au moins depuis l'antiquité grecque) que les longueurs (X, Y, Z) des côtés d'un triangle rectangle vérifient l'équation $Z^2 = Y^2 + X^2$. Ceci n'est rien d'autre que le fameux théorème de Pythagore. Trouver un triangle rectangle à coordonnées entières c'est donc résoudre dans $(\mathbf{Z}^*)^3$ l'équation suivante*

$$z^2 = y^2 + x^2$$

ou, ce qui est équivalent, résoudre dans $(\mathbf{Q}^*)^2$ l'équation suivante

$$y^2 + x^2 = 1$$

Par suite, $(3, 4, 5)$, par exemple, ou $(30, 40, 50)$ sont des solutions à l'équation précédente. Un guide pour monter la cloison s'obtient donc en plaçant un point B sur le mur à 40 centimètres du point A de départ de la cloison et en cherchant l'intersection du cercle de centre A, de rayon 30 centimètres et du cercle de centre B, de rayon 50 centimètres (ce qu'on pourra réaliser avec n'importe quel bout de ficelle d'au moins 51 centimètres de long).

Exemple 13.1.4 *Revenons à l'exemple du canon et cherchons cette fois le point d'impact de l'obus à vitesse initiale v_0 et angle α donnés, on est amené à résoudre l'équation $-g\frac{x^2}{v_0^2\cos^2(\alpha)} + x\tan(\alpha) = 0$. Une factorisation simple donne que l'unique solution (physiquement réalisable) est*

$$x = \frac{\tan(\alpha)v_0^2\cos^2(\alpha)}{g}$$

Maintenant, si v_0 est constante et si α varie, l'équation précédente peut être interprétée comme une équation à deux inconnues x et $\alpha \in [0, \pi/2[$. Chercher les points de la courbe revient ici à comprendre la dépendance du point d'impact par rapport à α, donc à "paramétrer" x en fonction de α.

*

Dans les exemples précédents, on a vu apparaître des ensembles de points du plan définis par des relations liant les coordonnées cartésiennes de la forme $E(x, y) = 0$ où E est une équation à deux variables x et y de la forme

$$E(x, y) = \sum_{0\leq i\leq m, 0\leq j\leq n} a_{i,j}x^i y^j, \quad a_{i,j} \in \mathbf{R} \ .$$

Ces équations sont appelées *équations algébriques* et E est un *polynôme* à deux variables. Dans un premier temps, on se contentera d'étudier ces sous-ensembles particuliers. Ceci nous amène à formuler la définition suivante :

On appelle *courbe (algébrique) plane* tout sous-ensemble Γ de $\mathbf{R}^2$ tel qu'il existe une fonction polynomiale $(x, y) \mapsto E(x, y)$ vérifiant

$$\Gamma := \big\{(x, y) \in I \times J \subset \mathbf{R}^2 \ \mid \ E(x, y) = 0\big\}$$

Étudier de tels ensembles, c'est d'abord se demander si l'on est capable de les dessiner. Est-il déjà facile de trouver des points du plan appartenant à ces ensembles, spécialement si l'on impose que les coordonnées soient rationnelles ou entières. Tel est le sujet du paragraphe suivant.

13.2 D'une approche diophantienne ...

[1]

Un des buts de ce paragraphe est de faire apparaître la notion de *paramétrisation*. Nous verrons notamment comment cette technique permet de trouver facilement des points (rationnels) sur les *coniques* (et parfois dans des cas plus généraux). Nous verrons à la fin du paragraphe que, pour les *cubiques*, la situation est à la fois plus compliquée et plus riche.

Un autre aspect de ce chapitre est de faire apparaître certaines "philosophies" en mathématiques : la classification et l'axiomatisation. À travers l'exemple des courbes, nous allons montrer pourquoi il peut être intéressant de savoir classer les objets, ou encore d'être capable d'axiomatiser pour développer une théorie générale.

[1] Diophante est un mathématicien grec qui vécut probablement aux alentours du 3e siècle après J.-C. En référence à ce mathématicien, on appelle équation diophantienne une équation algébrique pour laquelle on va chercher des solutions entières.

Plus largement, nous espérons que ce chapitre sera l'occasion pour le lecteur d'un premier contact avec des objets et certains énoncés (cf. sujets d'étude ci-après) de ce que l'on appelle la *géométrie algébrique*, dont les courbes algébriques sont des exemples.

*

La notion de paramétrisation intervient naturellement dans la recherche de points de ces ensembles et plus spécialement dans la recherche des *points rationnels*, *i.e.* des points du plan à coordonnées rationnelles.

Commençons par l'exemple le plus simple : celui des droites du plan. On sait, qu'un tel sous-ensemble de $\mathbf{R}^2$ est défini comme l'ensemble des points du plan dont les coordonnées vérifient une relation de la forme :

$$ax + by + c = 0,\ a, b, c \in \mathbf{R} \text{ (non tous nuls)}$$

Une telle droite contient-elle, contient-elle toujours, ou peut-elle ne pas contenir de points à coordonnées rationnelles ? Considérons quelques exemples : si la droite D_0 d'équation $x + y - \sqrt{2} = 0$ possède un point rationnel M_0 de coordonnées $(x_0, y_0) \in \mathbf{Q}^2$, alors la relation $x_0 + y_0 - \sqrt{2} = 0$ entraîne facilement que $\sqrt{2}$ est rationnel. Or on sait que cette assertion est fausse. En conséquence, notre droite ne peut contenir de points rationnels. Si l'on considère la droite D_1 d'équation $x + \sqrt{2}y - 1 = 0$, il est simple de vérifier que le point de coordonnées $(1, 0) \in \mathbf{Q}^2$ appartient à D_1. La droite D_1 peut-elle en posséder d'autres ? La réponse (négative) se déduit de l'exercice suivant :

Exercice 13.2.1 *a. Soient M_1 et M_2 deux points du plan à coordonnées rationnelles. Montrer que la droite qui relie M_1 et M_2 est rationnelle (une droite est dite rationnelle si elle peut être décrite par une équation cartésienne dont les coefficients sont des nombres rationnels).*

b. Soient L_1 et L_2 deux droites rationnelles du plan, montrer que, si leur intersection est non vide, leur point d'intersection est un point rationnel du plan.

Soit enfin la droite D_3 d'équation $x + y - 1 = 0$. Cette équation peut encore se réécrire en $y = 1 - x$, de sorte qu'un point de D_3 est rationnel si et seulement si son abscisse est un nombre rationnel. En particulier, il n'est pas difficile de se convaincre que, dans ce dernier cas, D_3 possède une infinité (dénombrable) de points rationnels. Une droite (comme D_3) qui possède une équation dont les coefficients sont rationnels s'appelle une *droite rationnelle*.

Remarque 13.2.2 *La droite D_3 possède aussi des points qui ne sont pas rationnels : par exemple,* $(1 - \sqrt{2}, \sqrt{2})$.

Exercice 13.2.3 *Soit D une droite rationnelle. Montrer qu'il existe une bijection $f : \mathbf{R} \to \mathbf{R}$ telle que l'application $\theta : \mathbf{R} \to \mathbf{R}^2$ définie par $x \mapsto (x, f(x))$ induise une bijection de $\mathbf{R}$ sur D et dont la restriction à $\mathbf{Q}$ induise une bijection sur l'ensemble des points rationnels de D.*

En se convainquant que ces résultats se généralisent, on peut résumer ces constatations dans l'énoncé suivant, dont nous laissons la démonstration en exercice :

Proposition 13.2.4 *Soit D une droite du plan. Alors :*

a. ou bien D ne possède aucun point rationnel ;

b. ou bien D possède un unique point rationnel ;

c. ou bien D possède une infinité (dénombrable) de points rationnels. Ce dernier cas se réalise si et seulement si D est rationnelle.

13.2.a Les coniques

Une conique C est un sous-ensemble de $\mathbf{R}^2$ défini par une relation de la forme

$$ax^2 + bxy + cy^2 + dx + ey + f = 0,\ (a, b, c, d, e, f) \in \mathbf{R}^6 \quad \text{(non tous nuls)}$$

Certains exemples de coniques sont connus depuis longtemps :

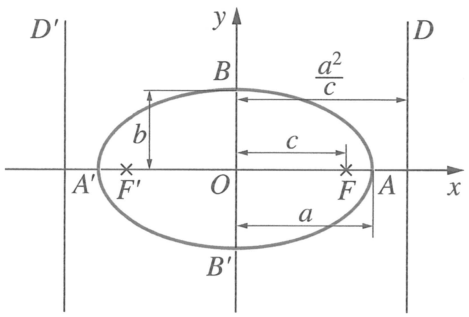


FIG. 13.6 – L'ellipse d'équation $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} - 1 = 0$

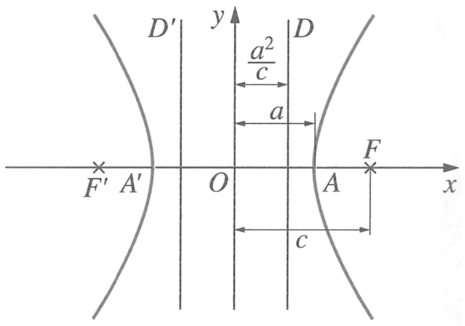


FIG. 13.7 – L'hyperbole d'équation $\dfrac{x^2}{a^2} - \dfrac{y^2}{b^2} - 1 = 0$

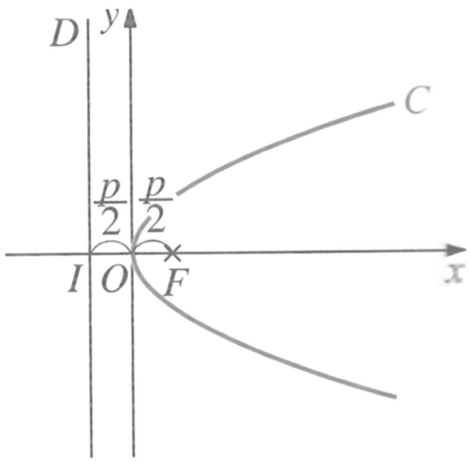


FIG. 13.8 – La parabole d'équation $y^2 - 2px = 0$

Tous ces ensembles contiennent-ils un point rationnel, ou même réel ? La question n'est pas aussi évidente que l'on peut le croire, puisque, par exemple, le sous-ensemble du plan défini par

$$x^2 + 1 = 0$$

est vide, alors que la conique $x^2 - 1 = 0$ est la réunion des deux droites $x - 1 = 0$ et $x + 1 = 0$.

Exercice 13.2.5 *Montrer que la conique C d'équation*

$$x^2 + y^2 - 3 = 0$$

contient un point réel, mais aucun point rationnel. Pour cela répondre aux questions suivantes :

a. *Trouver un couple $(x, y) \in \mathbf{R}^2$ tel que $x^2 + y^2 - 3 = 0$*
b. *Montrer que $C \cap \mathbf{Q}^2$ est non vide si et seulement s'il existe trois entiers naturels u, v, $w \neq 0$ tels que $u^2 + v^2 - 3w^2 = 0$ et tels qu'il n'existe pas d'entier naturel n qui divise à la fois u, v et w.*
c. *Montrer que, dans un tel triplet (u, v, w), 3 ne peut diviser ni u, ni v.*
d. *En déduire le résultat annoncé.*

Comme dans le cas des droites, nous allons nous intéresser aux coniques *rationnelles*, *i.e.* dont les six coefficients sont rationnels.

Changement de repère et classification des coniques

Dans ce paragraphe, on note (et on fixe) (e_1, e_2) la base canonique de $\mathbf{R}^2$ et $R := (O, e_1, e_2)$ le repère cartésien dans lequel les lettres x et y désignent l'abscisse et l'ordonnée du point $M = (x, y)$.

Les coniques se définissent à partir d'équations semblables, mais leurs représentations ont des allures bien différentes, comme le montrent les quelques exemples ci-dessus. Une idée assez naturelle (et que l'on retrouve assez souvent en mathématiques) est d'essayer pourtant de "classifier" ces objets.

Théorème 13.2.6 *Dans un système de coordonnées, toute conique est l'une des courbes suivante :*

a. *une ellipse d'équation* $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} - 1 = 0$;

b. une parabole d'équation $y = mx^2$ *;*

c. une hyperbole d'équation $\dfrac{x^2}{a^2} - \dfrac{y^2}{b^2} - 1 = 0$ *;*

ou encore, l'un des cas "dégénérés" suivants :

a. l'ensemble vide ;

b. un point ;

c. une droite ;

d. un couple de droites sécantes ;

e. un couple de droites parallèles.

Que peut signifier l'expression "dans un système de coordonnées" ? Considérons l'exemple de la conique C_1 définie par l'équation suivante :

$$y - ax^2 + bx + c = 0$$

avec $a \neq 0$. On peut alors écrire

$$ax^2 + bx + c = a\left(\left(x + \frac{b}{2}\right)^2 - \frac{b^2}{4a^2} + \frac{c}{a}\right)$$

Posons

$$\begin{cases} X = x + \dfrac{b}{2} \\ \\ Y = y - \dfrac{b^2 - 4ac}{4a} \end{cases}$$

L'équation de la conique C_1 peut encore se décrire dans les nouvelles *coordonnées* X et Y par

$$Y = aX^2$$

Ces opérations consistent à "changer de variables" ou à "changer de repère", en fixant comme origine du nouveau repère le point de coordonnées $(-b/2, (b^2 - 4ac)/4a)$ (dans l'ancien repère). Géométriquement, on a simplement effectué une *translation* du repère en déplaçant l'origine.

Remarque 13.2.7 *Ce calcul a déjà été rappelé en exercice en 6.2 pour le problème des solutions des équations du second degré. La forme factorisée précédente s'appelle la* forme canonique *du polynôme* $ax^2 + bx + c$*. C'est de cette manière qu'apparaît le discriminant.*

Soit C_2 la conique d'équation $x^2 + xy + y^2 - 2 = 0$. Considérons le changement de variables

$$\begin{cases} x &= \dfrac{1}{\sqrt{2}}(X - Y) \\ \\ y &= \dfrac{1}{\sqrt{2}}(X + Y) \end{cases}$$

Une équation cartésienne de C_2 en les coordonnées X et Y se déduit en écrivant :

$$\begin{aligned} x^2 + xy + y^2 - 1 &= 0 \\ \tfrac{1}{2}\left((X - Y)^2 + (X - Y)(X + Y) + (X + Y)^2\right) - 2 &= 0 \\ 3X^2 + Y^2 - 4 &= 0 \\ \frac{3X^2}{4} + \frac{Y^2}{4} - 1 &= 0 \end{aligned}$$

En remarquant que $\cos(\pi/4) = \sin(\pi/4) = \sqrt{2}/2$, on peut encore écrire *matriciellement* ce changement de coordonnées par

$$\begin{pmatrix} \cos\left(\frac{\pi}{4}\right) & -\sin\left(\frac{\pi}{4}\right) \\ \sin\left(\frac{\pi}{4}\right) & \cos\left(\frac{\pi}{4}\right) \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}$$

Qu'est devenu notre repère par ce changement de variables ? Géométriquement on a effectué une *rotation* de centre O et d'angle $\pi/4$.

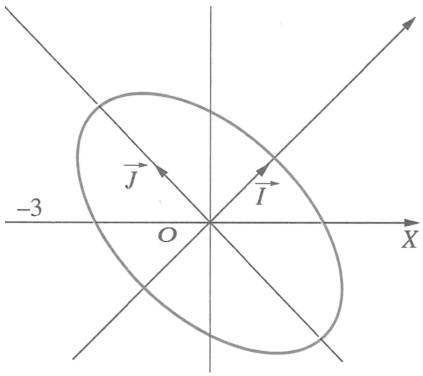


FIG. 13.9 – Le changement de repère

Avec les conventions du théorème 13.2.6, la conique C_2 est donc une ellipse.

Exercice 13.2.8 *Considérons la conique C d'équation*

$$xy - 1 = 0$$

Montrer qu'il existe un système de coordonnées dans lequel C est définie par l'équation cartésienne $X^2 - Y^2 - 1 = 0$. On donnera la matrice de changement de coordonnées.

Exercice 13.2.9 *Démontrer que le graphe de la fonction $h : \mathbf{R}\backslash\{1\} \to \mathbf{R}$ définie par $x \mapsto \dfrac{x+1}{x-1}$ est une hyperbole en déterminant un système de coordonnées dans lequel ce sous-ensemble de $\mathbf{R}^2$ se décrit par une équation de la forme $x^2/a^2 - y^2/b^2 - 1 = 0$.*

En s'inspirant de l'exemple de la parabole et de l'hyperbole, on peut trouver une démonstration générale du théorème 13.2.6 (cf. exercice 13.2.10 ci-dessous).

Exercice 13.2.10 *Le but de cet exercice est de prouver le théorème 13.2.6. On considère la conique C d'équation*

$$Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0 \ .$$

a. Si $B^2 - AC \neq 0$, montrer que par une translation puis une rotation, que l'on précisera, C est une ellipse ou une hyperbole (ou bien l'ensemble vide, un point ou deux droites sécantes).

b. Si $B^2 - AC = 0$, montrer que la courbe C est du type parabole.

*

Le théorème 13.2.6 a un intérêt évident dans l'étude des coniques. En effet, il ramène tous les ensembles que l'on peut décrire avec une équation de la forme

$$ax^2 + bxy + cy^2 + dx + ey + f = 0$$

(il y en a une infinité !) à huit types d'ensembles, dont on connaît la représentation graphique.

Exercice 13.2.11 *a. Dessiner la conique d'équation* $x^2 + 3y^2 - 1 = 0$.
b. Dessiner la conique d'équation $5x^2 - 2y^2 - 1 = 0$.

Cette dernière remarque justifie l'intérêt de la classification en même temps qu'elle fournit une première réponse à la question de la représentation graphique des coniques.

La question des points rationnels. Notion de paramétrisation

Dans ce paragraphe, nous allons adopter un autre point de vue pour aborder l'étude des coniques : la *paramétrisation*. Cette technique va consister à identifier les abscisse et ordonnée de tout point d'une conique donnée à des fonctions d'un même *paramètre* t, variant dans un intervalle de $\mathbf{R}$. Du point vue de la physique, si la conique représente la trajectoire d'un solide (cf. exemple 13.1.2), paramétrer la courbe c'est comprendre la dépendance de la position du solide en fonction du temps et être capable, à chaque instant t, de donner précisément cette position. Nous allons montrer comment cette technique permet d'établir le théorème 13.2.17 ci-après.

*

Dans ce paragraphe, on n'étudie que des coniques rationnelles, *i.e* des courbes définies par une équation de la forme

$$ax^2 + bxy + cy^2 + dx + ey + f = 0, \quad \text{avec} \quad a, b, c, d, e, f \in \mathbf{Q}.$$

Commençons par étudier l'exemple du cercle unité (cf. exemple 13.1.3). Les coordonnées x, y d'un point de ce cercle vérifient l'équation

$$x^2 + y^2 - 1 = 0 \ .$$

Le point A de coordonnées $(-1, 0)$ est clairement un point rationnel du cercle unité C, *i.e* un point du cercle unité dont les coordonnées sont des nombres rationnels. Considérons alors la droite D_t passant par les points A et $M_t = (0, t)$.

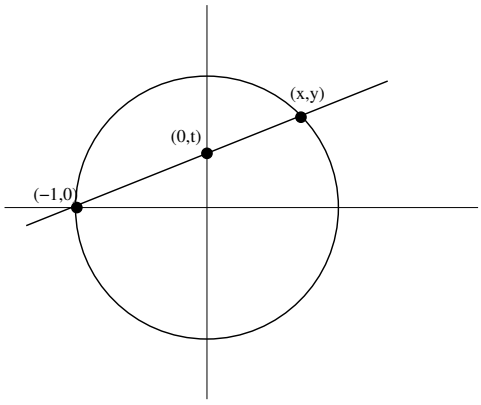


FIG. 13.10 – C et D_t

La droite D_t est alors décrite par l'équation $y - t(1 + x) = 0$. Géométriquement, on remarque que D_t coupe C en A, bien sûr, et en un autre point que l'on note A_t. Cela se vérifie également de manière algébrique en calculant les coordonnées des points d'intersection, *i.e.* en résolvant le système d'équations :

$$\begin{cases} x^2 + y^2 - 1 &= 0 \\ y - t(1 + x) &= 0 \end{cases}$$

Ces deux équations impliquent la relation

$$t^2(1 + x)^2 = y^2 = 1 - x^2$$

Ce qui se réécrit en l'équation du second degré en x

$$(1 + t^2)x^2 + 2t^2x + (t^2 - 1) = 0$$

La méthode de résolution de telles équations, qui a été rappelée au paragraphe précédent (cf. remarque 13.2.7), donne les deux solutions (distinctes) suivantes

$$\begin{cases} x = 0 \\ \text{ou} \\ x = \dfrac{1 - t^2}{1 + t^2} \end{cases}$$

Le point A_t a donc pour coordonnées le couple $\left(\dfrac{1 - t^2}{1 + t^2}, \dfrac{2t}{1 + t^2}\right)$.

Exercice 13.2.12 *Soit $B = (x, y)$ un point du cercle unité, différent de A. Montrer qu'il existe $t \in \mathbf{R}$ tel que B soit le second point A_t d'intersection de la droite D_t.*

À l'exception du point A, tous les points du cercle unité peuvent être décrits de cette manière (cf. exercice 13.2.12 ci-dessus). On peut donc décrire l'ensemble $C\backslash\{A\}$ comme l'ensemble des points du plan dont les coordonnées sont définies par

$$\begin{cases} x &= \dfrac{1 - t^2}{1 + t^2} \\ y &= \dfrac{2t}{1 + t^2} \end{cases}$$

avec $t \in \mathbf{R}$. Les coordonnées x et y d'un point de $C\backslash\{A\}$ sont des fonctions du *paramètre* t. On a donné une *paramétrisation rationnelle* du cercle C.

Remarque 13.2.13 *On peut remarquer que si l'on fait croître la valeur de t, la droite D_t se "rapproche" de la tangente de C en A et le point A_t se rapproche de A. En un certain sens, A est la "limite" des A_t. En conséquence on pourrait noter $A := A_\infty$ pour exprimer ce phénomène.*

Il est alors clair qu'un point de $C\backslash\{A\}$ est rationnel si $|t|$ est rationnel. Mais, du fait de la relation $t = y/(1 + x)$, il est également clair qu'un point rationnel de $C\backslash\{A\}$ est forcément défini par une valeur rationnelle du paramètre t. Autrement dit, avoir une paramétrisation rationnelle du cercle unité nous permet de constater que l'ensemble des points rationnels de C est infini dénombrable.

Remarque 13.2.14 *Il faut se convaincre que la méthode employée s'interprète géométriquement comme une projection de C sur l'axe des ordonnées par rapport au point A (cf. figure ci-dessus). En particulier, on peut remarquer que l'on a fait correspondre aux points rationnels de C (excepté A) les points rationnels d'une droite rationnelle (ici l'axe des ordonnées), qui contient une infinité dénombrable de points rationnels.*

Exercice 13.2.15 *Répondre aux questions suivantes.*

a. *Montrer qu'il existe un unique triplet $(X, Y, Z = 0) \in \mathbf{N}^3$ solution de l'équation $X^2 + Y^2 = Z^2$.*

b. *Montrer que l'ensemble des triplets (X, Y, Z) tels que*

$$\begin{cases} X^2 + Y^2 = Z^2 \\ \\ X,\ Y \in \mathbf{N},\ Z \in \mathbf{N}^* \end{cases}$$

est en bijection avec l'ensemble des points rationnels du cercle unité.

c. *Déduire de la paramétrisation rationnelle du cercle unité une description des triplets pythagoriciens primitifs.*

La méthode de *projection* utilisée pour le cercle peut être réemployée pour les coniques, que nous qualifions de *non dégénérées*, à savoir les ellipses (dont les cercles font partie) non ponctuelles, les hyperboles, et les paraboles.

Exercice 13.2.16 *Donner une paramétrisation des courbes suivantes :*

a. *La parabole d'équation $y - x^2 - x = 0$. Dessiner cette courbe.*

b. *L'hyperbole d'équation $x^2 - y^2 - 1 = 0$. Dessiner cette courbe.*

c. *L'ellipse d'équation $2x^2 + y^2 - 5 = 0$. Étudier la question des points rationnels. Dessiner la courbe.*

d. *La conique d'équation $x^2 - y^2 = (x - 2y)(x + y)$. Dessiner la courbe.*

e. *La conique d'équation $x^2 - y^2 = (x - 2y)(x + y) + 1$. À l'aide de transformations que vous expliciterez, donner un système de coordonnées dans lequel cette courbe se décrit par une équation de la forme de celles du théorème 13.2.6. Dessiner la courbe.*

Et l'on peut prouver le théorème suivant, qui résume la question des points rationnels des coniques rationnelles.

Théorème 13.2.17 *Soit C une conique rationnelle. Alors :*

a. *ou bien C ne possède aucun point rationnel ;*

b. *ou bien C possède un unique point rationnel ;*

c. *ou bien C possède une infinité dénombrable de points rationnels.*

Exercice 13.2.18 *Donner un exemple qui illustre chacun des trois cas du théorème 13.2.17.*

13.2.b Un exemple de cubique ou un exemple (non évident) de groupe

Dans ce paragraphe, nous allons nous intéresser au cas des *cubiques* (rationnelles), *i.e* aux sous-ensembles de $\mathbf{R}^2$ définis par une relation de la forme

$$ax^3 + bx^2y + cxy^2 + dy^3 + ex^2 + fxy + gy^2 + hx + uy + v = 0,\ (a, b, c, d, e, f, g, h, u, v) \in \mathbf{Q}^{10}$$

Nous verrons que la description que nous avions faite de l'ensemble des points rationnels des coniques, ne peut s'étendre au cas des cubiques de manière générale. Nous donnerons une explication et montrerons que le cas des cubiques est, par ailleurs, plus riche que celui des coniques.

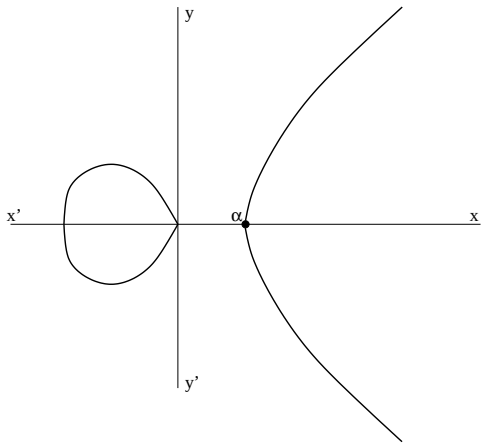


FIG. 13.11 – La cubique régulière d'équation $y^2 - x(x-\alpha)(x-\beta) = 0$

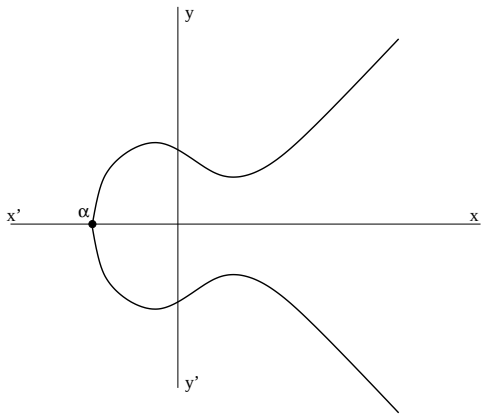


FIG. 13.12 – La cubique régulière d'équation $y^2 - x^3 - \alpha^3 = 0$

Des petites modifications de l'équation peuvent changer la nature de la courbe

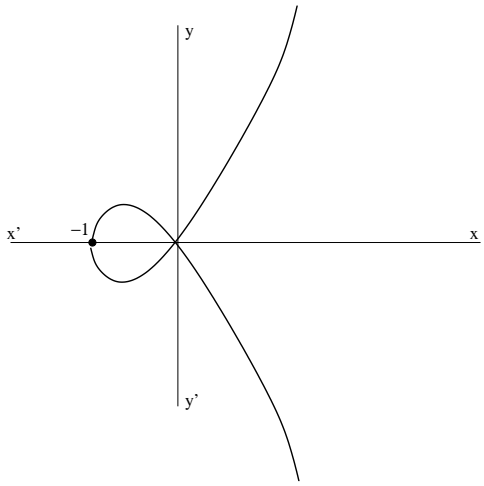


FIG. 13.13 – La cubique avec un point double d'équation $y^2 - x^3 - x^2 = 0$

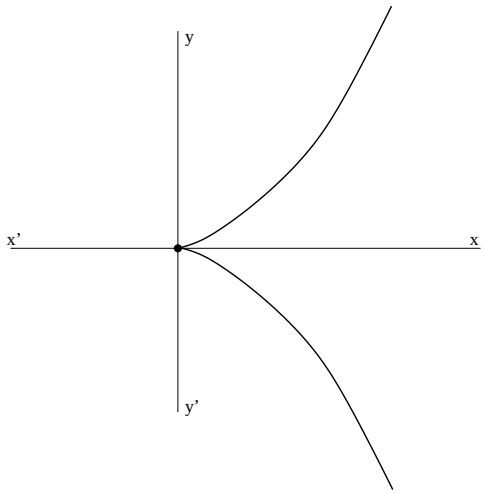


FIG. 13.14 – La cubique avec un point de rebroussement d'équation $y^2 - x^3 = 0$

étudions à nouveau la question des points rationnels de tels sous-ensembles. Il est parfois possible d'obtenir, comme dans le cas de coniques des paramétrisations rationnelles qui permettent de conclure.

Exercice 13.2.19 (Folium de Descartes) *Soit C la cubique d'équation*

$$x^3 + y^3 - 3xy = 0.$$

En considérant l'intersection de C avec la droite d'équation $y = tx$, déterminer une paramétrisation de C et l'ensemble de ses points rationnels.

Exercice 13.2.20 *Soit C la cubique d'équation*

$$(x-2y)(x^2+y^2)+y^2-x^2=0$$

En considérant l'intersection de C avec la droite d'équation $y=tx$, déterminer une paramétrisation rationnelle de C et l'ensemble de ses points rationnels.

Malheureusement, cette méthode ne permet pas de traiter le cas de toutes les cubiques. Moralement, cette difficulté découle du fait que l'intersection d'une droite et d'une conique possède (dans les bons cas) exactement deux points d'intersection (on a vu que calculer une telle intersection revient à résoudre une équation polynomiale du second degré). Mais dans le cas des cubiques ce n'est plus vrai en général et le théorème de Bézout assure que l'on peut avoir jusqu'à trois points d'intersection (cf. sujet d'étude 2). Voyons maintenant deux exemples.

• Dans le paragraphe précédent et dans l'exemple 13.1.3, on a vu que la conique d'équation x^2+y^2-1 possédait une infinité de points rationnels. Pourtant la cubique d'équation

$$x^3+y^3-1=0$$

ne possède pas d'autres points rationnels que $(0,1)$ et $(1,0)$. Cet énoncé peut être interprété comme un cas particulier du théorème de Fermat-Wiles, connu aussi sous le nom de "grand théorème de Fermat".[2]

• On peut montrer que la cubique d'équation $y^2-x^3-x=0$ ne possède pas de paramétrisation rationnelle.

Exercice 13.2.21 (TD) *Montrer que la cubique d'équation $y^2-x^3-x=0$ ne possède pas de paramétrisation rationnelle.*

La loi de groupes sur C

Dans ce paragraphe, C désigne la cubique d'équation

$$y^2-x^3-17=0\ .$$

Elle possède des points rationnels (le point $(2,5)$ par exemple). Nous allons étudier l'ensemble des points rationnels de C, ou plus exactement sa *structure de groupe*. Qu'est-ce-qu'un groupe ? On connaît des exemples de groupes (même si l'on a jamais employé ce mot). Par exemple, on sait additionner deux nombres réels et l'on sait que cette addition vérifie des règles :

la commutativité, c'est-à-dire le fait que $2+3=3+2$ par exemple ;

l'associativité, c'est-à-dire le fait $(2+3)+4=2+(3+4)$;

la règle du 0, c'est-à-dire le fait $0+2=2+0=2$

l'existence de l'inverse, c'est-à-dire le fait que l'on puisse soustraire.

[2] Pierre de Fermat (1601-1605) était un magistrat, conseiller du roi au Parlement de Toulouse. Il laisse notamment à la communauté mathématique deux énoncés, que l'on désigne comme les "petit" et "grand" théorèmes. L'énoncé du grand théorème de Fermat dit que l'équation

$$X^n+Y^n=Z^n$$

n'a pas de solutions $X,Y,Z\in\mathbf{Z}$ non nuls, si n est supérieur ou égal à 3. Fermat pensait avoir la preuve de ce théorème. On sait aujourd'hui que la preuve de Fermat devait être erronée, bien que l'on n'ait jamais retrouvé de preuve écrite d'une telle démonstration. C'est seulement en 1993(-1996) que Andrew Wiles a su trouver une démonstration (admise par la communauté mathématique).

L'addition et l'ensemble des règles de calculs forment ce que l'on appelle la *structure de groupe additif* de l'ensemble **R**. Cette structure donne à **R** une grande richesse (et lui confère le nom de groupe additif), comme on a pu le voir dans un chapitre précédent. Toutes ces règles semblent très naturelles quand il s'agit des nombres réels, rationnels, entiers relatifs... En fait, cette structure existe dans des situations insoupçonnées, comme le cas des points rationnels de C.

Comment "additionner" deux points rationnels de C ?

• Soient P et Q deux points rationnels de C. La droite rationnelle qui passe par P et Q va couper C en un troisième point d'intersection que l'on note $P*Q$. Si $P = Q$, on peut considérer la tangente en P à C, qui va couper C en $P * P$. Comme la courbe C est rationnelle, il n'est pas difficile de montrer que $P * Q$ (ou $P * P$) est encore rationnel.

Exercice 13.2.22 *Soient P et Q deux points rationnels de C. Montrer que $P * Q$ et $P * P$ sont rationnels.*

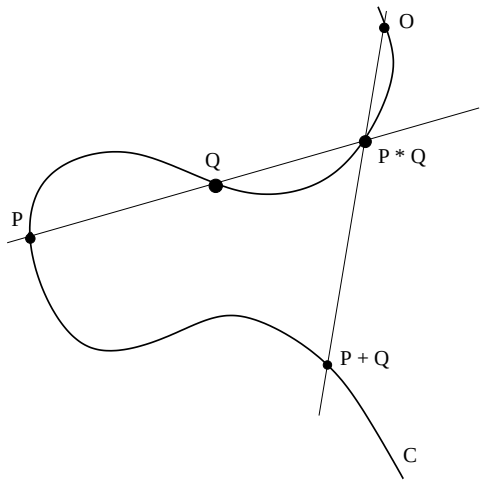


FIG. 13.15 – $P * Q$ et $P + Q$

La "loi de composition" $*$ permet donc d'associer à deux points rationnels de C (distincts ou non) un autre point rationnel de C. Toutefois ce n'est pas une *loi de groupe*. En effet, si l'on garde en mémoire l'exemple de **R**, la loi $*$ ne définit pas d'*élément neutre*, *i.e.* d'équivalent de 0. Pour palier cette carence, notons O un point rationnel de C (par exemple, $(2, 5)$). Si l'on joint $P * Q$ au point O, on obtient un autre point rationnel que l'on note cette fois $P + Q$. Par construction, le point O est bien un *élément neutre* de l'addition $+$ des points rationnels de C, *i.e* le point O vérifie

$$P + O = O + P = P$$

pour tout point rationnel P de C. Il suffit de s'en convaincre avec la figure ci-dessous

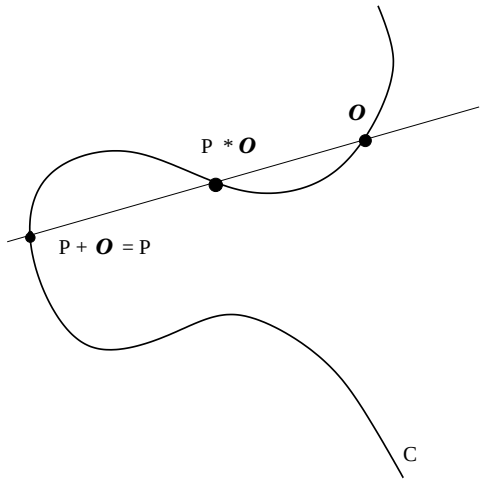


FIG. 13.16 – O l'élément neutre

• On va maintenant étudier les propriétés de cette addition. Commençons par la propriété de *commutativité*, *i.e.* le fait que

$$P + Q = Q + P$$

pour tout point rationnel P et tout point rationnel Q de C. Encore une fois cela découle de la construction, puisque la droite qui joint P à Q est la droite qui joint Q à P (cf. figures ci-dessus). Dans $\mathbf{R}$, tout élément x a un inverse pour l'addition (que l'on appelle l'opposé de x), c'est-à-dire un nombre $y \in \mathbf{R}$ tel que $x + y = y + x = 0$. L'addition sur l'ensemble des points rationnels de C que nous venons de définir possède également cette propriété. Soit Q un point rationnel de C. On définit son *inverse*, que l'on note $-Q$, de la manière suivante. La tangente en O à C recoupe C en S. Si l'on joint Q à S, on obtient encore un autre point rationnel que l'on note $-Q$.

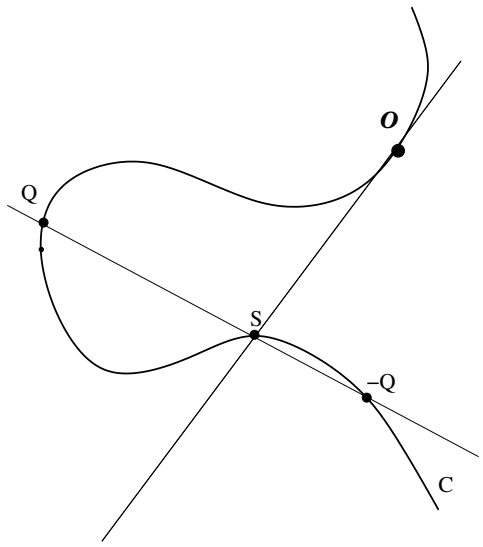


FIG. 13.17 – $-Q$ l'inverse de Q

Le point obtenu est-il bien l'inverse de Q, *i.e.* est-ce que

$$Q + (-Q) = O \ ?$$

Par construction, $Q * (-Q) = S$ et donc $Q + (-Q) = O$ (puisque la droite (OS) est tangente à C en O). Un point plus délicat est de vérifier l'*associativité* de l'addition, *i.e.* la propriété

$$(P + Q) + R = P + (Q + R)$$

pour tout triplet (P, Q, R) de points rationnels de C. Nous laissons cette vérification en exercice.

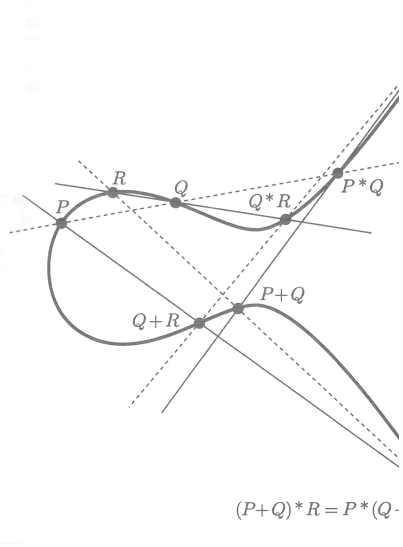


FIG. 13.18 – P, Q, R, $P * Q$, $P + Q$, $Q * R$, $Q + R$

Exercice 13.2.23 *a. Soient C_1 et C_2 deux cubiques qui se coupent en neuf points d'intersection. Soit D une cubique qui passe par huit de ces neuf points d'intersection. En raisonnant sur les coefficients, montrer que D passe forcément par le neuvième point d'intersection.*

b. En utilisant la question 1 et en s'inspirant de la figure ci-dessus, montrer que l'addition que l'on a définie sur l'ensemble des points rationnels de C est associative, i.e. montrer que $(P + Q) + R = P + (Q + R)$.

La loi + et ses propriétés font de l'ensemble des points rationnels de C un objet comparable, de ce point de vue, à l'ensemble $\mathbf{R}$ des nombres réels. On parlera désormais du *groupe* des points rationnels de C. Disposer d'une telle structure n'est pas du tout gratuit dans l'étude des points rationnels des cubiques qui, éventuellement, ne seraient pas paramétrisables. Une simple traduction des propriétés assure que l'on peut trouver un nouveau point rationnel en en additionnant deux autres. Mais en réalité on a bien plus :

Théorème 13.2.24 (Mordell,1923) *Si C_0 est une cubique (non singulière) qui possède un point rationnel, alors le groupe des points rationnels de C_0 est de type fini.*

Sans définir précisément les expressions “non singulières” et “de type fini”, cet énoncé dit simplement qu'il existe un nombre fini de points rationnels $P_1, \ldots, P_n$ à partir desquels on peut atteindre tout point rationnel de C_0 par un nombre fini d'additions et de soustractions.[3]

Exercice 13.2.25 *On pourra s'aider du sujet d'étude 2. Répondre aux questions suivantes :*

a. Vérifier que les points $P = (-2, 3)$ et $Q = (-1, 4)$ sont des points rationnels de C.

b. Calculer $P + Q$, $2P := P + P$ et $2Q$.

*

Avec l'exemple de la cubique C, on a vu que certains objets, de définitions très différentes (l'ensemble des nombres réels et l'ensemble des points rationnels de C), pouvaient partager des propriétés communes, ne pas être si différents du point de vue mathématique. Il est souvent intéressant d'exhiber ces propriétés communes pour définir, abstraitement, une classe d'objets que l'on va étudier pour eux-mêmes, *i.e* d'axiomatiser ces propriétés pour développer une théorie. Par exemple, on peut définir, de manière générale, la notion de *groupe* (abélien) comme suit :

Définition 13.2.26 *On dit qu'un ensemble non vide G est un groupe abélien si cet ensemble est muni d'une application*

$$\mu : G \times G \to G$$

qu'on appelle loi de composition et qui vérifie les propriétés suivantes :

a. $\forall f, g, h \in G$, $\mu(f, \mu(g, h)) = \mu(\mu(f, g), h)$;

b. $\forall f, g \in G$, $\mu(f, g) = \mu(g, f)$;

c. $\exists\ e \in G$ tel que $\mu(e, g) = \mu(g, e) = g$, $\forall g \in G$;

d. $\forall g \in G$, $\exists! g' \in G$, $\mu(gg') = \mu(g'g) = e$.

Exercice 13.2.27 *Vérifier que l'ensemble des points rationnels de C est un groupe au sens de cette définition.*

Sujet d'étude 1 : le théorème de Bézout

Dans cette étude, nous revenons à des courbes algébriques planes quelconques, *i.e.* définies par une équation polynomiale $P(x, y) = 0$.

Définition 13.2.28 *Si $P(x, y) = \sum_{0 \leq i \leq m, 0 \leq j \leq n} a_{i,j} x^i y^j$ avec $a_{i,j} \in \mathbf{R}$ est une fonction polynomiale, on appelle* degré *de P le plus grand degré des monômes qui la composent,* i.e. *le plus grand des entiers $i + j$ tel que $a_{i,j}$ soit non nul.*

• Soit C est une courbe algébrique définie par $P(x, y) = 0$. On dit qu'un polynôme non constant P_i est un facteur irréductible de P si l'on peut écrire P sous la forme

$$P = P_i^{\alpha_i} Q, \quad \alpha_i \in \mathbf{N}^*$$

[3]Louis J. Mordell (1888,1972) est un mathématicien américain qui s'est intéressé aux questions diophantiennes. Dans les années 1920, il démontre une conjecture de H. Poincaré, que nous avons énoncée ci-dessus (cf. théorème 13.2.24). à cette époque, il propose également une conjecture, appelée “conjecture de Mordell”, sur les points rationnels d'une classe de courbes algébriques. Cette conjecture célèbre ne deviendra un théorème (de Faltings) qu'en 1983.

et si P_i ne peut être décomposé, à son tour, comme produit de polynômes (non constants). Un théorème d'algèbre commutative (hors programme) assure que tout polynôme non constant $P(x,y)$ peut être décomposé en produit de facteurs irréductibles

$$P(x,y) = \prod_{i=1}^{n} P_i^{\alpha_i}(x,y)$$

Par suite, la relation $P(x,y) = 0$ équivaut à la condition $\exists 1 \leq i \leq n, \quad P_i(x,y) = 0$. Géométriquement, cela signifie exactement que la courbe C est la réunion des courbes C_i d'équations $P_i(x,y) = 0$. Dans ce cas, on appelle *équation minimale* de C l'équation

$$\prod_{i=1}^{n} P_i(x,y) = 0$$

Exemple 13.2.29 *Les facteurs irréductibles de $x^2 - y^2$ sont les polynômes $x - y$ et $x + y$, de sorte que la courbe C d'équation $x^2 - y^2 = 0$ est la réunion des droites $y - x = 0$ et $x + y = 0$.*

Exercice 13.2.30 *Montrer que la courbe d'équation $(x^2 - y^2)^2 - x^2 + y^2 = 0$ est la réunion de deux droites et d'une conique dont on précisera les équations.*

Exemple 13.2.31 *Les facteurs irréductibles de $x^4 + x^2y^2 - x^2 - yx^2 - y^3 + y$ sont les polynômes $y - x^2$ et $x^2 + y^2 - 1$ de sorte que la courbe d'équation $x^4 + x^2y^2 - x^2 - yx^2 - y^3 + y = 0$ est la réunion des courbes $y - x^2 = 0$ et $x^2 + y^2 - 1 = 0$.*

- Voici un énoncé (faible) du théorème de Bézout :

Théorème 13.2.32 *Soient C_1 et C_2 deux courbes algébriques planes d'équations minimales $P_1 = 0$ et $P_2 = 0$, sans facteur irréductible commun. On suppose que le degré de P_1 est n_1 et n_2 celui de P_2. Alors l'ensemble des points d'intersection de C_1 et C_2 est fini et son cardinal $\iota(C_1, C_2)$ (éventuellement nul) vérifie :*

$$\iota(C_1, C_2) \leq n_1 n_2.$$

Exercice 13.2.33

a. Combien peut-on obtenir de points d'intersection avec deux coniques sans facteurs irréductibles communs ?

b. Calculer les coordonnées des points d'intersection des courbes $C_1 : x^2 + 2x + y^2 = 0$ et $C_2 : x^2 - 2x + y^2 = 0$. Faire une figure.

c. Calculer les coordonnées des points d'intersection des courbes $C_1 : x^2 + y^2 - 1 = 0$ et $C_2 : x^2/2 + y^2 - 1 = 0$. Faire une figure.

d. Décrire l'ensemble des points d'intersection des courbes $C_1 : x^2 - y^2 = 0$ et $C_2 : x^2 - 2y^2 + xy = 0$. Pourquoi n'y-a-t'il pas de contradiction avec le théorème de Bézout. Faire une figure.

e. Calculer les coordonnées des points d'intersection des courbes $C_1 : x^2 + y^2 - 1 = 0$ et $C_2 : x^2/4 + 4y^2 - 1 = 0$. Faire une figure.

f. Calculer les coordonnées des points d'intersection des courbes $C_1 : x - 2y = 0$ et $C_2 : x^2 - y^2 - 1 = 0$. Faire une figure.

g. Deux coniques peuvent-elles avoir trois points d'intersection ? Si la réponse est positive, donner un exemple. Si la réponse est négative, justifier cette réponse.

Sujet d'étude 2 : les tangentes des courbes algébriques

Dans cette étude, nous revenons à des courbes algébriques planes quelconques, définies par une équation polynomiale $P(x, y) = 0$. On a vu au paragraphe 13.2.b la nécessité de savoir calculer l'équation d'une tangente à une courbe algébrique : c'est le sujet de cette étude.

*

Soient C une telle courbe et $M_0 = (x_0, y_0)$ un point de C d'équation *minimale* $P(x, y) = 0$ (cf. sujet d'étude 1 pour une définition). Soit $D_{\alpha,\beta}$ une droite passant par M_0 de vecteur directeur (α, β). On peut montrer que, pour tout $t \in \mathbf{R}$,

$$\begin{cases} x(t) &= x_0 + \alpha t \\ y(t) &= y_0 + \beta t \end{cases}$$

est une paramétrisation de la droite $D_{\alpha,\beta}$.

Exercice 13.2.34 *Démontrer ce résultat.*

Quitte à translater le repère (et à considérer plutôt l'équation $P(X + x_0, Y + y_0) = 0$), on peut supposer que M_0 est égal à l'origine du repère, donc que $x_0 = y_0 = 0$.

Définition 13.2.35 *On appelle degré du monôme $x^i y^j$ l'entier $i + j$.*

Définition 13.2.36 *On dit qu'une fonction polynomiale $P(x, y) = \sum_{0 \leq i \leq m, 0 \leq j \leq n} a_{i,j} x^i y^j$ est* homogène *de degré d, si tous les degrés des monômes qui la composent sont égaux à d.*

Exercice 13.2.37 *Montrer que si P est homogène de degré d alors, pour tout $\lambda \in \mathbf{R}$,*

$$P(\lambda x, \lambda y) = \lambda^d P(x, y)$$

En déduire que si le point $M = (x, y)$ appartient à la courbe définie par P, alors tous les points de la droite passant par M et l'origine y appartiennent aussi.

Il est clair que l'on peut alors écrire

$$P(x, y) = \sum_{\ell=r}^{d} P_\ell(x, y)$$

($r \geq 1$ et $P_r(x, y) \neq 0$) en groupant les monômes de même degré, de sorte que, pour tout $1 \leq \ell \leq r$, P_ℓ est homogène de degré ℓ. Si l'on calcule l'intersection de $D_{\alpha,\beta}$ avec C on est ramené à résoudre l'équation

$$P\left(x(t), y(t)\right) = t^r \left(\sum_{\ell=r}^{d} t^{\ell - r} P_\ell(\alpha, \beta) \right) = 0$$

puisque les P_ℓ sont homogènes. Si $P_r(\alpha, \beta) = 0$, on dit que $D_{\alpha,\beta}$ est *tangente* à C en M_0 et l'ensemble des tangentes à la courbe en M_0 est donné par l'équation

$$P_r(x, y) = 0$$

Exemple 13.2.38 *La conique d'équation*

$$x^2 + y^2 - 1 = 0$$

possède en $(1,0)$ *une unique tangente. En effet, on commence par changer de repère en considérant l'équation*

$$\begin{cases} (X+1)^2 + Y^2 - 1 &= 0 \\ \left(X^2 + Y^2\right) + 2X &= 0 \end{cases}$$

La tangente en $(1,0)$ *a pour équation* $X = 0$ *dans le nouveau repère et donc* $x = 1$ *dans l'ancien.*

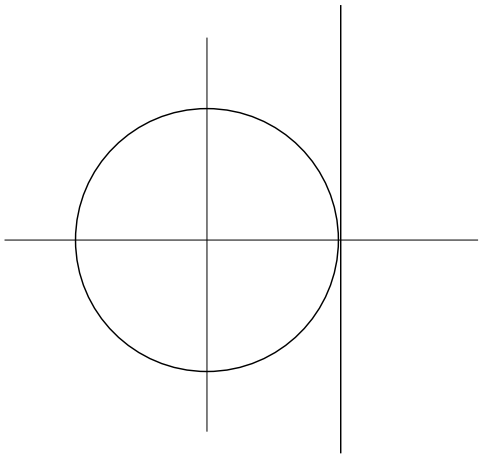


FIG. 13.19 – La tangente au cercle unité au $(1,0)$

Remarque 13.2.39 *Il découle de l'expression de* $P(x(t), y(t))$ *que*

$$P_r(\alpha, \beta) = \lim_{t \to 0, t \neq 0} \frac{P(\alpha t, \beta t)}{t^r}$$

Exercice 13.2.40 *Montrer que la cubique d'équation*

$$y^2 = x^3 + 17$$

possède en $(2,5)$ *une unique tangente d'équation* $5y - 6x + 20 = 0$.

Définition 13.2.41 *On dit qu'un point* M_0 *est non singulier si* $r = 1$.

Remarque 13.2.42 *La courbe* C *et la courbe* C_n *d'équation* $P^n(x,y) = 0$ *sont, du point de vue de notre définition, les mêmes objets. Par contre, si* $n \geq 2$, C_n *ne peut avoir de point non singulier. C'est pour cela que dans cette étude, on ne travaille qu'avec une équation minimale de* C.

Exercice 13.2.43 *Soit* $x \mapsto P(x)$ *une fonction polynomiale telle que P(0)=0. Son graphe a pour équation*

$$y - P(x) = 0$$

a. En utilisant la dérivabilité de la fonction, calculer une équation de la tangente en $(0,0)$.

b. En utilisant la description ci-dessus, montrer que la courbe d'équation $y - P(x) = 0$ possède une unique tangente dont on précisera une équation en posant $P(x) = \sum_{i\geq 1} a_i x^i$.

c. Vérifier que ces deux définitions coïncident.

Il faut bien noter que, contrairement au cas des graphes des fonctions polynomiales (à une variable), les courbes algébriques peuvent avoir plusieurs tangentes.

Exemple 13.2.44 *La cubique d'équation*

$$y^2 - x^3 - x^2 = (y - x)(y + x) - x^3 = 0$$

possède deux tangentes en $(0, 0)$ qui sont les deux bissectrices des axes.

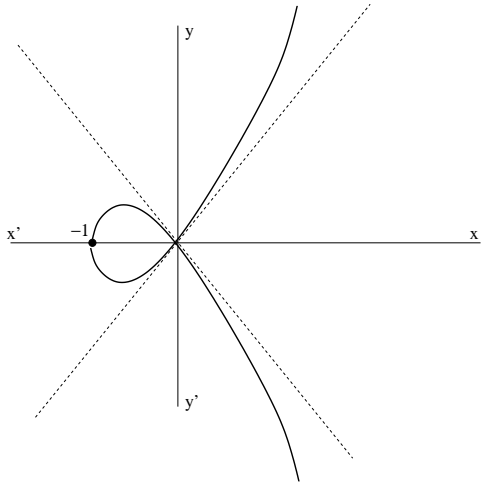


FIG. 13.20 – La cubique avec un point double d'équation $y^2 - x^3 - x^2 = 0$

• Plaçons-nous dans le cas où $r = 1$. Dans ce cas, l'intersection de la droite $D_{\alpha,\beta}$ et de la courbe C donne une équation de la forme

$$P(\alpha t, \beta t) = tP_1(\alpha, \beta) + t^2 \left(\sum_{\ell=2}^{d} t^{\ell-2} P_\ell(\alpha, \beta) \right)$$

Exemple 13.2.45 *C'est le cas par exemple d'une courbe d'équation de la forme $y - P(x) = 0$ (cf. exercice 13.2.43).*

On remarque également que, dans ce cas, C ne possède qu'une seule tangente. En outre, la condition $P_1(\alpha, \beta) = 0$ s'écrit encore

$$\alpha P_1(1, 0) + \beta P_1(0, 1) = 0$$

Remarque 13.2.46 *Cette dernière relation s'interprète géométriquement comme l'orthogonalité des vecteurs de coordonnées (α, β) et $(P_1(1, 0), P_1(0, 1))$. Ce dernier vecteur est appelé gradient de P en $(0, 0)$.*

Autrement dit, le vecteur (α, β) est le vecteur directeur de la tangente à C en l'origine si et seulement s'il est orthogonal au gradient de P en $(0,0)$. Il existe une autre manière, peut-être plus directe, de calculer le gradient de P, sans passer par une décomposition en composantes homogènes. En effet, on voit facilement que les réels $P_1(1,0)$ et $P_1(0,1)$ s'obtiennent aussi par un calcul de limites. Plus précisément, on a les formules

$$P_1(1,0) = \lim_{t\to 0, t\neq 0} \frac{P(t,0)-P(0,0)}{t} \quad \text{et} \quad P_1(0,1) = \lim_{t\to 0, t\neq 0} \frac{P(0,t)-P(0,0)}{t}$$

Exemple 13.2.47 *Considérons le cas du graphe d'une fonction polynomiale d'équation $y - P(x) = 0$ telle que $P(0) = 0$. De l'une ou l'autre des deux manières précédentes, on peut calculer le gradient de cette courbe C. Le calcul donne*

$$P_1(0,1) = 1 \quad et \quad P_1(1,0) = -P'(0)$$

L'équation de la tangente en $(0,0)$ est donnée par l'équation $y - P'(0)x = 0$.

On adoptera les notations suivantes :

$$\frac{\partial P}{\partial x}(0,0) := P_1(1,0) = \lim_{t\to 0, t\neq 0} \frac{P(t,0)-P(0,0)}{t}$$

et

$$\frac{\partial P}{\partial y}(0,0) := P_1(0,1) = \lim_{t\to 0, t\neq 0} \frac{P(0,t)-P(0,0)}{t}$$

Bien entendu, les définitions précédentes sont énoncées dans le cas où l'on a translaté M_0 en l'origine du repère. Mais, quitte à appliquer la translation inverse, on peut démontrer les formules suivantes :

∘ Soient C une courbe algébrique, définie par l'équation $P(x,y) = 0$ et $M_0 = (x_0, y_0) \in C$ un point non singulier. On appelle *dérivées partielles* de P en M_0 les réels

$$\frac{\partial P}{\partial x}(x_0,y_0) := \lim_{t\to 0, t\neq 0} \frac{P(x_0+t,y_0)-P(x_0,y_0)}{t}$$

et

$$\frac{\partial P}{\partial y}(x_0,y_0) := \lim_{t\to 0, t\neq 0} \frac{P(x_0,y_0+t)-P(x_0,y_0)}{t}$$

∘ La tangente en M_0 est donnée par l'équation

$$\frac{\partial P}{\partial x}(x_0,y_0)\,(x-x_0) + \frac{\partial P}{\partial y}(x_0,y_0)\,(y-y_0) = 0$$

Remarque 13.2.48

a. *L'appellation "dérivée partielle" n'est pas sans rapport avec l'opération de dérivation usuelle que l'on connaît pour les fonctions de la variable réelle. Moralement, avoir une dérivée partielle signifie être dérivable au sens usuel par rapport au vecteur de base correspondant (cf. exercice 13.3.31 pour une explication plus précise) ;*

b. *il faut bien noter que le x au dénominateur de $\dfrac{\partial P}{\partial x}$ est simplement une notation qui indique que l'on dérive par rapport au vecteur e_1 ;*

c. les dérivées partielles se calculent comme les dérivées usuelles, en considérant l'autre variable comme une constante.

Exercice 13.2.49 *Soit C la courbe d'équation $x^5 + yx - 1 = 0$.*

a. Montrer que le point $(1, 0)$ est non singulier.

b. Calculer une équation de la tangente de C en $(1, 0)$ en calculant une décomposition homogène de C.

c. Calculer une équation de la tangente de C en $(1, 0)$ en utilisant les dérivées partielles.

13.3 . . . à une approche dynamique des courbes planes

• Au paragraphe précédent on a vu apparaître la notion de paramétrisation (rationnelle) et avons pu apprécier son efficacité pour résoudre la question de la "description" des points rationnels des coniques, des droites rationnelles, ou de certaines cubiques. En outre, avec l'exemple de la cubique d'équation $y^2 - x^3 - x = 0$, on a également vu apparaître des limites qui ne sont pas à imputer à cette technique, mais au fait de ne considérer que des paramétrisations rationnelles (en effet, la cubique précédente possède une représentation paramétrique dite de *Weierstrass* - mathématicien allemand, 1815-1897 - de la forme $(\mathfrak{P}, \mathfrak{P}')$ où $\mathfrak{P}$ est une fonction qui n'est pas rationnelle).

• D'un point de vue cinématique (cf. l'exemple 13.1.2), si l'on veut considérer notre courbe comme la trajectoire d'un mobile qui se déplace sur une surface plane, il est important de connaître sa position, qui est modélisée par l'abscisse et l'ordonnée des points de la courbe, à chaque instant t donné. Autrement dit, on veut être capable de décrire l'abscisse x et l'ordonnée y d'un point M de la courbe comme des *fonctions* du temps $t \mapsto x := x(t)$ et $t \mapsto y := y(t)$. Là encore, comme tous les phénomènes physiques que l'on côtoie ne se modélisent pas avec des fractions rationnelles, il est nécessaire de considérer des fonctions $t \mapsto x(t)$ et $t \mapsto y(t)$ plus générales.

• Depuis le lycée, on sait représenter les graphes de fonctions de la variable réelle d'équation $y - f(x) = 0$, avec f une fonction dérivable. Pourquoi alors se limiter au cas des fonctions polynomiales ?

• Enfin, même dans le cas où une courbe est paramétrisable nous n'avons toujours pas donné de technique pour représenter graphiquement de tels sous-ensembles de $\mathbf{R}^2$.

Ce paragraphe va tâcher de répondre à cette dernière question. En outre, les remarques précédentes justifient que l'on adopte une nouvelle approche des courbes planes, en parlant de *courbes paramétrées*.

13.3.a La notion de courbe paramétrée : définitions et premiers exemples

• On a donné au paragraphe précédent une paramétrisation *rationnelle* du cercle unité, que l'on rappelle

$$\begin{cases} x(t) &= \dfrac{1-t^2}{1+t^2} \\ \\ y(t) &= \dfrac{2t}{1+t^2} \end{cases}$$

avec $t \in \mathbf{R}$ et en remarquant que le point de coordonnées $(-1, 0)$ peut être atteint comme limite lorsque la valeur du paramètre t tend vers l'infini. Toutefois, avec la trigonométrie du triangle rectangle, on peut définir une autre paramétrisation du cercle unité, qui n'est plus définie par des fonctions $x(t)$ et $y(t)$ rationnelles, mais par des fonctions trigonométriques (que l'on sait étudier)

$$\begin{cases} x(t) &= \cos(t) \\ \\ y(t) &= \sin(t) \end{cases}$$

Les relations trigonométriques assurent que $x(t)^2 + y(t)^2 - 1 = 0$, si l'on veut retrouver l'équation cartésienne du cercle (où le paramètre t n'apparaît plus qu'implicitement). Bien évidemment cette deuxième paramétrisation est moins utile dans l'étude des points rationnels.

• Soit $f : I \subset \mathbf{R} \to \mathbf{R}$ une fonction dérivable sur l'intervalle I. On sait représenter son graphe, *i.e.* le sous-ensemble de $\mathbf{R}^2$ défini par l'équation $y - f(x) = 0$. Pourtant, le cas où f n'est pas une fonction polynomiale n'a pas été pris en compte au paragraphe précédent. En outre, on peut encore décrire cet ensemble comme l'ensemble des points du plan dont les coordonnées x et y sont de la forme

$$\begin{cases} x := x(t) &= t \\ y := y(t) &= f(t) \end{cases}$$

pour un réel t appartenant à l'intervalle I.

Au regard de ces deux exemples, on est tenté de donner une nouvelle définition de la notion de courbe plane, peut-être plus générale et plus adaptée à la modélisation des phénomènes physiques.

On dira qu'une application $\varphi : I \subset \mathbf{R} \to \mathbf{R}^2$ définie par $t \mapsto (x(t), y(t))$ est une *paramétrisation* de l'ensemble $C \subset R^2$, si $\varphi(I) = C$ et si les fonctions $x : I \subset \mathbf{R} \to \mathbf{R}$ et $y : I \subset \mathbf{R} \to \mathbf{R}$ sont dérivables sur I. On dit alors qu'un tel sous-ensemble de $\mathbf{R}^2$ est une *courbe paramétrée* du plan.

Remarque 13.3.1 *Cette définition se justifie du point de vue de la cinématique. En effet, se donner une courbe paramétrée c'est se donner, tout d'abord, la position du mobile en fonction du temps, à chaque instant t. En outre, si l'on arrive à représenter graphiquement de tels objets, on aura la trajectoire du mobile. Enfin, l'hypothèse de dérivabilité se traduit, via la notion de vitesse, par la façon dont le mobile va décrire cette trajectoire.*

On notera $\varphi(t)$ le vecteur de coordonnées $(x(t), y(t))$ et $\varphi'(t)$ celui de coordonnées $(x'(t), y'(t))$, les fonctions x' et y' étant simplement les dérivées des fonctions x et de y.

• La notion de courbes paramétrées permet de décrire des situations plus générale que celles décrites par les courbes algébriques. Voici quelques exemples pour s'en convaincre.

Exemple 13.3.2 *Considérons la paramétrisation $\varphi : I \subset \mathbf{R} \to \mathbf{R}^2$ définie par*

$$\begin{cases} x(t) &= t + 1 \\ y(t) &= t - 2 \end{cases}$$

Si $I = \mathbf{R}$, on sait que l'ensemble $\varphi(I)$ est la droite D passant par le point de coordonnées $(1, -2)$ et parallèle à la première bissectrice des axes. Si $I = [0, 1]$, il est clair que $\varphi(I) \subset D$, mais cette inclusion n'est plus une égalité. L'ensemble $\varphi(I)$ est simplement un segment de la droite D. On peut éliminer t des équations et vérifier que les points de la courbe appartiennent bien à l'ensemble défini par l'équation $y - x + 3 = 0$.

Dans le même ordre d'idée, on peut considérer la paramétrisation $\psi : \mathbf{R}_+^* \to \mathbf{R}^2$ définie par

$$\begin{cases} x(t) &= t \\ y(t) &= \dfrac{1}{t} \end{cases}$$

En éliminant t, on trouve $xy - 1 = 0$. Notre courbe paramétrée est donc contenue dans la courbe algébrique d'équation $xy - 1 = 0$. Précisément, l'ensemble $\psi(\mathbf{R}_+^*)$ peut se décrire par le système :

$$\begin{cases} xy - 1 = 0 \\ \\ x > 0 \end{cases}$$

Autrement dit, on ne décrit qu'une seule branche de l'hyperbole. Cette dernière n'est pas une courbe algébrique.

Exercice 13.3.3 *Soit $\varphi : \mathbf{R}_+ \to \mathbf{R}^2$ la paramétrisation définie par*

$$\begin{cases} x(t) &= & t^2 + 1 \\ \\ y(t) &= & t^2 + t + 1 \end{cases}$$

En éliminant t, montrer que la courbe paramétrée par φ est un morceau d'une conique, dont on précisera l'équation. Faire une figure.

Exemple 13.3.4 *Soit $\varphi : [0, \pi/4] \to \mathbf{R}^2$ la paramétrisation définie par*

$$\begin{cases} x(t) &= & \cos(t) \\ \\ y(t) &= & t\cos(2t) \end{cases}$$

On peut encore éliminer t dans les équations, mais on trouve la relation

$$y - (2x^2 - 1)\arccos(x) = 0$$

Remarque 13.3.5 *Il n'est pas toujours facile (ou possible) de trouver une équation cartésienne d'une courbe paramétrée, en éliminant le paramètre dans les coordonnées. S'en convaincre avec l'exemple suivant*

$$\begin{cases} x(t) &= & e^t + \log(t) \\ \\ y(t) &= & t + \sin(t) \end{cases}$$

Exercice 13.3.6 (La cycloïde) *Soit $\varphi : \mathbf{R} \to \mathbf{R}^2$ la paramétrisation définie par*

$$\begin{cases} x(t) &= & t - \sin(t) \\ \\ y(t) &= & 1 - \cos(t) \end{cases}$$

Montrer que l'ensemble $\varphi(\mathbf{R}) \cap D$, où D est la droite d'équation $y = 0$, est infini. Déduire du théorème de Bézout (cf. sujet d'étude 1) que la courbe paramétrée par φ ne peut être algébrique.

13.3.b Le tracé des courbes paramétrées

Dans ce paragraphe, on note C une courbe paramétrée par $\varphi : I \subset \mathbf{R} \to \mathbf{R}^2$ définie par $t \mapsto (x(t), y(t))$ (I est un intervalle de $\mathbf{R}$). Notons que l'image de $t \in \mathbf{R}$ est un vecteur de $\mathbf{R}^2$. En particulier, on note

$$\|\varphi(t)\| := \sqrt{(x(t))^2 + (y(t))^2}$$

la longueur (on dit aussi la *norme*) de ce vecteur.

L'étude précédant la représentation graphique des courbes paramétrées ressemble à celle des fonctions de la variable réelle. Il faut quand même noter que, contrairement au cas des fonctions où seul y varie en fonction de x, les deux coordonnées varient simultanément. Nous allons maintenant donner un "protocole" pour mener cette étude.

L'étude des symétries

L'étude des symétries et de la périodicité permet de ramener l'étude à un intervalle plus petit que l'intervalle I donné au départ. Il y a beaucoup de façons de réduire l'intervalle d'étude (chercher une *isométrie* du plan laissant C invariante). La technique est d'essayer de changer de paramètre. Voici quelques premiers exemples :

• L'étude de la périodicité peut permettre une telle réduction.

Exemple 13.3.7 *Soit* $\varphi : t \in \mathbf{R} \mapsto (\sin(t)+1, \sin(t)+1)$. *On remarque facilement que*

$$\varphi(t) = \varphi(t+2\pi) \ ,$$

mais aussi que $\varphi(t) = \varphi(\pi - t)$. *Autrement dit, dans ce cas, on peut ramener l'étude à* $[-\pi, \pi]$, *puis à* $[-\pi/2, \pi/2]$, *par exemple. Cela signifie que la courbe toute entière est obtenue pour* t *variant dans cet intervalle. Remarquons enfin qu'elle forme un "morceau" de la droite d'équation* $y = x-2$ *(mais vérifie la condition supplémentaire* $0 \leq x \leq 2$*).*

• Passons à l'étude des symétries classiques.

Exemple 13.3.8 *Soit* $\varphi : t \in \mathbf{R} \mapsto \left(t^2+1, t^2-1\right)$. *On remarque facilement que*

$$\varphi(t) = \varphi(-t) \ .$$

Techniquement, il nous suffit donc d'étudier φ *sur* $\mathbf{R}_+$ *pour être capable de représenter la courbe. Cinématiquement, cela signifie que le mobile va dans un sens sur cette courbe jusqu'à l'instant* $t = 0$ *puis qu'il revient en arrière, repassant par les mêmes points. Il est par ailleurs, facile de voir que cette courbe est contenue dans la droite d'équation* $y = x - 2$ *(et vérifie la condition supplémentaire* $x > 0$*).*

Exemple 13.3.9 *Soit* $\varphi : t \in \mathbf{R}_+^* \mapsto \left(t + \dfrac{1}{t}, t^2 + \dfrac{1}{t^2}\right)$. *La transformation* $t \mapsto 1/t$ *laisse* φ *invariant. On peut donc ramener l'étude à l'intervalle* $]0, 1]$. *On peut également remarquer que cette courbe est une partie de la parabole d'équation* $y - x^2 + 2 = 0$. *La déterminer.*

Exemple 13.3.10 *Considérons la courbe paramétrée par* $\varphi : t \in \mathbf{R} \mapsto \left(\cos^2(t), 2\sin(t)\cos(t)\right)$. *Il est facile de voir que*

$$\varphi(t) = \varphi(t+\pi)$$

de sorte que l'on peut ramener l'étude à l'intervalle $]-\pi/2, \pi/2[$. *On peut remarquer aussi que*

$$\left\{ \begin{array}{rcl} x(t) & = & x(-t) \\ \\ y(t) & = & -y(t) \end{array} \right. .$$

Géométriquement, cela s'interprète en remarquant que les points de coordonnées $\varphi(t)$ *et* $\varphi(-t)$ *sont symétriques par rapport à l'axe des abscisses.*

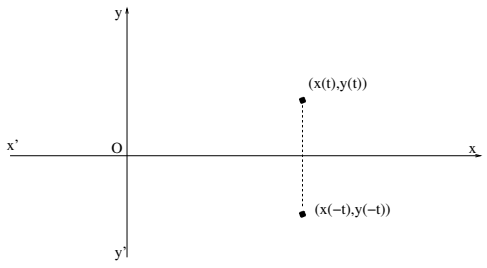


FIG. 13.21 – Symétrie par rapport à $y = 0$

Exemple 13.3.11 *Soit* $\varphi : t \in \mathbf{R}_+^* \to (3\cos(t) - \cos(3t), 3\sin(t) - \sin(3t))$. *Il est clair que l'on peut se ramener à l'étude de* φ *sur l'intervalle* $[0, 2\pi]$. *Par ailleurs, on remarque que*

$$\varphi(t) = -\varphi(t + \pi) \quad \textit{et que} \quad \varphi(-t) = -\varphi(t)$$

Les points de coordonnées $\varphi(t)$ *et* $\varphi(-t)$ *sont donc symétriques par rapport à l'origine* O. *De même, les points* $\varphi(t)$ *et* $\varphi(t + \pi)$ *sont symétriques par rapport à l'origine.*

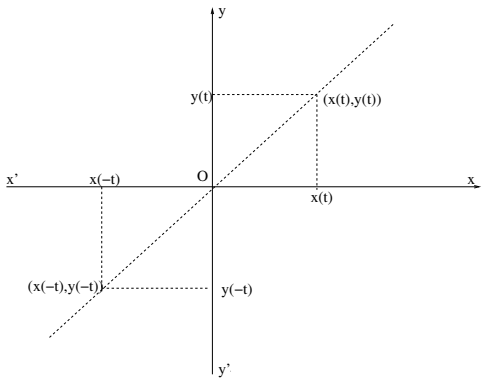


FIG. 13.22 – Symétrie par rapport à O

On peut donc ramener l'étude de φ *à l'intervalle* $[0, \pi/2]$.

• Résumons les quelques exemples recensés. Bien entendu cette liste n'est pas exhaustive. Soit $\psi : I \to I$ un changement de paramétrage (cf. ci-dessus) et I' un sous-intervalle de I tel que $I = I' \cup \psi(I')$ et $(I' \cap \psi(I') = \emptyset$ ou $I' \cap \psi(I') = \{t_0\})$. Par exemple,

$\psi : t \in]-a, a[\to -t \in]-a, a[$ et $I' = [0, a[$ (a pouvant éventuellement être infini)

$\psi : t \in \mathbf{R}_+^* \to 1/t \in \mathbf{R}_+^*$ et $I' =]0, 1]$

Hypothèse sur ψ	Isométrie correspondante
$\begin{cases} x(\psi(t)) = x(t) \\ y(\psi(t)) = y(t) \end{cases}$	Identité
$\begin{cases} x(\psi(t)) = x(t)+a \\ y(\psi(t)) = y(t)+b \end{cases}$	Translation de vecteur (a,b)
$\begin{cases} x(\psi(t)) = -x(t) \\ y(\psi(t)) = y(t) \end{cases}$	Symétrie orthogonale par rapport à l'axe $x=0$
$\begin{cases} x(\psi(t)) = x(t) \\ y(\psi(t)) = -y(t) \end{cases}$	Symétrie orthogonale par rapport à l'axe $y=0$
$\begin{cases} x(\psi(t)) = x(t) \\ y(\psi(t)) = -y(t) \end{cases}$	Symétrie orthogonale par rapport à O

Exercice 13.3.12 *Soit C la conique d'équation $x^2+y^2-1=0$. Montrer qu'il existe un paramétrage $\varphi : [0,\pi/4] \to \mathbf{R}^2$ de C. Préciser les symétries effectuées.*

L'étude des branches infinies

Les courbes paramétrées, algébriques... ont des allures bien différentes. Par exemple, une ellipse C est *bornée*, *i.e* la distance de tout point de l'ellipse à l'origine O du repère est finie, bornée par une même quantité k, ce qui se traduit en écrivant

$$\exists k \in \mathbf{R}, \quad \forall M \in C, \quad \|\vec{OM}\| \leq k$$

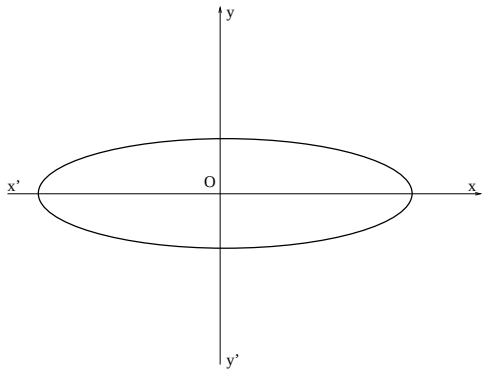


FIG. 13.23 – Une ellipse

Alors que cette propriété n'est pas vérifiée par la folium de Descartes (cf. exercice 13.2.19).

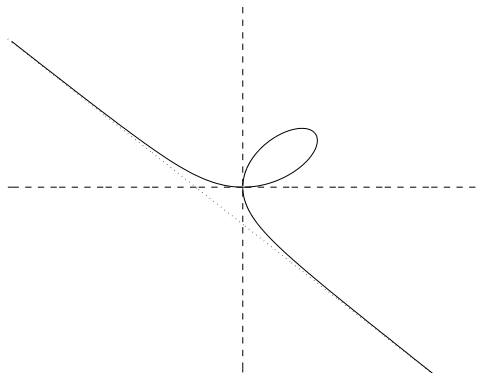


FIG. 13.24 – La folium de Descartes

Exercice 13.3.13 *Trouver $k \in \mathbf{R}_+^*$ tel que, si $M = (x, y)$ appartient à l'ellipse d'équation*

$$\frac{x^2}{2} + \frac{y^2}{3} - 1 = 0 \ ,$$

alors $\|O\vec{M}\| \leq k$.

On dit que la courbe paramétrée C possède une *branche infinie* en $t_0 \in \overline{I}$ si le réel $\|\varphi(t)\|$ tend vers $+\infty$ quand le paramètre t se rapproche de 0, ce qui se note $\lim_{t \to t_0} \|\varphi(t)\| = +\infty$. Avec la définition de la norme euclidienne, il est facile de se convaincre que cette définition équivaut au fait que, ou bien $|x(t)| \to +\infty$ ou bien $|y(t)| \to +\infty$.

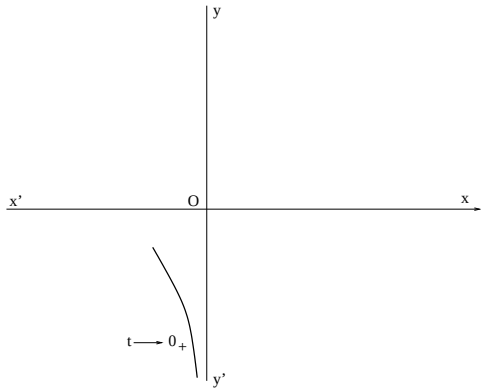
Plusieurs cas de figure peuvent se rencontrer.

- Si une seule des deux coordonnées tend vers l'infini, *i.e.*

$$\begin{cases} x(t) \to a \\ \\ y(t) \to \pm\infty \end{cases} \quad \text{ou} \quad \begin{cases} x(t) \to \pm\infty \\ \\ y(t) \to b \end{cases}$$

alors la courbe a une *asymptote* verticale ou horizontale respectivement.

Exemple 13.3.14 *Considérons* $\varphi : t \in \mathbf{R}_+^* \mapsto (t\log(t), t + \log(t))$. *Un calcul de limite montre que* $x(t) \to 0$ *et* $y(t) \to -\infty$. *Donc* C *possède une asymptote d'équation* $x = 0$.



- Si les deux coordonnées tendent vers l'infini, *i.e.*

$$\begin{cases} x(t) & \to & \pm\infty \\ y(t) & \to & \pm\infty \end{cases}$$

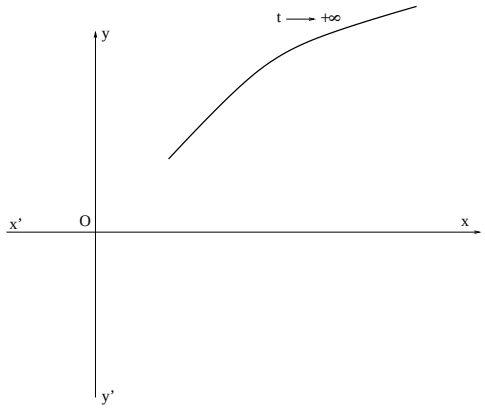
il faut étudier le quotient $\dfrac{y(t)}{x(t)}$. On a alors différents cas de figure.

  - Si $\dfrac{y(t)}{x(t)} \to \pm\infty$, on dit que C possède une *branche parabolique* de direction asymptotique l'axe des ordonnées.
  - Si $\dfrac{y(t)}{x(t)} \to 0$, on dit que C possède une *branche parabolique* asymptotique l'axe des abscisses.

    Exemple 13.3.15 *Reprenons l'exemple de* $\varphi : t \in \mathbf{R}_+^* \mapsto (t\log(t), t + \log(t))$. *Si l'on calcule la limite pour* t *tendant vers* $+\infty$, *on obtient* $x(t),\ y(t) \to +\infty$. *Par ailleurs, si* $t \neq 0$

$$\frac{y(t)}{x(t)} = \frac{1 + \frac{\log(t)}{t}}{\log(t)} \to 0$$

    La courbe admet donc une direction parabolique dans la direction de (Ox).



- Si $\dfrac{y(t)}{x(t)} \to a \in \mathbf{R}^*$ et si $y(t) - ax(t) \to \pm\infty$ on dit que C possède une *branche parabolique* de direction asymptotique la droite d'équation $y - ax = 0$.
- Si $\dfrac{y(t)}{x(t)} \to a \in \mathbf{R}^*$ et si $y(t) - ax(t) \to b \in \mathbf{R}$ on dit que C possède pour asymptote la droite d'équation $y - ax - b = 0$.

Exercice 13.3.16 *a. Étudier la branche infinie en $+\infty$ de la courbe C paramétrée par $\varphi : \mathbf{R} \to \mathbf{R}^2$*

$$t \mapsto \begin{cases} x(t) &= \dfrac{t^3+1}{t^2+1} \\ \\ y(t) &= \dfrac{t^4+1}{t^2+1} \end{cases}$$

b. Étudier la branche infinie en $+\infty$ de la courbe C paramétrée par $\varphi : \mathbf{R} \to \mathbf{R}^2$

$$t \mapsto \begin{cases} x(t) &= 4t^3 + 3t^2 - 6t \\ \\ y(t) &= 3t^2 + 2t + 1 \end{cases}$$

*

On fait l'hypothèse désormais que le vecteur $\varphi'(t) := (x'(t), y'(t))$ n'est jamais nul. Cette hypothèse nous restreint à l'étude de courbes paramétrées qui ne contiennent pas de points *singuliers*.

Remarque 13.3.17 *D'un point de vue dynamique, cette hypothèse signifie que l'on étudie uniquement les trajectoires de mobiles qui ne s'immobilisent pas (dont la vitesse ne s'annule pas) au cours de leur mouvement. Un point singulier d'une courbe correspond au contraire à l'immobilisation du mobile à un instant t_0 donné.*

L'étude des tangentes

Pour tracer la représentation graphique d'une courbe, avoir des points ne suffit pas en général. Il faut aussi savoir comment on arrive à ces points (ceci est déjà vrai pour les graphes de fonctions). Nous allons donc étudier la notion de *tangente* d'une courbe paramétrée.

Soient A un point de C de coordonnées $\varphi(t_0)$ et $\vec{v} = (v_1, v_2)$ un vecteur, limite d'une suite de vecteurs $\vec{v}^{(k)} = \left(v_1^{(k)}, v_2^{(k)}\right)$ de la forme $\lambda_k A\vec{M}_k$ (la suite $(M_k)_{k\in\mathbf{N}}$ est une suite de points de C qui tend vers A et $\lambda_k \in \mathbf{R}$). On rappelle que la suite $(\vec{v}^{(k)})_{k\in\mathbf{N}}$ converge vers v si la suite de terme général

$$\|\vec{v} - \vec{v}^{(k)}\|$$

tend vers 0 (quand k tend vers l'infini).

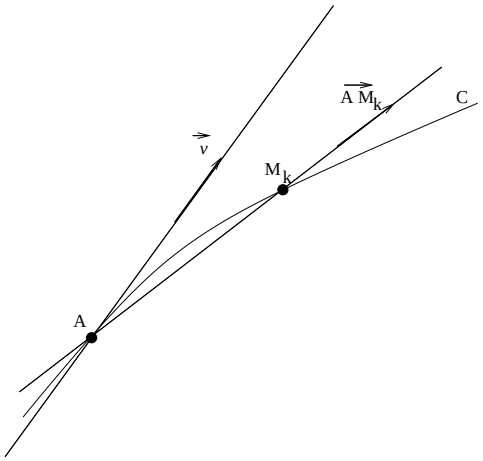


FIG. 13.25 – Le vecteur $\vec{v}$

On dit que la droite passant par A de vecteur directeur $\vec{v}$ est la *tangente de C en A*. Géométriquement, cela définit la tangente comme une "limite" de droites sécantes passant par A. Une description "paramétrique" donne encore

$$\vec{v} = \lim_{k \to +\infty} \lambda_k \left(\varphi(t_0 + h_k) - \varphi(t_0)\right)$$

avec $\lim h_k = 0$. Une manière d'exprimer qu'une fonction $f : J \subset \mathbf{R} \to \mathbf{R}$ est dérivable en $x_0 \in J$ est de dire que, au voisinage de x_0, on a l'égalité

$$f(x_0 + h) = f(x_0) + hf'(x_0) + h\varepsilon(h)$$

où ε est une fonction qui tend vers 0 quand h tend vers 0 (voir Chap. 8).

Grâce à ce point de vue, on pourrait démontrer qu'en fait $\vec{v}$ est colinéaire à $\varphi'(t_0)$. La proposition 13.3.18 ci-dessus donne un moyen facile de tracer les tangentes en A.

Proposition 13.3.18 *Soit C une courbe paramétré par $\varphi : I \subset \mathbf{R} \to \mathbf{R}^2$ telle que $\varphi'(t) \neq 0$ pour tout $t \in I$. Alors, en tout point $A = \varphi(t_0)$, C possède une tangente de coefficient directeur $\varphi'(t_0)$. Précisément,*

a. si $x'(t_0) = 0$, C possède une tangente parallèle à l'axe des ordonnées ;

b. si $y'(t_0) = 0$, C possède une tangente parallèle à l'axe des abscisses ;

c. si $x'(t_0) \neq 0$, l'équation de la tangente de C en A a pour vecteur directeur $\left(1, \dfrac{y'(t_0)}{x'(t_0)}\right)$.

Exercice 13.3.19 *En reprenant les notations ci-dessus, montrer que $\vec{v}$ est colinéaire à $\varphi'(t_0)$. On pourra démontrer que la suite $(\lambda_k h_k)_{k \in \mathbf{N}}$ converge.*

Remarque 13.3.20 *Du point de vue dynamique, il peut arriver qu'un mobile repasse par un même point. On dit dans ce cas que sa trajectoire possède un point multiple (double s'il passe deux fois...). Dans ce cas, la courbe paramétrée peut avoir plusieurs tangentes bien que le point soit régulier. Il est important de noter que dans la définition et la proposition ci-dessus la notion de tangente dépend de t_0 plus que du point A à proprement parler. En particulier, s'il existe $t_0 \neq t_1$ tels que $\varphi(t_0) = \varphi(t_1) = A$, il n'y a rien de choquant à obtenir, avec nos définitions, deux tangentes différentes.*

Exercice 13.3.21 *Montrer que la courbe C paramétrée par*

$$\begin{cases} x(t) &= t+1+\frac{1}{t-1} \\ y(t) &= t^2+1+\frac{1}{t} \end{cases}$$

possède un unique point double que l'on déterminera.

Exercice 13.3.22 *Soit C la courbe algébrique d'équation cartésienne*

$$4x^2 - y^2 - 4 = 0$$

a. *En utilisant la définition donnée au sujet d'étude 2, montrer que C possède une unique tangente dont vous donnerez une équation cartésienne.*

b. *Montrer que C est la réunion de C_1 la courbe paramétrée par $\varphi_1 : t \in \mathbf{R}_+^* \mapsto \left(t+\frac{1}{t}, 2(t+\frac{1}{t})\right) \in \mathbf{R}^2$ et de C_2 paramétrée par $\varphi_2 : t \in \mathbf{R}_-^* \mapsto \left(t+\frac{1}{t}, 2(t-\frac{1}{t})\right) \in \mathbf{R}^2$.*

c. *Retrouver le résultat de la question 1 en utilisant la proposition 13.3.18*

Remarque 13.3.23 *En général, quand un mobile se déplace il peut lui arriver de s'immobiliser au cours de son déplacement. De telles trajectoires ne sont pas étudiées ici de manière générale. La différence essentielle avec le cas régulier est la manière d'obtenir les tangentes. Pour avoir une théorie générale satisfaisante il est nécessaire de connaître les formules de Taylor aux ordres supérieurs, que nous n'allons pas traiter ici.*

Exercice 13.3.24 (étude d'un point de rebroussement de première espèce)

a. *Soit C la cubique d'équation $y^2 - x^3 = 0$. En utilisant les définitions du sujet d'étude 2, étudier le problème des tangentes au point $(0,0)$. En remarquant que tout nombre réel positif est un carré (dans $\mathbf{R}$), déterminer un paramétrage de C (penser à écrire, en le justifiant, que $x = t^2$) ; étudier les symétries de la courbe.*

b. *Trouver un paramétrage de $C\backslash\{0\}$ de la forme $x(t) = t$, $y(t) = f(t)$. Représenter C graphiquement.*

c. *Soit Γ la courbe paramétrée par $\varphi : \mathbf{R} \to \mathbf{R}^2$ définie par*

$$\begin{cases} x(t) &= \frac{1}{3}\left(2\cos(t)+\cos(2t)\right) \\ y(t) &= \frac{1}{3}\left(2\sin(t)-\sin(2t)\right) \end{cases}$$

Montrer qu'une équation cartésienne de Γ est donnée par

$$3(x^2+y^2)^2 + 8x(3y^2-x^2) + 6(x^2+y^2) - 1 = 0$$

d. *Vérifier que $(1,0)$ est point rationnel de Γ. Le point $(1,0)$ est-il régulier ? étudier la tangente de C en $(1,0)$.*

Les variations et le tracé de la courbe

La dernière étape avant la représentation graphique de C est l'étude de ses variations. Depuis le lycée, on sait utiliser la dérivée pour étudier les variations d'une fonction. Dans le cas des courbes paramétrées, il faut noter que l'abscisse *et* l'ordonnée sont deux fonctions qu'il faut étudier simultanément.

Exemple 13.3.25 (La lemniscate de Bernoulli) *Soit C la courbe paramétrée par $\varphi : \mathbf{R} \to \mathbf{R}^2$ définie par*

$$t \mapsto \begin{cases} x(t) &= \dfrac{t}{t^4+1} \\ y(t) &= \dfrac{t^3}{t^4+1} \end{cases}$$

• *x et y sont impaires. On étudiera donc φ sur $\mathbf{R}_+$ et l'on déduira la courbe C par symétrie centrale. Par ailleurs, comme*

$$x(1/t) = y(t) \quad et \quad y(1/t) = x(t)$$

on peut se limiter à l'intervalle $[0,1]$ (on déduira C par symétrie orthogonale par rapport à la première bissectrice des axes).

• *On vérifie facilement que x et y sont dérivables, que C est régulière et que*

$$t \mapsto \begin{cases} x(t) &= \dfrac{1-3t^4}{(t^4+1)^2} \\ y(t) &= \dfrac{t^2(3-t^4)}{(t^4+1)^2} \end{cases}$$

On en déduit le tableau de variations

t	*0*		$3^{-1/4}$		*1*
x'	*1*	$+$	*0*	$-$	*-1/2*
x	*0*	↗	$3/(43^{-1/4})$	↘	*1/2*
y'	*0*	$+$	$*$	$+$	*1/2*
y	*0*	↗		↗	*1/2*

et finalement la courbe C

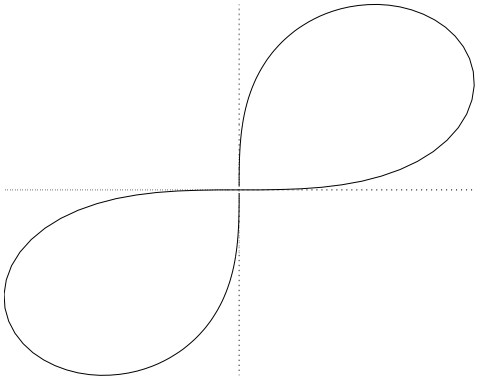


FIG. 13.26 – La lemniscate de Bernoulli

Exemple 13.3.26 (La strophoïde droite) *Soit C la courbe paramétrée par $\varphi : \mathbf{R} \to \mathbf{R}^2$ définie par*

$$t \mapsto \begin{cases} x(t) &= \frac{1-t^2}{1+t^2} \\ y(t) &= t\frac{1-t^2}{1+t^2} \end{cases}$$

• *Les applications x est paire et y est impaire. On se limite à étudier φ sur $[0, +\infty[$ et on déduira la courbe par symétrie par rapport à (Ox).*

• *Il est facile de vérifier que x et y sont dérivables sur $[0, +\infty[$ et*

$$\begin{cases} x'(t) &= -\frac{4t}{(1+t^2)^2} \\ y'(t) &= \frac{1-4t^2-t^4}{(1+t^2)^2} \end{cases}$$

• *Cette courbe possède une branche infinie en $+\infty$. En effet, on remarque que $\lim_{t\to+\infty} x(t) = -1$ et $\lim_{t\to+\infty} y(t) = +\infty$. La courbe C possède donc la droite $x + 1 = 0$ pour asymptote.*

• *On remarque que $(0, 0)$ est un point double (on peut montrer que c'est le seul) et la tangente en $t = 1$ a pour vecteur directeur $(1, 1)$.*

• *On peut étudier les variations de la courbe C*

t	*0*		$\sqrt{\sqrt{5}-2}$	$+\infty$
x'	*0*	$-$		$-$
x	*1*	↘		↘ *-1*
y'	*1*	$+$	0	$-$
y	*0*	↗		↘ *1/2* $-\infty$

et tracer sa représentation graphique

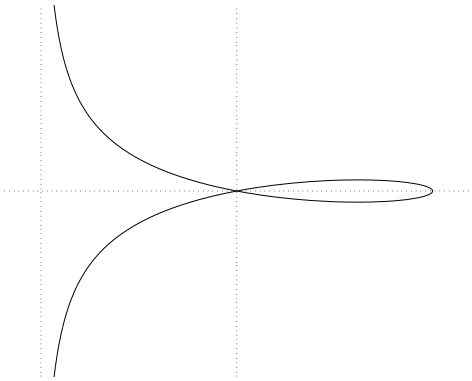


FIG. 13.27 – La strophoïde droite

Exercice 13.3.27 *Étudier la courbe C_1 paramétrée par $\varphi_1 : [0, \pi/6[\to \mathbf{R}^2$*

$$\begin{cases} x(t) &= \sin(2t) \\ y(t) &= \tan(3t) \end{cases}$$

Puis la courbe C_2 paramétrée par $\varphi_2 :]\pi/6, \pi/2[\to \mathbf{R}^2$ défini par les mêmes expressions. Tracer $C_1 \cup C_2$.

Exercice 13.3.28 *On définit le n-ième polynôme de Tchebycheff T_n par la relation*

$$\cos(nx) = T_n(\cos(x))$$

Démontrer les propriétés suivantes :

a. *$T_0 = 1$, $T_1(X) = X$ et, pour tout $n \geq 1$, $T_{n+1}(X) = 2XT_n(X) - T_{n-1}(X)$.*
b. *En déduire que $\deg(T_n) = n$, $T_n(1) = 1$ et $T_n(-1) = (-1)^n$, pour tout $n \geq 0$.*
c. *Trouver une expression de T_n en fonction de $t \in [-1, 1]$.*
d. *On définit les courbes $C_{m,n}$ paramétrées par l'application $\gamma_{m,n} :]-1, 1[\to \mathbf{R}^2$ définie par $t \mapsto (T_m(t), T_n(t))$. Ces courbes sont appelées courbes de Tchebycheff. Étudier $C_{2,3}$.*

Sujet d'étude 3 : le théorème des fonctions implicites

Une question naturelle est de se demander si les deux définitions, les deux points de vue, que l'on a des courbes, à savoir qu'une courbe est donnée par l'annulation d'une équation ou comme l'image d'un paramétrage, sont équivalents. Dans ce paragraphe nous donnons, à notre niveau, une réponse partielle à cette question. Précisément, le *théorème des fonctions implicites* assure que, *localement*, certaines courbes définies par l'annulation d'une équation peuvent être paramétrées.

*

Revenons donc à la première définition que l'on avait des courbes planes et considérons un sous-ensemble C de $\mathbf{R}^2$ défini par une équation de la forme

$$E(x,y) = \sum_{1\leq i,j\leq m} a_i(x)b_j(y)x^i y^j = 0$$

avec $a_i : I \subset \mathbf{R} \to \mathbf{R}$ et $b_j : I \subset \mathbf{R} \to \mathbf{R}$, pour tout $1 \leq i \leq m$ et tout $1 \leq i \leq m$, des fonctions *continûment dérivables* sur I, intervalle ouvert de $\mathbf{R}$. Cette définition est une généralisation de la notion de courbe algébrique, qui est le cas particulier avec les a_i et les b_j constantes.

Remarque 13.3.29 *On a vu (cf. exemple 13.3.4) que les courbes paramétrées pouvaient être des parties de tels ensembles. L'équation ci-dessus est appelée une équation implicite de C.*

Exercice 13.3.30 *Considérons $\gamma : \mathbf{R} \to \mathbf{R}^2$ l'application définie par*

$$t \mapsto \frac{e^t}{e^t+1}\left(\cos(t), \sin(t)\right)$$

Posons $\Gamma := \gamma(\mathbf{R})$. Montrer qu'il existe une suite $(x_n, y_n) \in \left(\mathbf{R}^2\right)^{\mathbf{N}}$ de points de Γ qui tend vers $(0,0)$ et que le point $(0,0)$ n'appartient pas à Γ. Conclure.

On a vu (cf. sujet d'étude 2) apparaître la notion de dérivées partielles d'une équation algébrique dans l'étude de ses tangentes. Cette notion se généralise.

Exercice 13.3.31 *Soit $M_0 = (x_0, y_0)$ un point de C. Montrer, en utilisant la dérivabilité des fonctions a_i et b_j pour tout i, j, que les applications $I \to \mathbf{R}$ définies par*

$$f : x \mapsto E(x, y_0) \quad et \quad g : y \mapsto E(x_0, y)$$

sont dérivables sur I. On note

$$\frac{\partial E}{\partial x}(x_0, y_0) := f'(x_0) \quad et \quad \frac{\partial E}{\partial y}(x_0, y_0) := g'(x_0)$$

On peut énoncer le théorème des fonctions implicites.

Théorème 13.3.32 *Soit C une courbe plane définie par une équation E de la forme*

$$\sum_{1\leq i,j\leq m} a_i(x)b_j(y)x^i y^j = 0$$

Pour tout $1 \leq i, j \leq m$, *les applications* $a_i,\ b_j : I \subset \mathbf{R} \to \mathbf{R}$ *sont supposées continûment dérivables sur l'intervalle ouvert* I *de* $\mathbf{R}$. *Soit* $M_0 = (x_0, y_0)$ *un point de* C. *On suppose que*

$$\frac{\partial E}{\partial y}(x_0, y_0) \neq 0$$

Alors l'équation $E(x, y)$ *peut être résolue localement en* y, *i.e. il existe un intervalle ouvert* V *de* $\mathbf{R}$ *contenu dans* I *et contenant* x_0, *un intervalle ouvert* W *de* $\mathbf{R}$ *contenu dans* I *et contenant* y_0, *une application* $\varphi : V \to W$ *(unique) continûment dérivable tels que*

$$(x \in V,\ y \in W \text{ et } E(x, y) = 0) \Longleftrightarrow (x \in V \text{ et } y = \varphi(x))$$

En outre, on a la relation

$$\varphi'(x) = -\frac{\frac{\partial E}{\partial x}(x, \varphi(x))}{\frac{\partial E}{\partial y}(x, \varphi(x))},\ \forall x \in V$$

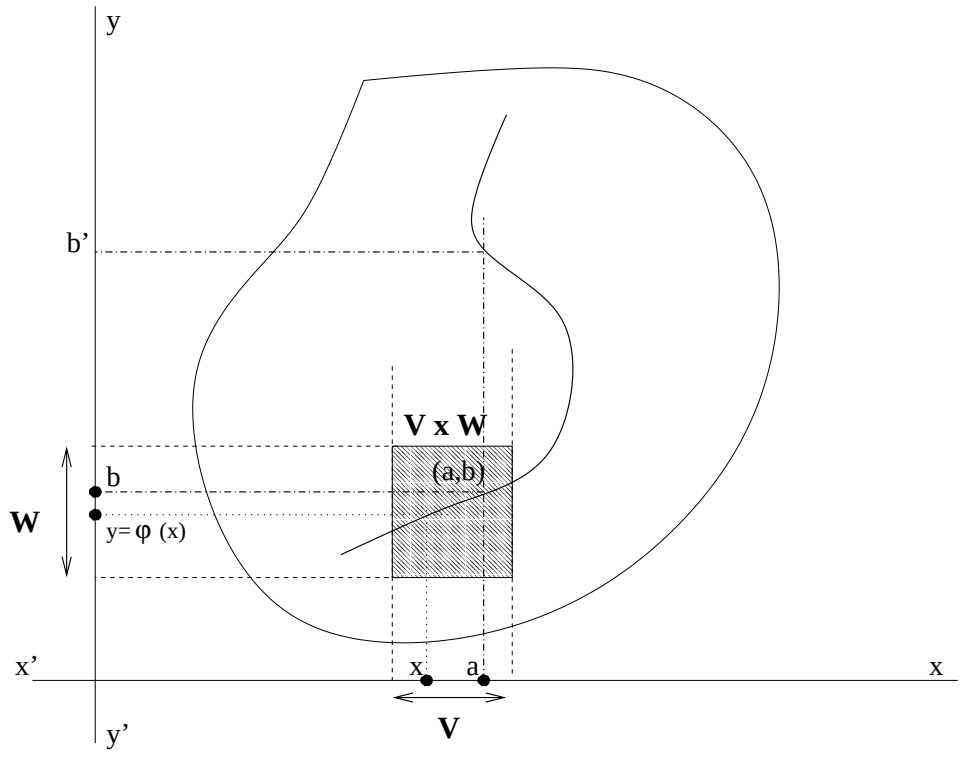


FIG. 13.28 – Le théorème des fonctions implicites

Autrement dit, ce théorème signifie que l'équation implicite $E(x, y) = 0$ de C peut être définie, au moins localement, comme le graphe d'une fonction $x \mapsto \varphi(x)$. Ceci implique en particulier que notre courbe C est (localement) paramétrisable par $\varphi : V \to \mathbf{R}^2$

$$\begin{cases} x(t) &= t \\ y(t) &= \varphi(t) \end{cases}$$

Exemple 13.3.33 *Pour l'équation $x^2+y^2-1=0$ et $M=(x,y)$ appartenant au cercle unité C, on a*

$$\frac{\partial E}{\partial y}(x,y)=2y$$

En prenant $M_0=(0,1)$, on voit que cette équation définit la fonction implicite

$$\varphi:]-1,1[\to \mathbf{R}_+^*$$

par $x\mapsto\sqrt{1-x^2}$.

Exercice 13.3.34 *Montrer que la courbe d'équation $y^2-x^3-x=0$ est paramétrisable au voisinage du point $(0,0)$.*

Remarque 13.3.35 *Le théorème des fonctions implicites est essentiellement un énoncé d'existence en ce sens qu'il est en général difficile d'expliciter la fonction implicite. Ce problème n'est pas sans relation avec la théorie des équations différentielles et plus précisément avec le Problème de Cauchy ou le théorème de Cauchy-Lipschitz. En effet, "expliciter" cette fonction, c'est résoudre l'équation différentielle*

$$\varphi'(x)=-\frac{\dfrac{\partial E}{\partial x}(x,\varphi(x))}{\dfrac{\partial E}{\partial y}(x,\varphi(x))}$$

avec la condition initiale $\varphi(x_0)=y_0$. Autrement dit, le théorème de Cauchy-Lipschitz assure que la fonction implicite est l'unique solution du problème de Cauchy

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

en posant $f:(x,y)\mapsto -\dfrac{\partial E}{\partial x}(x,y)/\dfrac{\partial E}{\partial y}(x,y)$.

Exercice 13.3.36 *Soit C la conique d'équation $x^2+y^2-1=0$. Retrouver le résultat de l'exemple 13.3.33 en résolvant une équation différentielle adaptée.*

Remarque 13.3.37 *La théorie des développements limités (formule de Taylor) permet également de s'attaquer au problème d'expliciter les fonctions implicites, en en donnant une approximation au voisinage du point de la courbe considéré.*

Exercice 13.3.38 *Montrer que la courbe d'équation*

$$\sin(y)+xy^4+x^4=0$$

peut être paramétrée au voisinage de $(0,0)$. Montrer que la fonction implicite définie par cette équation au voisinage de $(0,0)$ est de la forme $\varphi(x)=x^2-x^6/6+\varepsilon(x)$ avec $\varepsilon\to 0$ quand $x\to 0$.

Quatrième partie

Méthodes de calcul.

Calculer s'apprend par la pratique, mais il est préférable de le faire de manière méthodique. Réciproquement, les techniques exposées ici ne deviendrons familières, qu'en essayant de les appliquer.

Chapitre 14

La méthode du pivot pour la résolution de systèmes linéaires.

14.1 Les systèmes linéaires comme équations entre matrices

On s'intéresse à un *système* de n équations et m inconnues

$$\left\{\begin{array}{ccccccc} a_{11}x_1 & + & \cdots & + & a_{1m}x_m & = & b_1 \\ a_{21}x_1 & + & \cdots & + & a_{2m}x_m & = & b_2 \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ a_{n1}x_1 & + & \cdots & + & a_{nm}x_m & = & b_n \end{array}\right.$$

Ici les a_{ij} sont des éléments d'un ensemble de nombres R tel que $\mathbf{Q}$, $\mathbf{R}$ ou $\mathbf{C}$. [1] Les solutions $(x_1, \ldots, x_m)$ de ce système sont des éléments de R^m.

Nous allons voir comment transformer un tel système en un système sur lequel on voit clairement les solutions, *sans changer l'ensemble des solutions*. Pour cela nous allons nous inspirer du cas d'un "système" à une équation et à une inconnue : $ax = b$. Il est facile de trouver une solution de ce système, si a est inversible. Alors $x = a^{-1}b$. Or dans un corps tout $a \neq 0$ est inversible et ce cas est donc complètement traité. Mais la remarque qui va nous servir de guide est plutôt, que si c est inversible, alors $ax = b$ a les mêmes solutions que $cax = cb$ (montrer cette affirmation). Nous allons écrire le système qui nous intéresse comme une équation $AX = B$ dans un système de nombres généralisés : les matrices.

Une *matrice* à coefficients dans R n'est rien d'autre qu'un tableau de nombres $A = (a_{ij})$ de R. On dit qu'une matrice est de taille $n \times m$ si elle a n lignes et m colonnes. L'ensemble des matrices $n \times m$ à coefficients dans R est noté $M_{n,m}(R)$. Nous allons introduire un produit sur les matrices, tel que si A est une matrice de $M_{n,m}(R)$ et si B un élément de R^n, alors la résolution du système revient à trouver les solutions en X, élément de R^m, de l'équation

$$AX = B \ .$$

Ici, on identifie R^k à l'ensemble des matrices colonne $M_{k,1}(R)$.

Disons qu'une matrice carrée C de taille $n \times n$ est *inversible*, s'il existe une matrice (de même taille) D telle que

$$CD = I_n = DC \ ,$$

[1] Ce qui suit s'applique pour R un corps (voir l'Appendice).

où I_n est la *matrice identité*, dont les seuls coefficients non-nuls sont les termes a_{ii} sur la diagonale, qui sont égaux à 1 :

$$I_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ . & . & \dots & . \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

Nous affirmons que

si C est inversible alors les systèmes $AX = B$ et $CAX = CB$ ont les mêmes solutions.

On dit que deux systémes d'équations linéaires qui ont le même ensemble de solutions sont *équivalents*. Il est clair que si le produit de matrices obéit aux règles usuelles, alors le fait que X soit solution de $AX = B$, implique que X soit aussi solution de $CAX = CB$ (ici on n'utilise pas que C est inversible). Réciproquement, si X est solution de $CAX = CB$, alors vu que C est inversible il existe D telle que $DC = I$ et $D(CAX) = D(CB)$ donne bien $AX = B$.

Il s'agit donc de mettre en évidence des matrices inversibles. Celles qui vont nous intéresser seront obtenues à partir de la matrice identité par des *opérations élémentaires*.

Il y a trois types d'opérations élémentaires sur les lignes d'une matrice dans $M_{n,m}(R)$:

I) échanger deux lignes de la matrice

II) multiplier une ligne de la matrice par un élément a de R *non-nul*

III) ajouter un multiple d'une ligne de la matrice à une autre ligne de la matrice.

Les *matrices élémentaires* $n \times n$ sont les matrices obtenues à partir de la matrice identité I_n par une opération élémentaire.

Exemples. Voici des matrices élémentaires 3×3, pour $a \neq 0$ dans R

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} a & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 & a \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

On vérifie que les matrices élémentaires sont inversibles (l'inverse multiplicatif de a non-nul existe) et que de

faire une opération élémentaire sur les lignes d'une matrice revient à multiplier cette matrice à gauche par la matrice élémentaire correspondante.

Avec ce que nous avons dit précédemment on arrive à la conclusion, que

on ne change pas l'ensemble de solutions d'un système $AX = B$ en le transformant par une opération élémentaire sur les lignes.

Ce qui précède donne une méthode–la *méthode du pivot* (de Gauss)–, qui permet de transformer tout système d'équations linéaires en un système d'équations équivalent, et qui a la propriété d'être facilement résoluble. Pour la méthode du pivot on utilise seulement des opérations élémentaires sur les lignes.

Exemple. On considère le système $AX = B$ associé aux matrices

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

On travaille sur la matrice étendue

$$(A|B) = \left(\begin{array}{cccc|c} 3 & 2 & 3 & -2 & 1 \\ 1 & 1 & 1 & 0 & 3 \\ 1 & 2 & 1 & -1 & 2 \end{array}\right)$$

On va transformer cette matrice en la matrice

$$\left(\begin{array}{cccc|c} 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right)$$

Il est clair comment trouver les solutions du système associé! Les solutions $(x_1, \ldots, x_4)$ sont donnés par $x_4 = 3$, $x_2 = 2$ et $x_1 + x_3 = 1$. On peut vérifier qu'il s'agit bien de solutions du système initial. Une autre façon de présenter les solutions est

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 0 \\ 3 \end{pmatrix} + x_3 \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

Sous cette forme il apparaît clairement que le système homogène associé $AX = 0$ admet une droite de solutions, qui a pour base $(-1, 0, 1, 0)$.

Voici les étapes pour transformer $(A|B)$ avec la méthode du pivot. On échange la première et la troisième ligne ($L_1 \leftrightarrow L_3$), pour avoir un 1 en haut à gauche, ce qui donne la matrice

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & -1 & 2 \\ 1 & 1 & 1 & 0 & 3 \\ 3 & 2 & 3 & -2 & 1 \end{array}\right).$$

Le 1 ainsi obtenu est le premier pivot. On utilise maintenant ce pivot pour mettre des zéros ailleurs dans la première colonne. Pour ça on fait $L_2 \to L_2 - L_1$ et $L_3 \to L_3 - 3L_1$, ce qui donne la matrice

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & -1 & 2 \\ 0 & -1 & 0 & 1 & 1 \\ 0 & -4 & 0 & 1 & -5 \end{array}\right).$$

On multiplie la deuxième ligne par -1 pour obtenir le deuxième pivot ($L_2 \to -L_2$), et ensuite on fait $L_3 \to L_3 + 4L_2$, ce qui donne la matrice

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & -1 & 2 \\ 0 & 1 & 0 & -1 & -1 \\ 0 & 0 & 0 & -3 & -9 \end{array}\right).$$

Après avoir fait $L_3 \to -\frac{1}{3}L_3$ on obtient la matrice triangulaire

$$\left(\begin{array}{cccc|c} 1 & 2 & 1 & -1 & 2 \\ 0 & 1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 1 & 3 \end{array}\right).$$

On peut déjà lire les solutions à ce stade, mais il est mieux de continuer en remontant et en mettant aussi des zéros au-dessus des pivots. On obtient ainsi le résultat annoncé. [2]

Une matrice comme celle à laquelle nous nous sommes ramenés dans l'exemple est dite réduite. Plus généralement une matrice est dite *réduite* si :

1) toute ligne contenant un terme non-nul précède toute ligne identiquement nulle

[2] On voit que pour appliquer la méthode du pivot nous avons besoin de l'existence de l'inverse multiplicatif de tout élément non-nul de R : c'est exactement pour ça que nous demandons à R d'être un corps.

2) le premier terme non-nul d'une ligne non identiquement nulle est 1 et se trouve à la droite du premier terme non-nul de la ligne précédente (ce terme est appelé un *pivot*)
3) toute colonne qui contient un pivot ne contient que cet élément comme élément non-nul.

En résumé :

> *la méthode du pivot de Gauss permet de transformer une matrice, par des opérations élémentaires sur ses lignes, en une matrice réduite ; les sytèmes associés à la matrice de départ et à la matrice réduite sont équivalents.*

Par ailleurs on peut aussi montrer, que la matrice réduite associée par la méthode du pivot à une matrice donnée ne dépend pas du choix des opérations élémentaires employées dans la réduction.

On peut utiliser la méthode du pivot pour vérifier si une matrice carrée est inversible. Soit B une matrice $n \times n$. On considère la matrice étendue $(A|I_n)$, si la matrice réduite associée à cette matrice est $(I_n|B)$, alors B est l'inverse de A. Plus précisément, si E représente la matrice produit des matrices élémentaires utilisées dans la méthode du pivot, alors $E = B$. En particulier

> *une matrice carrée est inversible si et seulement si elle est le produit de matrices élémentaires.*

Ce qui nous reste à faire est de justifier le calcul matriciel.

14.2 Calcul matriciel I : suite de Fibonacci.

La *suite de Fibonacci* $\{u_n\}$ est la suite définie par les conditions

$$\left\{ \begin{array}{lcl} u_0 & = & 1 \\ u_1 & = & 1 \\ u_n & = & u_{n-1} + u_{n-2} \ . \end{array} \right.$$

Ainsi $u_2 = 2$, $u_3 = 3$, $u_4 = 5$, *etc.* On se propose de trouver une méthode pour déterminer la valeur de l'entier u_n, sans calculer tous les termes précédents. Une traduction du problème en termes concrets pourrait être la suivante : trouver de combien façons on peut vider un tonneau de n litres avec un pot de 1 litre et un pot de 2 litres. Ainsi un tonneau de 3 litres peut être vidé de $u_3 = 3$ manières différentes : en prélevant 1 litre puis 2 litres, ou bien 2 litres puis 1 litre, ou bien trois fois 1 litre. Fibonacci lui-même avait formulé le problème comme celui de la détermination de la croissance d'une population de lapins.

Il y a plusieurs manières de résoudre le problème. Ce que nous allons faire est étudier l'opération A qui, pour une valeur de n donnée, mène du couple $x_{n-2} = (u_{n-2}, u_{n-1})$ au couple $x_{n-1} = (u_{n-1}, u_n)$. L'opération A elle-même ne dépend pas de n. Il est clair que l'on arrive à u_n à partir du couple $x_0 = (1, 1)$ en itérant $n-1$ fois l'opération A . L'idée est de se dire qu'il y a des suites de même nature, que la suite de Fibonacci pour lesquelles la solution est facile et d'essayer de se ramener à une telle suite.

Dans ce qui suit nous allons écrire les couples sous forme de colonnes, mais il nous arrivera de ne pas faire la différence entre couples-lignes et couples-colonnes. En symboles on a donc l'égalité

$$\begin{pmatrix} u_n \\ u_{n+1} \end{pmatrix} = A^n \begin{pmatrix} 1 \\ 1 \end{pmatrix} \ ,$$

que nous lisons "on obtient (la colonne associée à) le couple (u_n, u_{n+1}) en appliquant n fois l'opération A (à la colonne associée) au couple $(1, 1)$". Ce qui va nous permettre de résoudre le problème posé est le fait que cette écriture symbolique correspond en fait à une égalité entre quantités sur lesquelles sont définies des opérations algébriques.

Nous avons défini une matrice comme étant un tableau de nombres. On peut considérer des tableaux de taille quelconque, mais pour nos besoins immédiats nous allons nous borner à des tableaux ayant au plus deux lignes et deux colonnes. Voici trois matrices :

$$\begin{pmatrix} 1 & 2 \\ 0 & 5 \end{pmatrix} \quad , \quad \begin{pmatrix} 1/2 \\ 4/7 \end{pmatrix} \quad , \quad (3\ 4)\ .$$

La première est une matrice à deux lignes et deux colonnes, la deuxième une matrice à deux lignes et une colonne, la troisième une matrice à une ligne et deux colonnes. Comme indiqué on parle de matrices 2×2, 2×1 et 1×2. (Faire attention à la convention qui compte d'abord le nombre de lignes et ensuite le nombre de colonnes.)

De manière assez banale on peut définir une *somme* sur les matrices : on somme les coefficients correspondants. Ainsi la somme de deux matrices 2×2 est définie par l'égalité

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} a+a' & b+b' \\ c+c' & d+d' \end{pmatrix}\ .$$

Il est facile de vérifier que cette opération somme a toutes les propriétés de l'opération somme sur les nombres. On pourrait aussi définir un produit banal en multipliant les coefficients de même indice. De manière beaucoup moins banale on peut définir le *produit* de matrices de taille convenable. Le produit de deux matrices 2×2 est défini par l'égalité

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} aa'+bc' & ab'+bd' \\ ca'+dc' & cb'+dd' \end{pmatrix}\ .$$

Ce qui n'est pas banal est le fait qu'avec ce produit on peut multiplier les matrices "sans parenthèses", c'est-à-dire que si A, B et C sont trois matrices telles que tous les produits considérés sont définis, alors

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

De plus $A(B+C) = AB + AC$ et $(A+B)C = AC + BC$. Par contre, en général, $AB \neq BA$, ainsi par exemple

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \neq \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

(faire le calcul !). Le produit d'une matrice 2×2 avec une matrice 2×1 est défini par l'égalité

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e \\ f \end{pmatrix} = \begin{pmatrix} ae+bf \\ ce+df \end{pmatrix}\ .$$

Noter que le produit d'une matrice 2×2 avec une autre matrice 2×2 peut être vu comme la juxtaposition des colonnes obtenues en faisant le produit de la première matrice par les deux colonnes de la deuxième.

Exercice. Vérifier qu'avec les conventions ci-dessus, la définition de la suite de Fibonacci s'écrit

$$\begin{pmatrix} u_{n-1} \\ u_n \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u_{n-2} \\ u_{n-1} \end{pmatrix}\ .$$

Notons par X_n la colonne correspondant au couple x_n et posons

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

Avec ces notations on a bien l'égalité

$$X_n = A^n \begin{pmatrix} 1 \\ 1 \end{pmatrix} ,$$

où ici la puissance n-ième de A est le produit de la matrice A avec elle-même n fois. L'idée est alors de se dire que, *si A était diagonale*, à savoir de la forme

$$D = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix} ,$$

alors ce serait facile, vu que la puissance n-ième D^n de D est simplement la matrice diagonale ayant les puissances n-ièmes des termes sur la diagonale comme éléments diagonaux

$$D^n = \begin{pmatrix} \alpha^n & 0 \\ 0 & \beta^n \end{pmatrix} .$$

Voici comment ramener A à une matrice diagonale. Posons

$$\alpha = \frac{1+\sqrt{5}}{2} \quad \text{et} \quad \beta = \frac{1-\sqrt{5}}{2}$$

puis

$$P = \begin{pmatrix} 1 & 1 \\ \alpha & \beta \end{pmatrix} \quad \text{et} \quad P^{-1} = \begin{pmatrix} -\beta/\sqrt{5} & 1/\sqrt{5} \\ \alpha/\sqrt{5} & -1/\sqrt{5} \end{pmatrix} .$$

Alors on vérifie que P^{-1} est *l'inverse* de P dans le sens que le produit PP^{-1} égale la matrice identité 2×2 I_2, et $A = PDP^{-1}$. On en déduit l'égalité clef

$$A^n = PD^nP^{-1} ,$$

qui nous permet d'écrire

$$X_n = \begin{pmatrix} \alpha^n & \beta^n \\ \alpha^{n+1} & \beta^{n+1} \end{pmatrix} \begin{pmatrix} \alpha/\sqrt{5} \\ -\beta/\sqrt{5} \end{pmatrix} ,$$

et de trouver la solution à notre problème sous la forme

$$u_n = \frac{\alpha^{n+1}}{\sqrt{5}} - \frac{\beta^{n+1}}{\sqrt{5}} .$$

Exercice. a) Calculer les valeurs de u_{10} et de u_{20}.
b) Vérifier que le membre de droite de cette égalité est bien un entier.
c) Utiliser le fait que $0 < -\beta < 1$, pour montrer que u_n égale la partie entière de $\alpha^{n+1}/\sqrt{5}$.

Le lecteur peut légitimement se demander comment nous sommes arrivés aux valeurs de α et β ci-dessus. Voici une approche possible. On observe que si M est une matrice diagonale, disons

$$M = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}, \text{et si } e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad , \quad e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} ,$$

alors $Me_1 = \lambda e_1$ et $Me_2 = \mu e_2$. On dit que e_1 et e_2 sont des *vecteurs propres* de M pour les *valeurs propres* λ et μ respectivement.(Pour qu'une colonne soit vecteur propre on demande que ses coefficients ne soient pas les deux nuls, ce qui est le cas de e_1 et e_2). Ainsi, une matrice diagonale a des valeurs

propres et nous allons trouver une condition pour qu'un nombre λ soit valeur propre d'une matrice. Considérons une matrice générale

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

et supposons que le nombre λ soit valeur propre de M, disons

$$M \begin{pmatrix} u \\ v \end{pmatrix} = \lambda \begin{pmatrix} u \\ v \end{pmatrix} .$$

Si on retranche le membre de gauche du membre de droite de cette dernière égalité (opérations sur les matrices !), on obtient le système d'équations

$$\begin{cases} (a-\lambda)u & + & bv & = & 0 \\ cu & + & (d-\lambda)v & = & 0 \end{cases} .$$

En multipliant la première équation par $-c$ et la deuxième par $(a-\lambda)$ et en additionnant les deux équations, on trouve l'équation

$$((a-\lambda)(d-\lambda) - bc)v = 0 .$$

En multipliant la première équation par $(d-\lambda)$ et la deuxième par $-b$ et en additionnant les deux équations, on trouve l'équation

$$((a-\lambda)(d-\lambda) - bc)u = 0 .$$

Donc si u ou v est différent de 0, on a $(a-\lambda)(d-\lambda) - bc = 0$ et en développant le membre de gauche $\lambda^2 - (a+d)\lambda + (ad-bc) = 0$, ainsi λ est racine du polynôme

$$X^2 - (a+d)X + (ad-bc) ,$$

d'où on tire que λ égale une des deux racines de ce polynôme, à savoir

$$\frac{a+d}{2} \pm \frac{1}{2}\sqrt{(\frac{a-d}{2})^2 + bc} .$$

Notons λ et μ ces racines. On peut vérifier par calcul direct, que ceux qui suivent sont vecteurs propres de M :

$$\begin{pmatrix} \lambda - d \\ c \end{pmatrix} \text{ et } \begin{pmatrix} b \\ \mu - a \end{pmatrix} .$$

La matrice Q ayant ces matrices 2×1 pour colonnes permet de transformer M en une matrice diagonale. On suppose que les deux colonnes ne sont pas multiple l'une de l'autre. Alors Q admet un inverse. Plus généralement toute matrice

$$N = \begin{pmatrix} e & f \\ g & h \end{pmatrix}$$

avec $eh - gf \neq 0$ a pour inverse

$$N^{-1} = \frac{1}{eh-gf} \begin{pmatrix} h & -f \\ -g & e \end{pmatrix}$$

(simplement effectuer le produit). En appliquant ceci à la matrice Q on vérifie que $Q^{-1}MQ$ est diagonale.

Exercice. Appliquer au cas de la suite de Fibonacci.

Dans ce qui précède nous avons vu apparaître un certain nombre de quantités associées à une matrice 2×2

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

- la *trace* de M : la somme des termes sur la (première) diagonale $\mathrm{tr}(M) = a + d$;
- le *déterminant* de M : $\det(M) = ad - bc$;
- le *polynôme caractéristique* de M : $X^2 - \mathrm{tr}(M)X + \det(M)$.

(A un signe près) le polynôme caractéristique a la trace et le déterminant pour coefficients et les valeurs propres sont les racines du polynôme caractéristique. De plus nous avons vu que la non-nullité du déterminant permet de donner une expression explicite pour l'inverse. (En fait si le déterminant d'une matrice est nul, alors la matrice n'admet pas d'inverse.) Ces observations se généralisent au cas des matrices de taille plus grande et sont à la base de leur étude.

Exercice. Montrer que le déterminant du produit de deux matrices égale le produit des déterminants des matrices.

14.3 Calcul matriciel II : nombres complexes.

Il est utile de penser aux matrices comme à des nombres généralisés, que l'on peut additionner et multiplier comme d'habitude, avec la seule différence, que (1) les opérations ne peuvent s'effectuer qu'à certaines conditions sur la taille et (2) il importe de surveiller l'ordre d'écriture des termes dans un produit.

Avant d'aborder le calcul matriciel dans le cas général, nous allons voir comment retrouver une copie de l'ensemble des nombres complexes dans l'ensemble $M_2(\mathbf{R})$ des matrices 2×2 à coefficients réels.

Tout d'abord, notons que l'on peut identifier l'ensemble des réels eux-mêmes au sous-ensemble de $M_2(\mathbf{R})$ des matrices diagonales

$$\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}$$

avec deux termes égaux sur la diagonale, et avec en particulier 1 identifié à la matrice

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

Ces matrices se comportent par rapport aux opérations de somme et de produit exactement comme des nombres. Le point de départ de l'identification des complexes dans $M_2(\mathbf{R})$ est l'observation que la matrice

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

est telle que $\mathbf{i}^2 = -\mathbf{1}$. Il est alors naturel de s'attendre que le nombre complexe $a + bi$ de partie réelle a et partie imaginaire b soit représenté par

$$a\mathbf{1} + b\mathbf{i} = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix} + \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix} = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} .$$

On vérifie les règles usuelles de multiplication des nombres complexes :

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} \begin{pmatrix} c & d \\ -d & c \end{pmatrix} = \begin{pmatrix} ac - bd & ad + bc \\ -(ad + bc) & ac - bd \end{pmatrix} = \begin{pmatrix} c & d \\ -d & c \end{pmatrix} \begin{pmatrix} a & b \\ -b & a \end{pmatrix} ,$$

c'est-à-dire, qu'en particulier, la partie réelle du produit des nombres complexes ayant pour partie réelle a et c, et partie imaginaire b et d égale $ac - bd$. On retrouve le module d'un nombre complexe grâce au déterminant :

$$\det \begin{pmatrix} a & b \\ -b & a \end{pmatrix} = a^2 + b^2 ,$$

et le conjugué grâce à la transposition :

$$\overline{\begin{pmatrix} a & b \\ -b & a \end{pmatrix}} = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} .$$

En résumé, avec ce qui précède, nous avons défini une application

$$\begin{array}{ccc} \mathbf{C} & \longrightarrow & M_2(\mathbf{R}) \\ z = a + bi & \mapsto & \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \end{array} ,$$

qui respecte les opérations et que l'on vérifie être une injection. Donc, d'une certaine manière l'ensemble des matrices 2×2 à coefficients réels peut être vu comme un système de nombres, qui est une extension du système de nombres que sont les nombres complexes : les mathématiciens des années trente du siècle passé parlaient de *systèmes de nombres hypercomplexes*.

14.4 Calcul matriciel III : matrices quelconques.

Pour des matrices plus générales on définit des opérations comme suit.

La somme de matrices. Soit $A = (a_{ij})$ et $B = (b_{ij})$ des matrices de $M_{m,n}(R)$, alors la somme $A + B$ est définie comme étant la matrice $C = (c_{ij})$ de $M_{m,n}(R)$ définie par

$$c_{ij} = a_{ij} + b_{ij} .$$

Le produit de matrices. Soit A une matrice de $M_{n,m}(R)$ et soit B une matrice de $M_{m,l}(R)$, alors le produit AB est la matrice $C = (c_{ik})$ de $M_{n,l}(R)$ définie par

$$c_{ik} = \sum_{j=1}^{m} a_{ij} b_{jk} .$$

On peut montrer que le produit de matrices est associatif : $A(BC) = (AB)C$ (c'est direct, mais il ne faut pas avoir peur des indices). Les autres propriétés mises en évidence pour les matrices 2×2 restent vrai en général.

Chapitre 15

Calcul de primitives

15.1 Primitives des fonctions usuelles ; antidérivation

Domaine	Fonction	Primitive
$\mathbf{R}$	$x^n,\ n \in \mathbf{N}$	$\frac{x^{n+1}}{n+1}$
$\mathbf{R}^*$	$x^n,\ n \in \mathbf{Z},\ n < -1$	$\frac{x^{n+1}}{n+1}$
$]0;+\infty[$	$x^a,\ a \in \mathbf{R},\ a \neq -1$	$\frac{x^{a+1}}{a+1}$
$\mathbf{R}^*$	$\frac{1}{x}$	$\ln\lvert x\rvert$
$\mathbf{R}$	e^x	e^x
$\mathbf{R}$	$\cos x$	$\sin x$
$\mathbf{R}$	$\sin x$	$-\cos x$
$\mathbf{R}$	$\operatorname{ch} x$	$\operatorname{sh} x$
$\mathbf{R}$	$\operatorname{sh} x$	$\operatorname{ch} x$
$\bigcup_{k\in\mathbf{Z}}]-\frac{\pi}{2}+k\pi, \frac{\pi}{2}+k\pi[$	$1+\tan^2 x = \frac{1}{\cos^2 x}$	$\tan x$
$\bigcup_{k\in\mathbf{Z}}]k\pi, (k+1)\pi[$	$1+\operatorname{cotan}^2 x = \frac{1}{\sin^2 x}$	$-\operatorname{cotan} x$
$]-1,1[$	$\frac{1}{\sqrt{1-x^2}}$	$\arcsin x$
$]-1,1[$	$\frac{1}{\sqrt{1-x^2}}$	$-\arccos x$
$\mathbf{R}$	$\frac{1}{1+x^2}$	$\arctan x$
$\mathbf{R}$	$1-\operatorname{th}^2 x = \frac{1}{\operatorname{ch}^2 x}$	$\operatorname{th} x$
$\mathbf{R}$	$1-\coth^2 x = -\frac{1}{\operatorname{sh}^2 x}$	$\coth x$
$\mathbf{R}$	$\frac{1}{\sqrt{x^2+1}}$	$\operatorname{argsh} x$ ou bien $\ln(x+\sqrt{x^2+1})$
$]-\infty,-1[\cup]1,+\infty[$	$\frac{1}{\sqrt{x^2-1}}$	$\varepsilon \operatorname{argch}(\varepsilon\, x)$ ou bien $\ln\left\lvert x+\sqrt{x^2-1}\right\rvert$ avec $\varepsilon = \operatorname{sign}(x)$

Ces affirmations se vérifient tout simplement en dérivant les fonctions dans la colonne de droite.

15.2 Techniques de calculs

15.2.a Intégration par parties

On réécrit la formule d'intégration par parties du Chap. 10.2. Soient f et g deux fonctions définies, continues, dérivables à dérivées continues sur un intervalle I. Alors on a :

$$\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx \ .$$

Exemples.

– En posant $f(x) = \ln x$ et $g(x) = x$, on obtient

$$\int \ln x dx = x \ln x - x + c, \text{pour tout } x > 0 \ .$$

– En posant $f(x) = \arctan x$ et $g(x) = x$, on obtient

$$\int \arctan x dx = x \arctan x - \frac{1}{2} \ln(1 + x^2) + c, \text{ pour tout } c \in \mathbf{R} \ .$$

15.2.b Changement de variable

La formule de changement de variable devient, pour u une fonction définie, continue, dérivable à dérivée continue sur un intervalle I, et f une fonction définie et continue sur l'intervalle $J = u(I)$. Si F est une primitive de f sur $u(I)$, alors

$$\int f(u(x))u'(x)dx = F(u(x)) \ .$$

15.3 Primitives classiques

15.3.a Fractions rationnelles

Décomposition en éléments simples
Les résultats de ce paragraphes seront admis pour la plupart : la théorie des fractions rationnelles fait intervenir des résultats d'algèbre et d'arithmétique qui ne sont pas abordés en détail ici.

Définition. On appelle *fraction rationnelle* une fonction numérique qui s'écrit comme le quotient de deux polynômes de $\mathbf{R}[x]$, c'est-à-dire $f(x) = P(x)/Q(x)$ avec $P(x)$ et $Q(x)$ éléments de $\mathbf{R}[x]$. L'ensemble des fractions rationnelles à coefficients dans $\mathbf{R}$ est noté $\mathbf{R}(x)$.

Commençons par énoncer et donner quelques exemples de propriétés des polynômes :

Définition. Étant donnés deux polynômes $A(x)$ et $B(x) \neq 0$ dans $\mathbf{R}[x]$, on appelle *division euclidienne* de $A(x)$ par $B(x)$, l'unique couple de polynômes $(Q(x), R(x))$ tel que :

$$A(x) = B(x)Q(x) + R(x) \ ,$$

avec $R(x) = 0$, ou degré$(R) <$ degré(B).

Proposition. Tout polynôme $P(x) \in \mathbf{R}[x]$ se décompose de manière unique (à l'ordre des facteurs près) en un produit de polynômes irréductibles unitaires (*i.e.* le coefficient du terme de plus haut degré est 1) et d'une constante. Pour préciser :

– Dans $\mathbf{R}[x]$, les seuls polynômes irréductibles unitaires sont les constantes, les polynômes de degré 1, et les polynômes du second degré à discriminant strictement négatif (ax^2+bx+c avec $b^2-4ac<0$).
– Tout polynôme de $\mathbf{R}[x]$ ($P(x)\neq 0$) s'écrit alors :

$$P(x) = a\prod_{1\leq k\leq p}(x-a_k)^{\alpha_k}\prod_{1\leq k\leq q}(x^2+b_kx+c_k)^{\beta_k}$$

les facteurs étant tous deux à deux distincts ; tous les exposants sont supérieurs ou égaux à 1 ; pour tout $k=1,\cdots,q,\ b_k^2-4c_k<0$ et a est le coefficient directeur de $P(x)$ (*i.e.* le coefficient du terme de plus haut degré).

Remarque. Étant donné un polynôme, il est difficile de le factoriser comme indiqué dans la proposition précédente. Cela revient essentiellement à calculer ses racines dans $\mathbf{C}$.

Exemple : $3x^6-6x^3+3 = 3(x-1)^2(x^2+x+1)^2$

Théorème. Soit $f(x)=P(x)/Q(x)$ une fraction rationnelle de $\mathbf{R}(x)$; on suppose que $P(x)$ et $Q(x)$ n'ont pas de diviseurs communs. Soit

$$Q(x) = a\prod_{1\leq k\leq p}(x-a_k)^{\alpha_k}\prod_{1\leq k\leq q}(x^2+b_kx+c_k)^{\beta_k}$$

la décomposition de $Q(x)$ en facteurs premiers. Alors, on a de manière unique :

$$f(x) = E(x)+\sum_{1\leq i\leq p}\sum_{1\leq j\leq\alpha_i}\frac{A_{i,j}}{(x-a_i)^j}+\sum_{1\leq i\leq q}\sum_{1\leq j\leq\beta_j}\frac{B_{i,j}x+C_{i,j}}{(x^2+b_ix+c_i)^j}$$

où les $A_{i,j}$, $B_{i,j}$ et $C_{i,j}$ sont des constantes réelles et $E(x)$ un polynôme de degré degré($P(x)$)-degré($Q(x)$) (c'est 0 si degré($P(x)$) <degré($Q(x)$), appelé *partie entière* de la fraction rationnelle $f(x)$.

Les $A_{i,j}/(x-a_i)^j$ sont appelés *éléments simples de première espèce.* Les $(B_{i,j}x+C_{i,j})/(x^2+b_ix+c_i)^j$ sont appelés *éléments simples de deuxième espèce.* L'écriture de f qui précède s'appelle la *décomposition de f en éléments simples* dans $\mathbf{R}(x)$.

Ce théorème assure l'existence et l'unicité de la décomposition, mais ne dit pas comment calculer $E(x)$ ainsi que les coefficients $A_{i,j}$, $B_{i,j}$ et $C_{i,j}$. C'est ce que nous allons voir maintenant.

Calcul de la décomposition de f en éléments simples. L'unicité de la décomposition en éléments simples permet d'utiliser un certains nombre de recettes pour en calculer les coefficients.

– Tout d'abord, la partie entière $E(x)$ n'est rien d'autre que le quotient de la division euclidienne de $P(x)$ par $Q(x)$. On commence donc par écrire :

$$P(x) = Q(x)E(x)+R(x), \text{avec degré}(R(x))<\text{degré}(Q(x)),\ \text{ou } R(x)=0.$$

On en déduit $f(x)=E(x)+R(x)/Q(x)$, avec les mêmes conditions.
– $R(x)/Q(x)$ s'écrit alors $\sum_{1\leq i\leq p}\sum_{1\leq j\leq\alpha_i}\frac{A_{i,j}}{(x-a_i)^j}+\sum_{1\leq i\leq q}\sum_{1\leq j\leq\beta_j}\frac{B_{i,j}x+C_{i,j}}{(x^2+b_ix+c_i)^j}$ comme dans le théorème.

Utilisons un exemple pour montrer quelques méthodes qui permettent de calculer simplement les coefficients. Soit

$$f(x) = (3x^7+9x+1)/(3x^6-6x^3+3)\ .$$

– la division euclidienne du numérateur par le dénominateur donne pour quotient $E(x)=x$ et pour reste $R(x)=6x^4+6x+1$. D'autre part $Q(x)$ se factorise en $Q(x)=3(x-1)^2(x^2+x+1)^2$ (vu auparavant). On a donc

$$f(x) = x+\frac{6x^4+6x+1}{3(x-1)^2(x^2+x+1)^2}$$

avec

$$\begin{aligned} g(x) &= \frac{6x^4+6x+1}{3(x-1)^2(x^2+x+1)^2} \\ &= \frac{a_1}{(x-1)}+\frac{a_2}{(x-1)^2}+\frac{b_1x+c_1}{(x^2+x+1)}+\frac{b_2x+c_2}{(x^2+x+1)^2} \end{aligned}$$

Nous avons donc 6 coefficients à calculer. La méthode consistant à procéder par identification ou à donner des valeurs particulières à x nous donne un système linéaire de 6 équations à 6 inconnues à résoudre ce qui est très fastidieux. Essayons donc d'être plus astucieux (ceci est valable dans le cas général !).

- Pour calculer a_2 : on multiplie les deux membres de l'inégalité par $(x-1)^2$ et on donne à x la valeur 1 ; on obtient $a_2 = 13/27$.
- Le calcul de (b_2, c_2) peut se faire en utilisant une méthode analogue quoique plus technique puisque utilisant les nombres complexes : on multiplie les deux membres par $(x^2+x+1)^2$ et on donne à x la valeur d'une des racines dans $\mathbf{C}$ de x^2+x+1. En effet ce trinôme dans $\mathbf{C}$ admet deux racines non-réelles *conjuguées* dont la somme vaut 1 et le produit -1. Soient x_0 et $\overline{x_0}$ ces deux racines. On obtient alors $(6x_0^4+6x_0+1)(3(x_0-1)^2) = b_2x_0+c_2$. Le membre de gauche peut s'exprimer sous la forme ax_0+c_2 uniquement en connaissant la somme et le produit des racines x_0 et $\overline{x_0}$. Ici on trouve
$$\frac{6x_0^4+6x_0+1}{3(x_0-1)^2} = -\frac{11}{9}+\frac{1}{9}x_0 \ ,$$
d'où $b_2 = 1/9$ et $c_2 = -11/9$ (en utilisant le fait que $(1, x_0)$ est une base de l'espace vectoriel $\mathbf{C}$ sur $\mathbf{R}$).
- Pour trouver les trois coefficients restants, on peut donner à x trois valeurs particulières et on obtient un système linéaire dont la solution est le triplet (a_1, b_1, c_1). Il est souvent plus rapide de procéder comme suit.
On calcule
$$h(x) = g(x) - \frac{b_2x+c_2}{(x^2+x+1)^2} = \frac{18x^2-19x+14}{9(x-1)^2(x^2+x+1)}.$$
En vertu de l'unicité, la décomposition de cette fraction rationnelle en éléments simples est
$$h(x) = \frac{a_1}{(x-1)}+\frac{a_2}{(x-1)^2}+\frac{b_1x+c_1}{(x^2+x+1)}$$
On multiplie les deux membres par x^2+x+1 et on donne à x la valeur x_0 ; on trouve $b_1 = -4/27$ et $c_1 = 33/27$.
- Il ne reste plus qu'à calculer a_1. On donne à x une valeur particulière, par exemple 0. On peut également (et c'est plus astucieux) multiplier les deux membres donnant $h(x)$ par x et de faire tendre x vers $+\infty$. On obtient $0 = a_1 + b_1$ soit $a_1 = \frac{4}{27}$.
- et donc

$$\begin{aligned} f(x) &= \frac{3x^7+9x+1}{3x^6-6x^3+3} \\ &= x+\frac{4}{27(x-1)}+\frac{13}{27(x-1)^2}+\frac{-4x+33}{27(x^2+x+1)}+\frac{x-11}{9(x^2+x+1)^2} \end{aligned}$$

Calcul de primitives de fractions rationnelles réelles.

Cas particulier de

$$\int \frac{P'(x)}{(P(x))^n}dx$$

où P est un polynôme et n un entier naturel.

- Si $n = 1$, on a $\int \frac{P'(x)}{P(x)}dx = \ln|P(x)|$ sur tout intervalle sur lequel P ne s'annule pas.
- Si $n > 1$, on a $\int \frac{P'(x)}{(P(x))^n}dx = \frac{1}{1-n}\frac{1}{(P(x))^{n-1}}$ sur tout intervalle sur lequel P ne s'annule pas.

Cas général

a. On décompose la fraction rationnelle en éléments simples dans $\mathbf{R}(X)$.

b. On est donc amené à chercher une primitive de chaque terme de la décomposition :
   - Pour la *partie entière* qui est un polynôme, une primitive est obtenue immédiatement.
   - Pour les *éléments simples de première espèce* :
     $\int \frac{1}{x-a}dx = \ln|x-a|$
     $\int \frac{1}{(x-a)^n}dx = \frac{1}{1-n}\frac{1}{(x-a)^{n-1}} \quad n \in \mathbf{N}, n \geq 2.$
   - Pour les *éléments simples de deuxième espèce* :
     $\frac{ax+b}{(x^2+px+q)^n} = \frac{a}{2}\frac{2x+p}{(x^2+px+q)^n} + \left(b - \frac{ap}{2}\right)\frac{1}{(x^2+px+q)^n}$
     Le premier terme est de la forme $\frac{a}{2}\frac{P'(x)}{(P(x))^n}$ dont on obtient immédiatement une primitive (cas particulier déjà vu).
     Il reste donc $\int \frac{1}{(x^2+px+q)^n}dx$. On met le trinôme sous forme canonique et après une changement de variable affine on est ramené à $J_n = \int \frac{1}{(t^2+1)^n}dt, n \in \mathbf{N}$.
     Par récurrence :
     - $J_1 = \arctan t$
     - $J_n = \int \frac{1+t^2}{(1+t^2)^{n+1}}dt = J_{n+1} + \int \frac{t^2}{(1+t^2)^{n+1}}dt$
       En intégrant par parties : $u'(t) = \frac{t}{(1+t^2)^{n+1}}$ $\quad u(t) = -\frac{1}{2n}\frac{1}{(1+t^2)^n}$
       $v(t) = t$ $\quad v'(t) = 1$
       $J_n = J_{n+1} - \frac{1}{2n}\frac{t}{(1+t^2)^n} + \frac{1}{2n}J_n$ d'où à une constante près :

$$J_{n+1} = \frac{1}{2n}\frac{t}{(1+t^2)^n} + \frac{2n-1}{2n}J_n$$

15.3.b Fractions rationnelles en les fonctions trigonométriques

Le but est de calculer la primitive $\int F(\sin x, \cos x)dx$ où $F(x, y)$ est une fraction rationnelle en les deux variables x et y (c'est-à-dire le quotient de deux polynômes à deux variables x et y).

Le changement de variable $t = \tan\frac{x}{2}$, pour $x \neq (2k+1)\pi$, ramène le calcul de cette primitive au calcul d'une primitive d'une fraction rationnelle en t.

Plus précisément et pour simplifier les calculs, si l'élément différentiel $F(\sin x, \cos x)dx$ est invariant par le changement de variable :

$\left.\begin{array}{l} u = -x, \text{ alors le changement de variable } t = \cos x \\ u = \pi - x, \text{ alors le changement de variable } t = \sin x \\ u = \pi + x, \text{ alors le changement de variable } t = \tan x \end{array}\right\}$ ramène le calcul au calcul d'une primitive d'une fraction rationnelle en t.

Les trois dernières assertions sont communément appelées *Règles de Bioche*.

Montrons la première assertion à titre d'exemple : Si $t = \tan\frac{x}{2}$, on a $\cos x = \frac{1-t^2}{1+t^2}$ et $\sin x = \frac{2t}{1+t^2}$. De plus $dt = \frac{1}{2}(1+t^2)dx$ et donc $dx = \frac{2}{1+t^2}dt$. Par conséquent :

$$\int F(\sin x, \cos x)dx = \int F\left(\frac{2t}{1+t^2}, \frac{1-t^2}{1+t^2}\right)\frac{2}{1+t^2}dt$$

qui est une primitive d'une fraction rationnelle en t.

15.3.c Fractions rationnelles en la fonction exponentielle

Le but est de calculer une primitive de la forme $\int f(e^x)dx$, ou $\int F(\operatorname{sh} x, \operatorname{ch} x)dx$, où $f(x)$ est une fraction rationnelle en x, et $F(x,y)$ est une fraction rationnelle en x et en y.

Le changement de variable $u = e^x$ ramène le calcul de ces primitives au calcul d'une primitive d'une fraction rationnelle en u.

Démonstration : En posant $u = e^x$, on a :

$$\int f(e^x)dx = \int \frac{f(u)}{u}du.$$

et

$$\int F(\operatorname{sh} x, \operatorname{ch} x)dx = \int F\left(\frac{u-\frac{1}{u}}{2}, \frac{u+\frac{1}{u}}{2}\right)\frac{du}{u}\ ,$$

qui sont bien des primitives de fractions rationnelles.

15.3.d Intégrales abéliennes

Proposition. Pour calculer $\int F\left(x, \sqrt[n]{\frac{ax+b}{cx+d}}\right)dx$ avec F fraction rationnelle à deux variables, n entier supérieur à 2 et $ad-bc \neq 0$, on pose $t = \sqrt[n]{\frac{ax+b}{cx+d}}$ et le changement de variable ramène le calcul à celui d'une primitive d'une fraction rationnelle en t.

Démonstration. On pose $t = \sqrt[n]{\frac{ax+b}{cx+d}}$, d'où $x = \frac{b-dt^n}{ct^n-a}$ et donc $dx = \frac{(ad-bc)nt^{n-1}}{(ct^n-a)^2}dt$.
On a alors $\int F\left(x, \sqrt[n]{\frac{ax+b}{cx+d}}\right)dx = \int F\left(\frac{b-dt^n}{ct^n-a}, t\right)\frac{(ad-bc)nt^{n-1}}{(ct^n-a)^2}dt$, ce qui est bien une primitive de fraction rationnelle en t.

Proposition. Pour calculer $\int F\left(x, \sqrt{ax^2+bx+c}\right)dx$ avec F fraction rationnelle à deux variables, on peut grâce à un changement de variable $t = h(x)$ ramener le calcul à celui d'une primitive d'une fraction rationnelle en t.

Démonstration succincte : Il suffit de paramétrer la conique $\Gamma = \{(x,y) \in \mathbf{R}^2;\ ax^2+bx+c-y^2 = 0\}$ à l'aide de fonctions trigonométriques ou hyperboliques.

- 1er cas : $a > 0$, Γ est une hyperbole, on discute selon la réalité des racines.
  - Le trinôme ax^2+bx+c a deux racines réelles. Par un changement de variable, on peut alors se ramener à u^2-1 puis on pose $u = \varepsilon\operatorname{ch} t$ avec $t \in \mathbf{R}^+$ et $\varepsilon \in \{-1,1\}$.
  - Le trinôme ax^2+bx+c n'a pas de racine réelle. Par un changement de variable, on se ramène à u^2+1 puis on pose $u = \operatorname{sh} t$, $t \in \mathbf{R}$.
    Remarque : Le cas où le trinôme admet une racine double est trivial puisqu'alors l'expression est directement simplifiable et le radical disparaît.

– 2ème cas : $a < 0$, Γ est une ellipse.
Le trinôme a nécessairement deux racines réelles distinctes α et β pour que la fonction $x \longmapsto \sqrt{ax^2+bx+c}$ soit définie sur un intervalle de $\mathbf{R}$ d'intérieur non-vide.
Par un changement de variable on se ramène d'abord à $1-u^2$ puis on pose $u = \sin t,\ t \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, ou bien $u = \cos t,\ t \in [0, \pi]$

Exemple. Afin de bien comprendre comment on obtient l'une des trois formes annoncées dans la preuve, un exemple est plus parlant qu'une démonstration formelle. On remarquera l'utilisation essentielle de la forme canonique du trinôme. Calcul de

$$\int \sqrt{2x^2-6x+4}dx\ .$$

On remarque ici que le trinôme $2x^2-6x+4$ a deux racines distinctes 1 et 2 et que donc la fonction $x \longmapsto \sqrt{2x^2-6x+4}$ n'est définie que pour $x \in]-\infty, 1] \cup [2, +\infty[$.

$$\begin{aligned}\int \sqrt{2x^2-6x+4}dx &= \sqrt{2}\int \sqrt{x^2-3x+2}dx \\ &= \sqrt{2}\int \sqrt{\left(x-\tfrac{3}{2}\right)^2-\tfrac{1}{4}}dx \\ &= \tfrac{1}{\sqrt{2}}\int \sqrt{(2x-3)^2-1}dx\end{aligned}$$

En posant $u = 2x-3$, on obtient alors $\int \sqrt{2x^2-6x+4}dx = \frac{1}{2\sqrt{2}}\int \sqrt{u^2-1}du$ qui est bien la forme annoncée dans la proposition.

Puis on pose $u = \begin{cases} \operatorname{ch} t, \text{ si } u \geq 0 \\ -\operatorname{ch} t, \text{ si } u \leq 0 \end{cases}$
Soit $\varepsilon = \pm 1$ selon que $u \geq 0$ ou $u \leq 0$ i.e. $x \geq 2$ ou $x \leq 1$. On a alors :

$$\begin{aligned}\int \sqrt{2x^2-6x+4}dx &= \tfrac{\varepsilon}{2\sqrt{2}}\int \operatorname{sh}^2 t dt \\ &= \tfrac{\varepsilon}{4\sqrt{2}}\int (\operatorname{ch} 2t - 1)\, dt \\ &= \tfrac{\varepsilon}{4\sqrt{2}}\left(\tfrac{1}{2}\operatorname{sh} 2t - t\right)\end{aligned}$$

Il reste alors à remplacer t par sa valeur en fonction de x, soit $t = \operatorname{argch} \varepsilon(2x-3)$.

Remarque : on peut directement poser $2x-3 = \begin{cases} \operatorname{ch} t, \text{ si } 2x-3 \geq 0 \text{ i.e. si } x \geq 2 \\ -\operatorname{ch} t, \text{ si } 2x-3 \leq 0 \text{ i.e. si } x \leq 1 \end{cases}$ ce qui raccourcit un peu les calculs.

15.3.e Intégrales définies

Soit f une fonction définie sur l'intervalle $[a, b]$ et admettant une primitive F définie sur $[a, b]$. On appelle intégrale définie de f sur l'intervalle $[a, b]$, le nombre réel :

$$\int_a^b f(x)dx = F(b) - F(a),$$

c'est à dire la valeur en b de l'unique primitive de f qui s'annule en a.

Le calcul d'une intégrale se ramène à celui d'une primitive quelconque de la fonction f (la quantité $F(b) - F(a)$ ne dépend pas de la primitive choisie).

Exemple. $\int_0^1 \frac{dx}{1+x^2} = \arctan 1 - \arctan 0 = \frac{\pi}{4}$.

Annexe A

Algèbre linéaire

Combinaisons linéaires. Commençons par deux exemples. Il vous est certainement arrivé lors d'une *chasse au trésor* de devoir suivre des indications comme "40 pas direction Sud, 10 pas direction Nord-Est, 5 pas direction Est, vous y êtes !". Il est clair qu'avec de telles indications on peut diriger une personne de n'importe quel endroit de la surface de la Terre à n'importe quel autre. De plus, si il n'y avait pas d'obstacles, deux directions suffiraient : par exemple Nord et Est, Sud et Sud-Est (mais pas Nord et Sud). En termes de Nord et de Est, 1 pas dans la direction Nord-Est serait remplacé dans les instructions par un certain nombre de pas dans la direction Nord *et* un certain nombre de pas dans la direction Est (ce n'est *pas* un nombre entier de pas ; combien de pas faut-il ?). Ce que nous faisons ici est manipuler des combinaisons linéaires de pas dans les directions du plan : en notation symbolique l'instruction plus haut peut s'écrire

$$40 \cdot \vec{S} + 10 \cdot \vec{NE} + 5 \cdot \vec{E} = (40 - 10\frac{1}{\sqrt{2}}) \cdot \vec{S} + (5 + 10\frac{1}{\sqrt{2}}) \cdot \vec{E} \ .$$

Une autre situation où l'on peut sommer et amplifier des objets d'une même nature est celle des *solutions d'équations différentielles linéaires* comme $y'' + by' + cy = 0$, que nous avons rencontré en 12.6. En effet si f_1 et f_2 sont des fonctions satisfaisant cette équation et a est un nombre réel, alors la somme $f_1 + f_2$ et la fonction af_1 en sont aussi solution : ceci découle du fait que $(f_1 + f_2)' = f_1^{'} + f_2^{'}$ et $(af_1)' = af_1^{'}$.

De manière imprécise étant donné des objets mathématiques f_1, ... , f_r, une *combinaison linéaire* de ces éléments à coefficients dans un ensemble de nombres R est une *somme de multiples* par des éléments de R de ces éléments :

$$a_1 f_1 + \cdots + a_r f_r \quad \text{avec } a_i \in R \ . \qquad (CL)$$

Espaces vectoriels. Pour donner un sens précis à la notion de combinaison linéaire on introduit le concept d'espace vectoriel :

> *un espace vectoriel (sur R) est un ensemble dans lequel la notion de combinaison linéaire (à coefficients dans R) a un sens.*

Pour qu'une expression comme (CL) ci-dessus ait un sens il faut disposer sur l'ensemble contenant les f_i d'une "opération somme" et d'une "opération produit-par-une-constante". Pour la définition de la notion d'espace vectoriel il faut donc spécifier l'ensemble R des constantes. Dans les exemples de la chasse au trésor et des solutions de l'équation différentielle considérée on s'autorise à multiplier par des nombres réels, mais en général il est utile de considérer d'autres ensembles de constantes : l'important

est que ces ensembles de constantes aient un certain nombre de propriétés algébriques (en particulier la propriété du sup n'intervient pas ici). On peut faire l'étude des combinaisons linéaires à coefficients entiers, mais déjà l'exemple de la chasse au trésor montre qu'il serait plus compliqué : comme noté $\vec{NE}$ n'est pas une combinaison linéaire de $\vec{N}$ et de $\vec{E}$ à coefficients entiers.

On demande que l'ensemble des constantes soit un corps.

Des exemples de corps R sont les ensembles de nombres tels que $\mathbf{Q}$, $\mathbf{R}$ ou $\mathbf{C}$. L'ensemble $\mathbf{Z}$ des entiers relatifs n'est pas un corps. Un autre exemple de corps est le corps $\mathbf{F}_2$ à deux éléments : c'est l'ensemble contenant deux éléments 0 et 1 muni de la somme + et du produit · définis par les tables

+	0	1
0	0	1
1	1	0

.	0	1
0	0	0
1	0	1

S'il vous paraît surprenant de considérer un ensemble de nombres dans lequel $1+1=0$ pensez à la règle "impair plus impair est pair" (1 = impair et 0 = pair). Le corps $\mathbf{F}_2$ est le plus petit corps : un corps R doit toujours contenir au moins deux éléments 0_R et 1_R, qui ont les propriétés caractéristiques... Les espaces vectoriels sur le corps $\mathbf{F}_2$ sont très utilisés dans la théorie des codes correcteurs d'erreurs, qui est appliquée par exemple dans le codage des informations audio sur les CD (compact-disks).

Plus généralement un corps est un ensemble muni de deux opérations "somme" et "produit", qui satisfont aux propriétés usuelles... Ce qui est important, et qui explique que l'algèbre linéaire que nous allons développer ne marche pas avec $R=\mathbf{Z}$ est que

dans un corps tout élément différent de 0 *admet un inverse pour le produit.*

Définissons formellement ce que l'on entend par un *espace vectoriel sur un corps* R. Il s'agit d'un ensemble V muni d'une opération "somme"

$$+ : V \times V \to V$$

et sur lequel R agit par "multiplication"

$$\cdot : R \times V \to V \ .$$

L'opération + est donc une opération interne : à deux éléments de V elle associe un élément de V. Par contre l'opération · est une opération externe : R n'est pas considéré comme un sous-ensemble de V. D'ailleurs l'opération somme dans V n'est pas forcément définie en termes de la somme dans R. D'habitude on ne note pas le · et on écrit av pour $a \cdot v$. Ces opérations doivent satisfaire une liste de propriétés (axiomes). Les quatre premières ne concernent que la somme dans V. Les quatre autres demandent que les deux opérations sur V et les deux opérations dans le corps R soient compatibles. Soit u, v et w éléments de V et a, b éléments de R, alors on demande :

1) (associativité) $(u+v)+w=u+(v+w)$
2) (existence d'un élément neutre) il existe un élément dans V, noté 0 (ou 0_V) tel que $0+v=v+0=v$
3) (existence d'un inverse pour la somme) il existe v' dans V tel $v'+v=v'+v=0$
4) (commutativité) $v+w=w+v$
5) $a(v+w)=av+aw$
6) $(a+b)v=av+bv$
7) $a(bv)=(ab)v$
8) $1_R v=v$

De ces axiomes on tire facilement les conséquences suivantes :

- *Unicité de* 0 : si $0'$ satisfait le (2), alors $0=0+0'=0'$.

- *Propriété de simplification* : si $v + w = v + w'$, alors $w = w'$. En effet, d'après (3) il existe v' tel que $v' + v = 0$, donc en utilisant (1) et (2) on obtient $v' + (v + w) = (v' + v) + w = 0 + w = w$ et $v' + (v + w') = (v' + v) + w' = 0 + w' = w'$, d'où $w = w'$.
- *Unicité de l'inverse* : si w et w' sont tels que $v + w = v + w' = 0$, alors par la propriété de simplification $w = w'$. Souvent on note l'inverse $-v$, d'ailleurs $(-1)v$ est l'inverse de v (voir ci-après ; ici (-1) est l'inverse de 1 dans R, c'est-à-dire le nombre a tel que $a + 1 = 0_R$).
- $0_R \cdot v = 0_V$: utilisons (8) et (6) et écrivons $0v + v = 0v + 1v = (0 + 1)v = 1v = v$, alors le résultat suit par simplification.
- $(-1)v$ *est l'inverse de* v : en effet $(-1)v + v = (-1)v + 1v = ((-1) + 1)v = 0v = 0$.

Exemples. 1) Étant donné un corps R et un entier naturel n on munit l'ensemble R^n des n-uplets d'éléments de R d'une structure d'espace vectoriel sur R en posant :

$$(a_1, \ldots, a_n) + (b_1, \ldots, b_n) = (a_1 + b_1, \ldots, a_n + b_n)$$

et

$$a \cdot (a_1, \ldots, a_n) = (aa_1, \ldots, aa_n) \ .$$

2) Soit I un ensemble et soit $\mathcal{F}(I, R)$ l'ensemble des fonctions $f : I \to R$. Cet ensemble peut être muni d'une structure d'espace vectoriel sur R en utilisant les opérations de R :

$$(f + g)(x) = f(x) + g(x)$$

et

$$(af)(x) = af(x) \ .$$

Ceci généralise le cas bien connu des fonctions réelles ou complexes (I un intervalle). En fait, si l'on devait définir R^n on ne le définirait pas comme un produit cartésien itéré, mais plutôt comme étant $\mathcal{F}(n, R)$ où n représente l'entier naturel n (qui est un ensemble à n éléments !). Donc on peut penser à I comme à un ensemble d'indices.

3) Soit R un corps. L'ensemble R^2 muni des opérations

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 - b_2)$$

et

$$a(a_1, a_2) = (aa_1, aa_2)$$

n'est pas un espace vectoriel sur R. De même R^2 muni des opérations

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2)$$

et

$$a(a_1, a_2) = (aa_1, 0)$$

n'est pas non plus un espace vectoriel sur R (dans ce dernier exemple seul l'axiome (8) n'est pas vérifié, ce qui montre que cet axiome n'est pas conséquence des autres).

4) Pour un sous-ensemble fini $S = \{f_1, \ldots, f_r\}$ d'un espace vectoriel V sur R on obtient une fonction

$$\begin{array}{rcl} F_S : R^r & \to & V \\ (a_1, \ldots, a_r) & \mapsto & a_1 f_1 + \cdots + a_r f_r \end{array}$$

Une grande partie de ce qui suit sera consacré à expliquer le fait, que dans tout espace vectoriel V sur un corps R on peut choisir un sous-ensemble S tel que la fonction F_S est bijective.

Sous-espaces vectoriels. Soit R un corps et soit V un espace vectoriel sur R. Considérons W un sous-ensemble non-vide de V. On peut se demander si la somme (dans V) de deux éléments de W est encore dans W et de même si la multiplication d'un élément de W par un élément de R, qui a priori est dans V, est encore dans W. On obtiendrait alors deux opérations sur W. Si ces opérations font de W un espace vectoriel sur R, on dit que W est un *sous-espace vectoriel* de V.

On peut aussi caractériser les sous-espaces d'un espace vectoriel V comme suit : un sous-ensemble W de V est un sous-espace si et seulement si

a) $W \neq \emptyset$

b) $(\forall w, w' \in W) \Rightarrow w +_V w' \in W$

c) $(w \in W \wedge a \in R) \Rightarrow a \cdot_V w \in W$

On peut aussi remplacer (a) par

a') $0_V \in W$

Exemples. 1) Soit $I \subset \mathbf{R}$ un intervalle, alors l'ensemble $\mathcal{C}^0(I, \mathbf{R})$ des fonctions $f : I \to \mathbf{R}$ partout *continues* sur I est un sous-espace vectoriel de $\mathcal{F}(I, \mathbf{R})$. De même pour l'ensemble $\mathcal{C}^r(I, \mathbf{R})$ des fonctions r fois dérivables sur I et dont la r-ième à dérivée est continue ou pour le sous-ensemble de $\mathcal{C}^2(I, \mathbf{R})$ des solutions de l'équation différentielle $y'' + by' + cy = 0$ en est un sous-espace vectoriel. Ces affirmations découlent des propriétés, que nous avons montrées, des fonctions continues et dérivables.

2) Soit R un corps et soit V un espace vectoriel sur R. Soit f_1 , ... , f_r des éléments de V. L'ensemble des combinaisons linéaires des f_i à coefficients dans R est un sous-espace vectoriel de V, noté

$$< \{f_1, \ldots, f_r\} > \quad \text{ou} \quad < f_1, \ldots, f_r > \ .$$

En fait il n'est nul besoin de se restreindre à un ensemble fini d'éléments de V. Si S est un sous-ensemble quelconque de V alors l'ensemble de toutes les combinaisons linéaires *finies* des éléments de S à coefficients dans R est un sous-espace vectoriel de V, noté $< S >$. On a donc :

$$< S >= \{v \in V : \exists r \in \mathbf{N}\ \exists a_1, \ldots, a_r\ \exists f_1, \ldots, f_r\ (v = a_1 f_1 + \cdots + a_r f_r)\} \ .$$

Le sous-espace $< S >$ est appelé le *sous-espace de V engendré par S*.

Si $V =< S >$ on dit que S *engendre* V, ou S est un *système de générateurs* de V, ou encore une *famille génératrice* de V.

Dire que S engendre V revient à dire que la fonction

$$\begin{array}{rcl} F_S : R^r & \to & V \\ (a_1, \ldots, a_r) & \mapsto & a_1 f_1 + \cdots + a_r f_r \end{array}$$

est surjective.

3) Tout sous-ensemble d'un espace vectoriel qui contient un ensemble de générateurs de l'espace est un ensemble de générateurs. D'ailleurs $V =< V >$.

4) L'intersection de sous-espaces d'un même espace vectoriel est encore un sous-espace. Par contre, en général, la réunion de deux sous-espaces n'est pas un sous-espace. Si U et W sont sous-espaces d'un espace vectoriel V, alors la *somme* de U et W est définie par

$$U + W :=< U \cup W > \ .$$

Il s'agit du plus petit sous-espace de V qui contient U et W.

5) Simplement en utilisant les axiomes de définition, on peut essayer de se faire une image de ce qu'est un espace vectoriel. Par exemple on peut faire *la liste des sous-espaces vectoriels de R^n* pour n petit : noter d'abord qu'un espace vectoriel V est toujours non-vide car un des axiomes demande l'existence d'un "zéro" 0_V dans V ; il se peut que le zéro soit le seul élément du sous-espace, on a

alors affaire avec l'espace nul ; si x est un élément non-nul du sous-espace, alors par les axiomes, le sous-espace doit contenir tous les multiples $\lambda \cdot x$ avec λ dans R, c'est-à-dire que le sous-espace contient la "droite" passant par 0 et x ; si le sous-espace contient un élément y qui n'est pas multiple de x, c'est-à-dire qui n'est pas sur la droite que nous venons de considérer, alors le sous-espace doit contenir toutes les expressions de la forme $\lambda \cdot x + \mu \cdot y$: c'est le "plan" passant par 0, x et y ; etc.

Bases, dimension. La droite des multiples λx de x ressemble beaucoup à R. Un plan ressemble beaucoup à R^2 : une fois choisis x et y dans le plan et pas sur la même droite on obtient tous les autres points du plan sous la forme $\lambda \cdot x + \mu \cdot y$ avec λ et μ dans R *uniques*. En fait la situation générale n'est guère plus compliquée :

> *pour tout espace vectoriel V sur R on arrive à trouver un sous-ensemble S de V tel que tout élément v de V s'écrive de façon unique comme combinaison linéaire à coefficients dans R des éléments de S ; donc non seulement $V =< S >$, mais l'écriture d'un élément de V comme combinaison linéaire des éléments de S est unique :*
>
> $$\forall v \in V \ \exists r \in \mathbf{N} \ \exists!(s_1, \ldots, s_r) \in S^r \ , \exists!(\lambda_1, \cdots, \lambda_r) \in R^r : v = \lambda_1 s_1 + \cdots + \lambda_r s_r \ .$$

Un tel sous-ensemble S est appelé une *base* de V sur R. Nous allons montrer plus bas comment trouver une base S sous l'hypothèse qu'il existe dans V un sous-ensemble fini S' avec $V =< S' >$, c'est-à-dire tel que tout élément de V est combinaison linéaire à coefficients dans R des éléments de S'. Dans ce cas S sera fini. Noter qu'un tel S' existe pour tout sous-espace vectoriel d'un R^N. On peut montrer qu'en fait tout espace vectoriel possède une base et que

> *le nombre d'éléments de n'importe quelle base est indépendant du choix de la base.*

On voit alors que l'existence d'une base dans un espace vectoriel permet de décrire tout élément de l'espace en termes d'un certain nombre de constantes et que le nombre d'éléments d'une base représente les *degrés de libertés* dont on dispose dans le choix des constantes, pour déterminer un élément de l'espace vectoriel

Donnons une définition formelle de ce qu'est une base d'un espace vectoriel V sur R. Un sous-ensemble S de V est appelé une *famille libre* (de V sur R), si toute écriture

$$\lambda_1 s_1 + \cdots + \lambda_r s_r = 0_V$$

avec les λ_i dans R entraîne $\lambda_1 = \lambda_2 = \cdots = \lambda_r = 0$. Un S libre est aussi appelé une *famille linéairement indépendante*. Une *base* de V (sur R) est un sous-ensemble S de V qui engendre V (sur R) et qui est une famille libre (de V sur R). Noter que si une famille S est libre, alors *toute* écriture $v = \lambda_1 s_1 + \cdots + \lambda_r s_r$ est unique : c'est la définition pour $v = 0_V$, pour v quelconque prendre deux écritures et les soustraire...

> *Un sous-ensemble $S = \{f_1, \ldots, f_r\}$ est une base de V si et seulement si la fonction*
>
> $$\begin{array}{rcl} F_S : R^r & \to & V \\ (a_1, \ldots, a_r) & \mapsto & a_1 f_1 + \cdots + a_r f_r \end{array}$$
>
> *est une bijection : elle est surjective ssi S engendre et elle est injective ssi S est libre.*

Exemples. 1) Une base de R^n est donnée par les n éléments $(1, 0, \ldots, 0)$, $(0, 1, \ldots, 0)$, ... , $(0, 0, \ldots, 1)$. Cette base est appelée la *base canonique* de R^n. Se souvenir qu'il y en a plein d'autres (mais elles ont toutes cardinal n) !

2) Toute solution dans $\mathcal{C}^2(I, \mathbf{R})$ de l'équation différentielle

$$y'' + by' + cy = 0$$

s'écrit sous la forme

$$y = Ate^{r_1 t} + Be^{r_1 t} \quad \text{ou} \quad y = Ae^{r_2 t} + Be^{r_1 t}$$

suivant que les racines r_1 et r_2 de l'équation algébrique $r^2 + br + c = 0$ coïncident ou pas. Ici A et B sont des constantes réelles. On montre assez facilement que les ensembles de fonctions $\{te^{r_1 t}, e^{r_1 t}\}$ et $\{e^{r_2 t}, e^{r_1 t}\}$ sont des familles libres. On voit donc qu'il s'agit de bases de l'ensemble des solutions de l'équation différentielle en question (se souvenir que l'exponentielle complexe d'un imaginaire pur défini les fonctions trigonométriques).

3) Tout sous-ensemble d'un ensemble libre est libre.

Exercice. Déterminer tous les sous-espaces vectoriels de l'espace vectoriel sur le corps à deux éléments $V = \mathbf{F}_2^3$ (noter que cet espace est un ensemble à 8 éléments). Donner une base pour chacun de ces sous-espaces. (Si vous avez des difficultés à faire cet exercice à partir des définitions essayez de le faire après avoir lu ce qui suit.)

Voici l'énoncé assurant l'existence des bases et l'unicité de leur cardinal :

Théorème-Définition. *Soit V un espace vectoriel sur R et S un sous-ensemble fini de V qui engendre V sur R. Alors V possède une base avec un nombre fini d'éléments. Plus précisément, on peut trouver un sous-ensemble de S qui est une base de V. Le nombre d'éléments d'une base quelconque de V est le même, ce nombre est appelé la* dimension *de V. Il est noté :* $\dim_R(V)$.

Pour démontrer l'existence d'une base on utilise le résultat suivant.

Lemme. *Si L est une famille libre dans V et si G est un sous-ensemble fini de V qui l'engendre avec*

$$L \subset G \ ,$$

alors il existe une base B de V avec $L \subset B \subset G$.

Si L engendre V on prend $B = L$. Sinon il existe g_1 dans G avec $g_1 \notin < L >$: posons $L_1 = L \cup \{g_1\}$. Alors L est sous-ensemble strict de L et L_1 est encore libre (!). De plus $L_1 \subset G$. Si L_1 engendre on pose $B = L_1$, sinon on obtient L_2 libre avec L_1 comme sous-ensemble strict et $L_2 \subset G$. Cette construction doit s'arrêter car G est fini...

Exercice. Détailler cette démonstration et identifier le passage où l'on utilise le fait que tout élément non-nul du corps R admet un inverse multiplicatif.

Pour démontrer l'unicité du cardinal des bases de l'espace vectoriel V, on montre que si L et G sont comme dans le lemme, *mais avec L non forcément contenu dans G*, alors

$$\mathrm{card}(L) \leq \mathrm{card}(G) \ .$$

Ceci entraîne l'unicité car si B et B' sont des bases en appliquant l'inégalité précédente aux couples $(L, G) = (B, B')$ et $(L, G) = (B', B)$ on obtient bien $\mathrm{card}(B) = \mathrm{card}(B')$. Pour montrer l'inégalité on utilise le résultat suivant.

Lemme d'échange. *Avec les notations précédentes, supposons que*

$$L = \{l_1, \ldots, l_n\} \quad \textit{et} \quad G = \{g_1, \ldots, g_p\} \ .$$

Alors, on peut remplacer n des éléments de G par les éléments de L de façon à ce que l'ensemble obtenu engendre encore.

Vu que L est libre l_1 est non-nul. Vu que G engendre on peut écrire

$$l_1 = \sum_j \lambda_j g_j \ ,$$

et il existe un j tel que λ_j soit différent de 0. Alors

$$g_j = \frac{-1}{\lambda_j}(\sum_{i \neq j} \lambda_i g_i - l_1) \ .$$

Posons $G_1 = \{g_1, \ldots, g_{j-1}, l_1, g_{j+1}, \ldots, g_p\}$. On voit que G_1 engendre encore. Donc

$$l_2 = \mu l_1 + \sum_{i \neq j} \mu_i g_i \ ,$$

et il existe k tel que μ_k soit différent de 0. Sinon on aurait $l_2 = \mu l_1$, qui contredit le fait que L est libre. On obtient ainsi

$$G_2 = \{g_1, \ldots, g_{j-1}, l_1, g_{j+1}, \ldots, g_{k-1}, l_2, g_{k+1}, \ldots g_p\}$$

(disons), qui engendre encore. On réitère...

Applications linéaires et matrices.

Une application linéaire est une fonction entre espaces vectoriels, qui préserve les combinaisons linéaires.

Plus formellement, soit V et W des espaces vectoriels sur le corps R. Une fonction $f : V \to W$ est dite *R-linéaire* si pour tout v_1, v_2 dans V et pour tout a dans R

$$f(av_1 + v_2) = af(v_1) + f(v_2) \ .$$

Exemples. 1) Les transformations géométriques du plan suivantes sont linéaires : la symétrie de centre l'origine et toutes les rotations et homothéties qui préservent l'origine. Les translations ne sont pas linéaires.

2) La dérivée et l'intégrale définissent des applications linéaires.

3) L'image (resp. la fibre en 0_W) d'une application linéaire $f : V \to W$ définit un sous-espace de W (resp. V). On pose

$$\ker(f) := f^{-1}(0_W) = \{v \in V : f(v) = 0_W\} \ .$$

Ce sous-espace de V est appelé le *noyau* de f (en Allemand ou en Anglais *kernel*).

Lemme. *Soit $f : V \to W$ une application linéaire. Alors $f(0_V) = 0_W$ et f est injective si et seulement si* $\ker(f) = 0_V$.

En effet pour tout v dans V on a $f(0_V) = f(v + (-v)) = f(v) + (-1)f(v) = 0_W$. Si f est injective alors toute fibre de f contient au plus un élément et comme on vient de le voir la fibre de 0_W contient au moins 0_V. Soit $\ker(f) = 0_V$ et supposons que $f(v) = f(v')$, alors $f(v - v') = 0_W$ et donc $v - v' \in \ker(f) = 0_V$, c'est-à-dire $v = v'$.

Choix de bases. Si $S = \{f_1, \ldots, f_r\}$ est une base de V, alors nous avons vu que la fonction

$$\begin{array}{rcl} F_S : R^r & \to & V \\ (a_1, \ldots, a_r) & \mapsto & a_1 f_1 + \cdots + a_r f_r \end{array}$$

est une bijection.

Il est facile de voir que F_S est une application linéaire.

Une application linéaire bijective est souvent appelée un *isomorphisme* : une telle application exhibe le fait que le domaine et l'image se ressemblent beaucoup, ils ont la même structure d'espace vectoriel, la même forme ("iso = même", "morphie = forme"). (Souvent les applications linéaires sont aussi appelées des *morphismes*. S'il s'agit d'applications d'un espace dans lui-même on parle d'*endomorphisme* et un endomorphisme bijectif est appelé un *automorphisme*.)

Réciproquement, un isomorphisme $f : R^r \to V$ donne une base de V : il suffit de considérer l'image des éléments de la base canonique.

Soit V et W des espaces vectoriels sur le corps R. Montrons que

une application R-linéaire $f : V \to W$ est déterminée par les images $f(e_j)$ des éléments d'une base $\{e_j\}$ de V et l'image de f est engendrée par les $f(e_j)$.

En effet si $B = \{e_1, \ldots, e_m\}$ est une base de V et v est un élément de V il existe un unique m-uple $(a_1, \ldots, a_m)$ tel que $v = a_1 e_1 + \cdots + a_m e_m$, donc

$$f(v) = a_1 f(e_1) + \cdots + a_r f(e_m)$$

est bien déterminé par les $f(e_i)$.

Si maintenant on choisit une base $C = \{\epsilon_1, \ldots \epsilon_n\}$ de W, et si on écrit les $f(e_j)$ dans la base C comme

$$f(e_j) = \sum_{i=1}^{n} a_{ij} \epsilon_i$$

avec a_{ij} dans R, alors on obtient une matrice $M(f) = M(f)_B^C := (a_{ij})$ à n lignes et m colonnes d'éléments de R dont la j-ème colonne correspond à $f(e_j)$:

$$M(f) = M(f)_B^C = (f(e_1)| \cdots | f(e_m)) \ .$$

En résumé :

à toute application linéaire $f : V \to W$ et à tout choix de bases B de V et C de W on associe une matrice $M(f) = M(f)_B^C$, à $\dim_R(V)$ colonnes et $\dim_R(W)$ lignes.

Remarque : noter que pour définir la matrice associée à une application linéaire on a choisi un ordre sur les bases.

Exemple. La matrice de la fonction identité $id : R^n \to R^n$ pour le même choix de base à la source et au but donne la *matrice identité* I_n de taille $n \times n$. Si on choisit deux bases différentes d'un même espace vectoriel on obtient pour l'identité la matrice dite de changement d'une base à l'autre, ou plus simplement *matrice de changement de base.*

Aux opérations avec les applications correspondent des opérations avec les matrices.

Soit $f, g : V \to W$ des applications linéaires, alors pour un choix de bases de V et W fixé

$$M(f + g) = M(f) + M(g) \ .$$

Si $f : V \to W$ et $g : W \to U$ sont des applications linéaires, alors pour un choix de bases B de V, C de W et D de U fixé

$$M(g \circ f)_B^D = M(g)_C^D \cdot M(f)_B^C \ ,$$

c'est-à-dire que par la correspondance $f \mapsto M(f)$ la composition d'applications correspond au produit de matrices. On peut utiliser la dernière égalité pour montrer que le produit de matrices est associatif : $A(BC) = (AB)C$.

Réciproquement étant donné une matrice A de $M_{n,m}(R)$ on définit une application

$$L_A : R^m \to R^n$$

en posant $L_A(X) = AX$, où X est vu comme un élément de $M_{m,1}(R)$. On vérifie que A *est la matrice associée à l'application L_A, si l'on choisit comme bases pour R^m et R^n les bases canoniques.* Aussi, si l'on identifie V à R^m et W à R^n par un choix de bases B et C, alors

$$L_{M_B^C(f)} = f \ .$$

On vérifie qu'une matrice A est inversible si et seulement si l'application linéaire associée est bijective.

Rang. On définit le *rang d'une matrice* A à n ligne et m colonnes à coefficients dans R comme étant la dimension du sous-espace de R^n engendré par les m colonnes de la matrice. En particulier le rang est au plus n. On voit que pour le rang de A, noté $\mathrm{rang}(A)$, on a

$$\mathrm{rang}(A) = \dim_R(L_A(R^m)) \ .$$

On ne change pas le rang d'une matrice par des opérations élémentaires sur les lignes. En particulier le rang d'une matrice et de sa forme réduite est le même.

En effet, pour toute matrice E de taille $n \times n$ on a

$$\mathrm{rang}(EA) = \dim_R(L_{EA}(R^m)) = \dim_R((L_E \circ L_A)(R^m)) = \dim_R(L_E(L_A(R^m))) \ .$$

Posons $V = L_A(R^m)$, alors pour E inversible L_E est inversible et

$$\dim_R(L_E(V)) = \dim_R(V) \ .$$

(Il est clair que $\dim_R(L_E(V)) \leq \dim_R(V)$ et pour E inversible on ne peut pas avoir inégalité stricte car alors on aurait une relation de dépendance linéaire comme $L_E(e) = \sum a_j L_E(e_j)$, qui donnerait, par soustraction, un élément non-nul du noyau de L_E.) Donc on a bien $\mathrm{rang}(EA) = \mathrm{rang}(A)$.

On montre de même, que si E est inversible, alors

$$\mathrm{rang}(AE) = \mathrm{rang}(A) \ .$$

On définit le *rang* d'une application linéaire $f : V \to W$ comme étant la dimension de l'espace vectoriel image de f :

$$\mathrm{rang}(f) := \dim_R \mathrm{im}(f) \ .$$

Par ce qui précède on obtient que

le rang d'une application linéaire égale le rang de n'importe quelle matrice qui lui est associée par un choix de bases dans le domaine et dans le but.

En particulier on peut calculer le rang d'une application linéaire par la méthode du pivot. En fait la méthode du pivot permet de montrer le théorème suivant.

Théorème du rang. *Soit V et W des espaces vectoriels de dimension finie sur le corps R et soit $f : V \to W$ une application R-linéaire. Alors*

$$\dim_R(V) = \dim_R \ker(f) + \dim_R \mathit{im}(f) \ .$$

En effet, si avec un choix de bases on identifie V à R^m et W à R^n, alors le rang de f sera égal au nombre de colonnes linéairement indépendantes de la matrice A de taille $n \times m$ associée. Or la méthode du pivot permet de calculer ce nombre sur la matrice réduite A' associée à A. On a vu que la méthode du pivot sépare les m coordonnées en deux sous-ensembles, qui correspondent à la forme de la matrice réduite A' : l'ensemble des coordonnées-pivot et celui des coordonnées-non-pivot. Le nombre de non-pivots donne la dimension de l'espace des solutions du système homogène $A'X = 0$, c'est-à-dire $\dim_R \ker(f)$ et le nombre de pivot donne le rang de f, ce qu'il fallait démontrer !

En conclusion, voici d'autres énoncés utiles :

Théorème. *Sont équivalents :*

a) B est une base de V.

b) B engendre V et si on lui enlève un élément, alors B n'engendre plus V.

c) B est une famille libre de V et on lui rajoute un élément, alors B n'est plus libre.

Théorème. *Soit V un espace vectoriel de dimension n (sur R).*
a) Tout sous-ensemble de V contenant $n+1$ éléments n'est pas libre.
b) $B \subset V$ est une base de V si il possède deux des trois des propriétés suivantes :
i) card$(B) = n$
ii) B engendre V
iii) B est libre dans V.

Théorème de la base incomplète. *Soit L libre dans V (supposé de génération finie), alors il existe une base B de V qui contient L.*

(Utiliser par exemple le Lemme d'échange.)
Formule sur les dimensions. *Soit W et W' deux sous-espaces d'un espace vectoriel (sur R) de dimension finie V, alors*

$$\dim_R(W + W') = \dim_R W + \dim_R W' - \dim_R(W \cap W') \ .$$

Dans la situation du dernier théorème, on dit que W et W' sont *supplémentaires* si $W \cap W' = \{0_V\}$. La somme de deux sous-espaces supplémentaires est dite *somme directe*, noté $W \oplus W'$. On a donc :

$$\dim_R(W \oplus W') = \dim_R W + \dim_R W' \ .$$

Un exemple de deux sous-espaces supplémentaires est fourni par les ensembles W des fonctions paires ($f(-x) = x$ comme cos) et W' des fonctions impaires ($f(-x) = -x$ comme sin).

Exercice : un endomorphisme d'un R-espace vectoriel de dimension finie dans lui-même qui est injectif est bijectif.