Mathématiques

Option Spécifique

Stéphane Perret

Version 3.600

Table des matières

Chapitre 0

Les principes de base de la logique

En mathématique, une *expression bien formée* ou *proposition* est une expression qui a du sens et qui peut être vraie ou fausse.

0.1 Le principe de non-contradiction

La logique (et donc les mathématiques) est basée sur le *principe de non-contradiction*. Ce principe dit qu'une expression bien formée ne peut pas être vraie et fausse à la fois.

0.2 Le principe du tiers exclu

Le *principe du tiers exclu* stipule que si une expression bien formée n'est pas vraie, alors elle est fausse (ou que si elle n'est pas fausse, alors elle est vraie).

Ce principe est vrai pour la plupart des expressions bien formées, bien qu'il y ait des expressions qui ne vérifient pas le principe du tiers exclu (voir l'énigme du cyclope ci-dessous). Ces expressions très particulières se prononcent, en général, sur leur propre valeur de vérité. Dans la suite du cours, on admettra que nos propositions vont satisfaire ce principe.

L'énigme du cyclope

Vous voilà enfermé dans une caverne en compagnie d'un cyclope qui veut votre mort. Il vous donne néanmoins un choix : soit vous dites une proposition vraie et vous serez bouilli ; soit vous dites une proposition fausse et vous serez roti.

Que dire ?

1. On peut dire : «Vous allez me rotir !» (ou «Vous n'allez pas me bouillir !»)
   Si cette proposition était vraie, alors vous finiriez bouilli et ainsi cette proposition serait fausse ; il s'agit d'une contradiction, donc cette proposition ne peut pas être vraie.
   Si cette proposition était fausse, alors vous finiriez roti et ainsi cette proposition serait vraie ; il s'agit d'une contradiction, donc cette proposition ne peut pas être fausse.
   Cette proposition n'est donc ni vraie, ni fausse.
2. On peut aussi dire : «Je suis en train de mentir !»
   Si cette proposition était vraie, alors vous seriez en train de dire la vérité et ainsi cette proposition serait fausse ; il s'agit d'une contradiction, donc cette proposition ne peut pas être vraie.
   Si cette proposition était fausse, alors vous seriez en train de mentir et ainsi cette proposition serait vraie ; il s'agit d'une contradiction, donc cette proposition ne peut pas être fausse.
   Cette proposition n'est donc ni vraie, ni fausse.
3. On peut aussi dire : «Cette phrase est fausse !»

Réponse : il y a plusieurs propositions possibles. Voici deux exemples.

0.3 Les implications

Lorsqu'on a deux expressions bien formées P et Q, on écrit

$$\boxed{P \Rightarrow Q}$$

pour dire que l'expression P *implique* l'expression Q. Dans ce cas, P est l'*hypothèse* et Q est la *conclusion*.

Il y a différentes façons de lire $P \Rightarrow Q$. On peut dire :

Si P, alors Q	Si la proposition P est vraie, alors la proposition Q est vraie
Q si P	La proposition Q est vraie si la proposition P est vraie
P seulement si Q	La proposition P est vraie seulement si la proposition Q est vraie

Lorsque l'expression P n'implique pas l'expression Q, on note $P \not\Rightarrow Q$. C'est le cas lorsque Q est fausse quand P est vraie.

Remarques importantes

1. En mathématiques, on n'écrit jamais d'expressions bien formées fausses (sauf si on s'est trompé en toute bonne foi).
2. En mathématiques, lorsqu'on dit qu'une proposition (ou implication) est vraie, cela signifie qu'elle est TOUJOURS vraie (l'expression «l'exception qui confirme la règle» n'a pas sa place en mathématiques). Ainsi une proposition (ou implication) est fausse lorsqu'elle n'est pas toujours vraie.

Exemples d'implications

1. Jean a gagné au loto $\Rightarrow$ Jean a joué au loto.

   On lit : a) Le fait que Jean a gagné au loto implique le fait qu'il a joué au loto.
   b) Si Jean a gagné au loto, alors il a joué au loto.
   c) Jean a joué au loto, s'il a gagné.
   d) Jean a gagné au loto seulement s'il a joué.

   Cette implication est vraie, car on ne peut pas gagner sans jouer.

2. $2x = 6 \overset{:2}{\Longrightarrow} x = 3$.

   Cette implication est vraie, car si le double d'un nombre x vaut 6, alors le nombre x est égal à 3 (on divise chaque côté de l'égalité par 2).

3. Si un enseignant vous dit : «Les cancres s'asseyent au fond de la classe», il pense que :

   $$\text{Un élève est un cancre} \implies \text{Il s'assied au fond de la classe}$$

   Non seulement cela ne signifie pas qu'il y a des cancres dans la classe, mais surtout cela ne signifie en aucun cas que tous les élèves du fond de la classe sont des cancres. Ainsi, l'enseignant n'a pas affirmé que : «Ceux qui s'asseyent au fond de la classe sont des cancres». D'ailleurs, même cet enseignant sera d'accord de penser que :

   $$\text{Un élève s'assied au fond de la classe} \not\Longrightarrow \text{C'est un cancre}$$

0.4 La réciproque d'une implication

La *réciproque* d'une implication $P \Rightarrow Q$ est l'implication $P \Leftarrow Q$.
Lorsque la réciproque n'est pas vraie, on trace l'implication : $P \not\Leftarrow Q$.

Exemples Regardons les réciproques des deux premiers exemples précédents.

1. Jean a gagné au loto $\not\Longleftarrow$ Jean a joué au loto.

   En effet, Jean est comme tout le monde, il ne gagne pas systématiquement au loto quand il y joue.

2. $2x = 6 \overset{\cdot 2}{\Longleftarrow} x = 3$.

   En effet, si un nombre x vaut 3, alors son double vaut 6 (on multiplie chaque côté de l'égalité par 2).

Moralité

La valeur de vérité de la réciproque d'une implication est indépendante de celle de l'implication.

En effet, la première implication de l'exemple est vraie, alors que sa réciproque est fausse. Tandis que la deuxième implication de l'exemple est vraie et que sa réciproque est vraie.

0.5 Les équivalences

Lorsqu'on a deux expressions bien formées P et Q telles que $P \Rightarrow Q$ et $P \Leftarrow Q$, on écrit :

$$\boxed{P \Longleftrightarrow Q}$$

et on dit que la proposition P *est équivalente à* la proposition Q.

Lorsque la proposition P n'est pas équivalente à la propostion Q, on note $P \not\Leftrightarrow Q$. C'est le cas lorsque $P \not\Rightarrow Q$ ou $P \not\Leftarrow Q$.

Au lieu de dire que P est équivalent à Q, on peut aussi dire que

$$P \text{ si et seulement si } Q$$

Exemples d'équivalence

1. Georges est le frère de Sophie si et seulement si Sophie est la sœur de Georges.

   Il est évident que «Georges est le frère de Sophie» et «Sophie est la sœur de Georges» sont des propositions synonymes.

2. Jean a gagné au loto $\not\Longleftrightarrow$ Jean a joué au loto.

   En effet, l'implication '$\Leftarrow$' est fausse, donc l'équivalence est fausse (malgré le fait que '$\Rightarrow$' est vraie).

3. $2x = 6 \Longleftrightarrow x = 3$.

   En effet, les deux implications '$\Leftarrow$' et '$\Rightarrow$' sont vraies.

0.6 Le contraire d'une expression bien formée

Si P est une proposition, alors sa *proposition contraire* est notée non P, $\neg P$ ou $\sim P$.

Par exemple

Si P est la proposition «Il pleut», alors non P est la proposition «Il ne pleut pas» (et non pas «Il fait beau», car il peut aussi neiger, grêler, etc.).

Remarques

1. Le principe de non-contradiction affirme que P et non P ne peuvent pas être vraies en même temps. De même, elles ne peuvent pas être fausses en même temps.
2. Le principe du tiers exclu permet d'affirmer que :

$$\begin{cases} P \text{ est vraie} & \iff & \text{non } P \text{ est fausse} \\ P \text{ est fausse} & \iff & \text{non } P \text{ est vraie} \end{cases}$$

   On voit l'importance du principe du tiers exclu, car les contraires des phrases de l'énigme du cyclope, qui ne sont ni vraies, ni fausses, sont des phrases vraies.

0.7 La contraposée

La *contraposée* d'une implication $P \Rightarrow Q$ est l'implication non $Q \Rightarrow$ non P.

Théorème

La contraposée d'une implication I est une implication qui a la même valeur de vérité que l'implication I.

$$\boxed{\underbrace{P \Rightarrow Q}_{\text{implication } I} \iff \underbrace{\text{non } Q \Rightarrow \text{non } P}_{\text{contraposée de l'implication } I}} \quad (\bigstar)$$

Interprétations

1. Le sens '$\Longrightarrow$' de ($\bigstar$) signifie que
   Si l'implication $P \Rightarrow Q$ est vraie, alors sa contraposée non $Q \Rightarrow$ non P est vraie.
2. Le sens '$\Longleftarrow$' de ($\bigstar$) signifie que
   Si la contraposée non $Q \Rightarrow$ non P est vraie, alors l'implication $P \Rightarrow Q$ est vraie.
3. La contraposée du sens '$\Longrightarrow$' de ($\bigstar$) signifie que
   Si la contraposée non $Q \Rightarrow$ non P est fausse, alors l'implication $P \Rightarrow Q$ est fausse.
4. La contraposée du sens '$\Longleftarrow$' de ($\bigstar$) signifie que
   Si l'implication $P \Rightarrow Q$ est fausse, alors sa contraposée non $Q \Rightarrow$ non P est fausse.

Moralité

Quelque soit la valeur de vérité d'une implication, sa contraposée a exactement la même valeur de vérité et inversement.

Exemples

1. La contraposée de l'implication

   $$\text{Jean a gagné au loto} \implies \text{Jean a joué au loto}$$

   est

   $$\text{Jean n'a pas joué au loto} \implies \text{Jean n'a pas gagné au loto}$$

   Comme la première implication est vraie, le théorème affirme que la deuxième implication est aussi vraie.

2. La contraposée de la proposition

   $$\text{Jean a joué au loto} \not\Longrightarrow \text{Jean a gagné au loto}$$

   est

   $$\text{Jean n'a pas gagné au loto} \not\Longrightarrow \text{Jean n'a pas joué au loto}$$

   Comme la première proposition est vraie (l'implication «Jean a joué au loto $\Rightarrow$ Jean a gagné au loto» est fausse), le théorème affirme que la deuxième proposition est aussi vraie (l'implication «Jean n'a pas gagné au loto $\Rightarrow$ Jean n'a pas joué au loto» est fausse).

3. La contraposée de l'équivalence $2x = 6 \Leftrightarrow x = 3$ est $x \neq 3 \Leftrightarrow 2x \neq 6$.

   C'est la raison principale pour laquelle on résout rarement des équations où le symbole '=' est remplacé par le symbole '$\neq$'.

Remarque

Si on contrapose la contraposée d'une implication, on retrouve cette implication.

Preuve du théorème

'$\Longrightarrow$' On suppose que $P \Rightarrow Q$ est vraie. On doit montrer que non $Q \Rightarrow$ non P est vraie, donc encore supposer que non Q est vraie, afin de montrer que non P est vraie.

On remarque que si P était vraie, alors l'implication $P \Rightarrow Q$ nous permettrait d'affirmer que Q serait vraie, ce qui est impossible (principe de non contradiction) car Q est fausse (puisque non Q est supposé vraie (principe du tiers exclu)).

Par conséquent, P n'est pas vraie, donc non P est vraie (principe du tiers exclu).

On vient donc de montrer, grâce aux principes de non-contradiction et du tiers exclu, que :

$$\Big(P \Rightarrow Q\Big) \Longrightarrow \Big(\text{non } Q \Rightarrow \text{non } P\Big)$$

'$\Longleftarrow$' En refaisant le raisonnement '$\Longrightarrow$' en remplaçant P par non Q et Q par non P, on a :

$$\Big(\text{non } Q \Rightarrow \text{non } P\Big) \Longrightarrow \Big(\text{non (non } P) \Rightarrow \text{non (non } Q)\Big) \Longleftrightarrow \Big(P \Rightarrow Q\Big)$$

□

0.8 Trois méthodes pour démontrer des implications

Pour montrer que l'implication ci-dessous est vraie

$$P \Rightarrow Q$$

on peut utiliser l'une des trois méthodes ci-dessous.

1. La première est la *méthode directe* : on suppose que P est vraie et on essaie de démontrer que Q est aussi vraie.

2. La deuxième façon utilise la contraposée, c'est la *preuve par contraposée* : on montre l'implication équivalente non $Q \Rightarrow$ non P de manière directe. C'est-à-dire que l'on suppose que non Q est vraie et on cherche à démontrer que non P est vraie.

3. La troisième façon de faire, c'est de procéder *par l'absurde*. Cela consiste à faire comme si la conclusion Q était fausse et à essayer d'en dégager une contradiction (c'est-à-dire une proposition vraie et fausse en même temps). Par le principe de non-contradiction, cela signifie donc qu'il y a une erreur quelque part et, si la preuve est bien ficelée, que cette erreur ne peut être que le fait que Q est fausse. Ainsi, Q doit donc être vraie (si Q satisfait le principe du tiers exclu).

   Voici un exemple d'une *preuve par l'absurde* :

   > Montrons qu'il n'existe pas de nombre réel x tel que $x^2 = -1$.
   >
   > Par l'absurde, on suppose que la conclusion est fausse, c'est-à-dire qu'il existe un nombre réel x tel que $x^2 = -1$. Or, grâce à la règle des signes, on sait que $x^2 \geqslant 0$. Ainsi, on a $-1 = x^2 \geqslant 0$.
   >
   > On a une contradiction : $-1 \geqslant 0$.
   > Donc, il n'existe pas de nombre réel x tel que $x^2 = -1$.

0.9 Contre-exemples

Pour montrer que l'implication $P \Rightarrow Q$ est fausse, il faut un *contre-exemple*, c'est-à-dire un cas particulier pour lequel P est vraie et Q est fausse.

Exemple

On a :

$$x \text{ est un nombre pair } \not\Longrightarrow \frac{x}{2} \text{ est un nombre pair}$$

En effet, $x = 2$ fournit un contre-exemple, car 2 est un nombre pair et que $\frac{2}{2} = 1$ n'est pas un nombre pair. Ici, le nombre 2 est un contre-exemple.

Attention

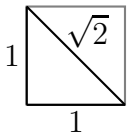
On ne démontre pas une implication à l'aide d'un exemple.

En effet, x est un nombre pair $\not\Rightarrow$ $\frac{x}{2}$ est un nombre pair. Pourtant, si on essaye avec $x = 4$, alors $\frac{x}{2} = \frac{4}{2} = 2$ est bien un nombre pair.

0.10 La découverte des nombres irrationnels

À la fin du VI^{e} siècle, les mathématiciens grecs, membres de l'école pythagoricienne, pensaient que deux grandeurs a et b étaient toujours *commensurables*, c'est-à-dire qu'il existait un nombre réel u (u comme unité) et deux nombres entiers m et n tels que $a = mu$ et $b = nu$, donc que $\frac{a}{b}$ est une fraction (car $\frac{a}{b} = \frac{mu}{nu} = \frac{m}{n}$).

Ils furent troublés de découvrir qu'ils avaient tort en étudiant un objet pourtant très simple, la diagonale du carré de côté 1, qui se trouve aussi être l'hypoténuse du triangle rectangle isocèle dont les cathètes sont de longueur 1.



car $\sqrt{2} = \sqrt{1^2 + 1^2}$
(par le théorème de Pythagore)

En effet, il se trouve que 1 et $\sqrt{2}$ sont incommensurables, car $\frac{\sqrt{2}}{1} = \sqrt{2}$ n'est pas une fraction. Puisque, pour les grecs, l'existence de tels nombres dépassait la raison, ces nombres furent appelés *irrationnels*.

Théorème

Le nombre $\sqrt{2}$ est un nombre irrationnel, c'est-à-dire $\sqrt{2} \notin \mathbb{Q}$.

Preuve

On note $2\mathbb{Z}$ l'ensemble des nombres qui sont des multiples de 2.

1. **Ingrédient** : Soit $n \in \mathbb{Z}$. Si $n^2 \in 2\mathbb{Z}$, alors $n \in 2\mathbb{Z}$.

   Il est équivalent de montrer la contraposée : si $n \notin 2\mathbb{Z}$, alors $n^2 \notin 2\mathbb{Z}$.

   Si n n'est pas un multiple de 2, alors n s'écrit $n = 2k + 1$ avec $k \in \mathbb{Z}$. On a

   $$n^2 = 4k^2 + 4k + 1 = 2(\overbrace{2k^2 + 2k}^{\in\mathbb{Z}}) + 1$$

   Ainsi, n^2 n'est pas un multiple de 2.

2. **La preuve par l'absurde.**

   Par l'absurde, on suppose que $\sqrt{2} \in \mathbb{Q}$. Donc $\sqrt{2} = \frac{a}{b}$ avec a, $b \in \mathbb{Z}$ et $b \neq 0$. On peut encore supposer que $\frac{a}{b}$ est irréductible.

   On a ainsi : $$\sqrt{2} = \frac{a}{b} \implies 2 = \frac{a^2}{b^2} \implies \boxed{a^2 = 2b^2} (\bigstar)$$

   Ainsi, on constate que $a^2 \in 2\mathbb{Z}$. Par l'**ingrédient**, on sait que $a \in 2\mathbb{Z}$.

   Par conséquent $a = 2k$ avec $k \in \mathbb{Z}$. En substituant ce résultat dans l'équation ($\bigstar$), on obtient

   $$(2k)^2 = 2b^2 \implies 4k^2 = 2b^2 \implies b^2 = 2k^2$$

   Ainsi, on constate que $b^2 \in 2\mathbb{Z}$. Par l'**ingrédient**, on sait que $b \in 2\mathbb{Z}$.

   Par conséquent, la fraction est réductible par 2. ↯ contradiction avec l'irréductibilité de $\frac{a}{b}$.

   Donc $\sqrt{2}$ est un nombre irrationnel.

Chapitre 1

Le théorème fondamental de l'arithmétique et sa preuve

1.1 Les nombres premiers

Lorsqu'on cherche à factoriser des nombres naturels le plus possible, certains nombres se distinguent. Afin de mettre en évidence ce phénomène, factorisons quelques nombres.

$$12 = 2 \cdot 6 = 2 \cdot 2 \cdot 3 \qquad 30 = 2 \cdot 15 = 2 \cdot 3 \cdot 5$$
$$39 = 3 \cdot 13 \qquad 175 = 5 \cdot 35 = 5 \cdot 5 \cdot 7$$
$$187 = 11 \cdot 17 \qquad 67 = 1 \cdot 67$$

On voit émerger certains nombres qui ne se factorisent pas : ce sont les *nombres premiers*.

Définition

Un *nombre premier* est un nombre p dont la seule factorisation possible est $p = 1 \cdot p$.

Remarques

1. Convention : on déclare que le nombre 1 n'est pas un nombre premier.
2. Voici les 18 premiers éléments de l'ensemble des nombres premiers.

$$\mathcal{P} = \{2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, \ldots\}$$

1.2 Le théorème fondamental de l'arithmétique

Théorème fondamental de l'arithmétique

Tout nombre naturel n plus grand que 1 se factorise de façon essentiellement unique en produit de nombres premiers.

1.2.1 Existence de la décomposition

Proposition

Soit $n \in \mathbb{N}$, $n > 1$.
Le plus petit diviseur de n différent de 1 est un nombre premier.

Preuve

Par l'absurde, supposons que le plus petit diviseur de n différent de 1, noté d, n'est pas premier. Ainsi, n ne pourrait pas être premier non plus (car si n est premier, alors son plus petit diviseur différent de 1 est lui-même) et le fait que d divise n se traduirait par

$$n = d \cdot m \qquad \text{avec } 1 < d \leqslant m < n$$

En effet, d étant le plus petit diviseur de n différent de 1, on a bien $d \leqslant m$. De plus, puisque d n'est pas premier, on a

$$d = a \cdot b \qquad \text{avec } 1 < a, b < d$$

Ainsi, on aurait

$$n = a \cdot b \cdot m \qquad \text{avec } 1 < a, b < d \leqslant m$$

Ce qui montre que a et b seraient des diviseurs de n différents de 1 plus petits que d. C'est une contradiction avec le fait que d est le plus petit diviseur de n différent de 1. □

Preuve de l'existence d'une décomposition en nombres premiers

Soit $n \in \mathbb{N}$, $n > 1$.

Si n est premier, alors n est sa propre décomposition en nombres premiers et la preuve est finie.

Par contre, si n n'est pas premier, alors son plus petit diviseur différent de 1, noté p_1 est premier. Ainsi

$$n = p_1 \cdot n_1 \qquad \text{avec } 1 < p_1 \leqslant n_1 < n$$

Si n_1 est premier, alors la décomposition en nombres premiers de n est

$$n = p_1 \cdot n_1$$

et la preuve est finie.

Par contre, si n_1 n'est pas premier, alors son plus petit diviseur différent de 1, noté p_2 est premier. Ainsi

$$n_1 = p_2 \cdot n_2 \qquad \text{avec } 1 < p_2 \leqslant n_2 < n_1$$

Si n_2 est premier, alors la décomposition en nombres premiers de n est

$$n = p_1 \cdot p_2 \cdot n_2$$

et la preuve est finie.

Par contre, si n_2 n'est pas premier, alors son plus petit diviseur différent de 1, noté p_3 est premier. Ainsi

$$n_2 = p_3 \cdot n_3 \qquad \text{avec } 1 < p_3 \leqslant n_3 < n_2$$

...

On se rend compte que l'on ne peut pas continuer ainsi indéfiniment puisqu'on aurait construit une suite décroissante de nombres n_i naturels tous plus grands que 1. Cette suite ne pourrait pas être infinie, donc il existe forcément un moment où le n_k sera premier et dans ce cas, la preuve sera finie. □

1.2.2 Unicité de la décomposition

Le lemme d'Euclide

Soit a et $b \in \mathbb{Z}$. Si p est un nombre premier qui divise ab, alors p divise a ou b.

Preuve

Cette preuve nécessite l'utilisation du théorème de Bezout (voir page 23). On distingue :

1. p divise a. Dans ce cas, c'est démontré !
2. p ne divise pas a. Dans ce cas, il faut démontrer que p divise b. Comme p est premier et que p ne divise pas a, alors $\text{pgcd}(p, a) = 1$. Par le théorème de Bezout, il existe deux nombres entiers x et y tels que $x \cdot p + y \cdot a = 1$.

   En multipliant cette équation par b, on obtient :

$$\underbrace{x \cdot p \cdot b}_{\text{divisible par } p} + \underbrace{y \cdot a \cdot b}_{\text{divisible par } p, \text{ car } p \text{ divise } ab} = b$$

   Donc b est divisible par p. □

Remarque

Soit p et q deux nombres premiers. Si p divise q, alors $p = q$.

Preuve

Si p divise q, alors il existe $m \in \mathbb{N}$ tel que $q = p \cdot m$. Comme q est premier et que $p \neq 1$ (car 1 n'est pas premier), on a $m = 1$ et $p = q$. □

Preuve de l'unicité de la décomposition en nombres premiers

Soit $n \in \mathbb{N}$, $n > 1$.
On suppose que n admet deux décompositions

$$n = p_1 \cdots p_m = q_1 \cdots q_{m'} \qquad \text{avec } m \leqslant m' \quad (m, m' \in \mathbb{N} \setminus \{0\})$$

On va montrer qu'à une permutation près, on retrouve les mêmes nombres premiers et qu'il y en a autant (c'est-à-dire $m = m'$).

On voit que p_1 divise $q_1 \cdots q_{m'}$. Par le lemme d'Euclide, p_1 divise un des q_i. Sans nuire à la généralité, on peut supposer que p_1 divise q_1. Par la remarque, on a donc $p_1 = q_1$.

En simplifiant par p_1 l'équation ci-dessus, on trouve

$$p_2 \cdots p_m = q_2 \cdots q_{m'}$$

Sans nuire à la généralité, on montre comme précédemment que $p_2 = q_2$, $p_3 = q_3$, etc. Finalement, il va rester

$$p_m = q_m \cdots q_{m'}$$

Cela signifie en même temps que $p_m = q_m$ et que $m = m'$, car aucun nombre premier n'est égal à 1. □

Remarque

Ce sont les "sans nuire à la généralité" qui sont la cause du mot "essentiellement" qui se trouve dans l'énoncé du théorème fondamental de l'arithmétique.

1.3 Il y a une infinité de nombres premiers

Le théorème fondamental de l'arithmétique nous permet de montrer qu'il existe une infinité de nombres premiers.

1.3.1 Première démonstration

Cette très belle preuve inventée par Euclide, s'est retrouvée dans un poème de Brian D. Beasley, adapté de Robert Frost.

Stopping By Woods
on a Snowny Evening
by Robert Frost

Whose woods these are I think I know.
His house is in the village though ;
He will not see me stopping here
To watch his woods fill up with snow.

My little horse must think it queer
To stop without a farmhouse near
Between the woods and frozen lake
The darkest evening of the year.

He gives his harness bells a shake
To ask if there is some mistake.
The only other sound's the sweep
Of easy wind and downy flake.

The woods are lovely, dark and deep.
But I have promises to keep,
And miles to go before I sleep,
And miles to go before I sleep.

Stopping by Euclid's Proof
of the Infinitude of Primes
by Brian D. Beasley

Whose proof this is I think I know.
I can't improve upon it, though ;
You will not see me trying here
To offer up a better show.

His demonstration is quite clear :
For contradiction, take the mere
n primes (no more), then multiply ;
Add one to that...the end is near.

In vain one seeks a prime to try
To split this number — thus, a lie !
The first assumption was a leap ;
Instead, the primes will reach the sky.

This proof is lovely, sharp, and deep.
But I have promises to keep,
And tests to grade before I sleep,
And tests to grade before I sleep.

Démonstration d'Euclide

On suppose par l'absurde qu'il y a un nombre fini de nombres premiers, disons n nombres premiers. Dans ce cas, l'ensemble $\mathcal{P}$ s'écrit

$$\mathcal{P} = \{p_1, p_2, p_3, \ldots, p_n\}$$

On examine le nombre

$$N = p_1 \cdot p_2 \cdot p_3 \cdots p_n + 1$$

Comme $N > 1$, il existe, grâce au théorème fondamental de l'arithmétique, un nombre premier p_k (avec $k \in \{1, 2, \ldots, n\}$) qui divise N. Or, ce nombre premier divise aussi $p_1 \cdot p_2 \cdot p_3 \cdots p_n$.

Par conséquent, p_k divise 1 car $1 = n - p_1 \cdot p_2 \cdot p_3 \cdots p_m$. Mais le seul nombre naturel qui divise 1 est 1. De ce fait, on a $p_k = 1$. C'est une contradiction (avec le fait que 1 n'est pas un nombre premier). □

1.3.2 Deuxième démonstration

Cette démonstration a l'élégance de ne pas être une démonstration par l'absurde.

Notation

On note

$$n! = 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$$

Par exemple, $1! = 1$; $2! = 1 \cdot 2 = 2$; $3! = 1 \cdot 2 \cdot 3 = 6$.

Théorème

Il existe une infinité de nombres premiers.

Preuve

Il suffit de démontrer que, pour tout nombre entier $n \geqslant 3$, il existe un nombre premier entre n et $n!$.

En effet, si c'est le cas, on sait qu'il y a un nombre premier entre 3 et 3!, un autre entre 3! et (3!)!, encore un autre entre (3!)! et ((3!)!)!, etc.

Montrons donc cette affirmation :

Dans ce but, on considère le nombre $n! - 1$. Puisque $n \geqslant 3$, on a $n! - 1 > 1$.

Par conséquent, ce nombre s'écrit de manière essentiellement unique comme produit de nombres premiers (grâce au théorème fondamental de l'arithmétique). On peut donc prendre un nombre premier p qui divise $n! - 1$.

Montrons par l'absurde que ce nombre premier p satisfait : $p > n$.

> Par l'absurde, on suppose que $p \leqslant n$. Dans ce cas p divise $n! = 1 \cdots p \cdots n$.
>
> Par conséquent, comme p divise $n!$ et $n!-1$, p divise leur différence qui vaut 1. Or le seul nombre entier positif qui divise 1 est 1 lui-même. Cela voudrait dire que $p = 1$. C'est impossible, car 1 n'est pas premier.

Ainsi p est un nombre premier entre n et $n!$ (en effet, puisque p divise $n! - 1$, on a $p \leqslant n! - 1 < n!$ et on vient de montrer que $p > n$). □

Chapitre 2

Les anneaux

2.1 Définitions

Définition

Un *anneau*[1] est un ensemble A muni de deux opérations internes, notées $\oplus$ et $\odot$. Un anneau est donc un triplet $(A, \oplus, \odot)$.

L'opération $\oplus$ satisfait les propriétés suivantes[2].

1. À chaque paire d'éléments de A, notés a_1 et a_2, on associe un unique élément de l'anneau A, noté $a_1 \oplus a_2$. En d'autres termes, $\oplus$ est une application bien définie
$$\oplus : A \times A \to A;\ (a_1; a_2) \mapsto a_1 \oplus a_2$$
2. Quelque soit a_1, a_2 et a_3 dans A, on a $(a_1 \oplus a_2) \oplus a_3 = a_1 \oplus (a_2 \oplus a_3)$.
3. Il existe un élément spécial de A, appelé *neutre additif* et noté 0 tel que
$$a \oplus 0 = 0 \oplus a = a \quad \text{quelque soit } a \in A$$
4. Pour chaque $a \in A$, il existe un *opposé*, noté $-a$ tel que $a \oplus -a = 0$.
5. Pour chaque paire d'éléments de A, notés a_1 et a_2, on a $a_1 \oplus a_2 = a_2 \oplus a_1$.

L'opération $\odot$ satisfait les propriétés suivantes[3].

6. Il existe un élément spécial de A, appelé *neutre multiplicatif* et noté 1 tel que
$$1 \odot a = a \ \text{ et } \ a \odot 1 = a \quad \text{quelque soit } a \in A$$
7. Quelque soit a_1, a_2 et a_3 dans A, on a $(a_1 \odot a_2) \odot a_3 = a_1 \odot (a_2 \odot a_3)$.

Il y a encore deux règles de compatibilité entre les opérations $\oplus$ et $\odot$. Il s'agit des règles de distributivité ou de mise en évidence.

$$a_1 \odot (a_2 \oplus a_3) \ = \ a_1 \odot a_2 \ \oplus \ a_1 \odot a_3$$
$$(a_2 \oplus a_3) \odot a_1 \ = \ a_2 \odot a_1 \ \oplus \ a_3 \odot a_1$$

1. Dans ce cours, il s'agit en fait d'un anneau unitaire.
2. Le couple $(A, \oplus)$ est d'un groupe additif commutatif (donnée par le cinquième axiome).
3. Ces propriétés sont proches de celles que l'on trouve dans les axiomes d'espaces vectoriels. Elles font penser à une action de groupe.

Définition

Un anneau $(A, \oplus, \odot)$ est dit *commutatif* si la propriété suivante est satisfaite.

Pour chaque paire d'éléments de A, notés a_1 et a_2, on a $a_1 \odot a_2 = a_2 \odot a_1$.

Conséquences

Des règles ci-dessus, on peut en DÉDUIRE les trois règles suivantes.

$$0 \odot a = 0 \qquad a \odot 0 = 0 \qquad (-1) \odot a = -a$$

La deuxième règle n'est pas superflue si l'anneau n'est pas commutatif. Pour la troisième règle, l'élément -1 est l'opposé du neutre multiplicatif, c'est-à-dire $-1 \oplus 1 = 0$.

Preuve

Montrons d'abord que pour chaque élément de l'anneau, il n'y a qu'un opposé possible (les règles disent a priori qu'il y a en au moins un).

Pour cela, supposons que si on prend deux opposés d'un élément $a \in A$, appelés a_1 et a_2, alors ils sont forcément égaux, c'est-à-dire $a_1 = a_2$.

En effet, puisque a_1 et a_2 sont des opposés de a, par définition, on a

$$a_1 \oplus a = 0 = a_2 \oplus a$$

Ainsi, on a

$$a_1 \oplus a = a_2 \oplus a$$

En faisant $-a$ de chaque côté, on trouve $a_1 = a_2$ (on a le droit car l'application $\oplus$ est bien définie et tout élément de l'anneau possède un opposé).

Déduisons maintenant les trois règles énoncées ci-dessus.

1. Soit $a \in A$, on a

   $$0 \odot a \ \oplus \ a \ = \ 0 \odot a \ \oplus \ 1 \odot a \ = \ (0 \oplus 1) \odot a \ = \ 1 \odot a \ = \ a$$

   En faisant $-a$ de chaque côté, on trouve $0 \odot a = 0$ (on a le droit car l'application $\oplus$ est bien définie et tout élément possède un opposé).

2. Soit $a \in A$, on a

   $$a \odot 0 \ \oplus \ a \ = \ a \odot 0 \ \oplus \ a \odot 1 \ = \ a \odot (0 \oplus 1) \ = \ a \odot 1 \ = \ a$$

   En faisant $-a$ de chaque côté, on trouve $0 \odot a = 0$ (on a le droit car l'application $\oplus$ est bien définie et tout élément possède un opposé).

3. Soit $a \in A$, on a

   $$(-1) \odot a \ \oplus \ a \ = \ (-1) \odot a \ \oplus \ 1 \odot a \ = \ \big((-1) \oplus 1\big) \odot a \ = \ 0 \odot a \ = \ 0$$

   Donc, par unicité de l'opposé, on a $(-1) \odot a = -a$ pour tout $a \in A$.

Exemples d'anneaux

1. Les nombres entiers $(\mathbb{Z}, +, \cdot)$, les nombres rationnels $(\mathbb{Q}, +, \cdot)$, les nombres réels $(\mathbb{R}, +, \cdot)$ et les nombres complexes $(\mathbb{C}, +, \cdot)$ sont tous des anneaux commutatifs. Le neutre additif est le 0 et le neutre multiplicatif est le 1.
2. Les fonctions réelles, dont le domaine de définition et le domaine d'arrivée est $\mathbb{R}$, forment un anneau commutatifs pour les opérations $+$ et $\cdot$ conventionnelles. Le neutre additif est la fonction $0 : \mathbb{R} \to \mathbb{R};\ x \mapsto 0$ (c'est la fonction qui vaut 0 quelque soit la valeur de x) et le neutre multiplicatif est la fonction $1 : \mathbb{R} \to \mathbb{R};\ x \mapsto 1$ (c'est la fonction qui vaut 1 quelque soit la valeur de x).
3. Les fonctions réelles, dont le domaine de définition et le domaine d'arrivée est $\mathbb{R}$, forment un anneau non commutatif pour les opérations $+$ et $\circ$ (composition de fonctions) conventionnelles. Le neutre additif est la fonction $0 : \mathbb{R} \to \mathbb{R};\ x \mapsto 0$ (c'est la fonction qui vaut 0 quelque soit la valeur de x) et le neutre multiplicatif est la fonction $\mathrm{id} : \mathbb{R} \to \mathbb{R};\ x \mapsto x$ (c'est la fonction qui dont l'image est la même que l'élément de départ).

Les nombres naturels ne forment pas un anneau

En effet, dans $\mathbb{N}$ muni de l'addition et de la multiplication usuelles, aucun nombre non nul n'admet d'opposé (dans $\mathbb{N}$). Ce qui contredit la propriété 4.

Définitions

Soit $(A, \oplus, \odot)$ un anneau.

1. Un élément a de A est dit *inversible* s'il existe b dans A tel que $a \odot b = 1$ et $b \odot a = 1$. Dans ce cas, on dit que b est l'*inverse de* a et on note $b = a^{-1}$.
2. On note $A^\times$ l'ensemble des éléments inversibles de A.
3. On dit que A est un anneau *intègre* si pour tout élément a, b dans A, on a

$$a \odot b = 0 \quad \Longrightarrow \quad a = 0 \ \text{ ou } \ b = 0$$

Définition

Un *corps* est un anneau A tel que $A^\times = A \setminus \{0\}$. C'est-à-dire lorsque tout élément non nul est inversible.

Théorème

Tout anneau A fini et intègre est un corps.

Preuve

On doit montrer que tout élément non nul de l'anneau est inversible. Soit $a \in A$ tel que $a \neq 0$. Disons que A possède exactement n éléments différents (c'est possible puisque A est supposé fini). Ainsi

$$A = \{a_1, \ldots, a_n\}$$

Regardons l'ensemble

$$B = \{a \odot a_1, \ldots, a \odot a_n\}$$

Cet ensemble a n éléments distincts (en exercice). Ainsi $A = B$ et par conséquent, il existe $a_i \in A$ tel que $a \odot a_i = 1$. Cela signifie que a est inversible, ce qu'on voulait montrer. □

2.2 L'anneau des matrices de taille 2 fois 2

Considérons un nouvel objet mathématique appelé *matrice*. Pour simplifier, on ne va étudier que les matrices à coefficients réels de taille 2 fois 2. Voici l'ensemble de telles matrices.

$$M_2(\mathbb{R}) = \left\{ \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} : a_{i,j} \in \mathbb{R} \right\}$$

On définit l'addition de deux matrices de la manière suivante.

$$\begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} + \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix} = \begin{pmatrix} a_{1,1}+b_{1,1} & a_{1,2}+b_{1,2} \\ a_{2,1}+b_{2,1} & a_{2,2}+b_{2,2} \end{pmatrix}$$

On définit la multiplication de deux matrices de la manière suivante. Elle est effectuée ligne par colonne.

$$\begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \cdot \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix} = \begin{pmatrix} a_{1,1}b_{1,1}+a_{1,2}b_{2,1} & a_{1,1}b_{1,2}+a_{1,2}b_{2,2} \\ a_{2,1}b_{1,1}+a_{2,2}b_{2,1} & a_{2,1}b_{1,2}+a_{2,2}b_{2,2} \end{pmatrix}$$

Les matrices sont très importantes en mathématiques et sont utilisées dans beaucoup de sujets de mathématiques appliquées (le moteur de recherche de `Google`, la programmation linéaire, les stratégies de jeux (en tandem avec la théorie des probabilités), les modèles économiques, l'imagerie par ordinateur, les modèles de populations animales, ...).

2.3 Les anneaux de congruences

2.3.1 Division euclidienne

On considère deux entiers a, b de l'anneau des nombres entiers $(\mathbb{Z}, +, \cdot)$. Si b n'est pas nul, on peut effectuer une division euclidienne de a par b. Cela permet d'obtenir un *quotient* q et un *reste* r tels que :

$$a = b \cdot q + r$$

Afin d'avoir l'unicité pour le quotient et pour le reste, on va toujours choisir le plus petit reste positif possible ! Cela signifie que l'on impose $r \geqslant 0$ et $r < |b|$.

Exemple

Si on veut distribuer 20 pièces de 5 centimes à 7 personnes, on va donner 2 pièces à chacune et il en restera 6. La division euclidienne de 20 par 7 livre donc un quotient de 2 et un reste de 6.

$$20 = 7 \cdot 2 + 6$$

2.3.2 Les congruences

Soit a et b deux nombres entiers et m un nombre naturel positif. Si la division euclidienne de a par m donne le même reste que celle de b par m, on dit que a *est congru à* b *modulo* m et on note

$$a \equiv b \pmod{m}$$

Exemple. L'exemple ci-dessus montre que $20 \equiv 6 \pmod{7}$.

Proposition

On a l'équivalence : $a \equiv b \pmod m \iff a - b$ est divisible par m
Autrement dit : $a \equiv b \pmod m \iff a - b \equiv 0 \pmod m$

Preuve

On effectue la division euclidienne de a par m et celle de b par m. On obtient :

$$a = mq_a + r_a \quad \text{avec} \quad 0 \leqslant r_a < m \quad \text{et} \quad b = mq_b + r_b \quad \text{avec} \quad 0 \leqslant r_b < m$$

Ainsi, on a $\boxed{a - b = m(q_a - q_b) + r_a - r_b}$(♠) avec $-m < r_a - r_b < m$ (♣).
Remarquons que $r_a - r_b$ n'est pas forcément le reste de la division euclidienne de $a - b$ par m. En effet, on n'a pas forcément $0 \leqslant r_a - r_b < m$, puisque $r_a - r_b$ peut être négatif.

Ainsi, on a : $a \equiv b \pmod m \iff a$ et b ont les mêmes restes de division par m
$\iff r_a = r_b \iff r_a - r_b = 0 \overset{(♣)}{\iff} r_a - r_b$ est divisible par m
$\overset{(♠)}{\iff} a - b$ est divisible par m □

Proposition

Si $a \equiv \alpha \pmod m$ et $b \equiv \beta \pmod m$. Alors :

a) $a + b \equiv \alpha + \beta \pmod m$ b) $a \cdot b \equiv \alpha \cdot \beta \pmod m$

Preuve

Par la proposition précédente, on sait que $a - \alpha$ et $b - \beta$ sont divisibles par m. Ainsi, il existe k_a et k_b dans $\mathbb{Z}$ tels que :

$$a - \alpha = k_a m \quad \text{et} \quad b - \beta = k_b m$$

a) Il faut s'assurer que $a + b - (\alpha + \beta)$ soit divisible par m. C'est bien le cas car :

$$a + b - (\alpha + \beta) = a - \alpha + b - \beta = k_a m + k_b m = (k_a + k_b)m$$

b) Il faut s'assurer que $a \cdot b - (\alpha \cdot \beta)$ soit divisible par m. Ici c'est un peu plus subtil. On a

$$a = k_a m + \alpha \quad \text{et} \quad b = k_b m + \beta$$

Donc

$$ab = (k_a m + \alpha) \cdot (k_b m + \beta) = k_a k_b m^2 + \alpha k_b m + \beta k_a m + \alpha\beta$$

Ce qui est équivalent à dire que $ab - \alpha\beta$ est bien divisible par m, car

$$ab - \alpha\beta = (k_a k_b m + \alpha k_b + \beta k_a)m$$

□

Remarque

On vient de montrer que l'on peut utiliser l'addition et la multiplication des nombres entiers dans le contexte des congruences. Cela permet de simplifier les calculs. Par exemple, on a :

$$\begin{cases} 49 \equiv 5 \pmod{11} \\ 118 \equiv 8 \pmod{11} \end{cases} \implies \begin{cases} 49 + 118 \equiv 5 + 8 \equiv 13 \pmod{11} \\ 49 \cdot 118 \equiv 5 \cdot 8 \equiv 40 \pmod{11} \end{cases}$$

Et ceci sans avoir eu à calculer les valeurs de $49 + 118$ et de $49 \cdot 118$ dans $\mathbb{Z}$.

Principe

Lorsqu'on calcule des congruences, on s'arrange toujours pour inscrire le plus petit nombre positif ou nul possible. Par exemples :

c) $125 \equiv 0 \pmod 5$ d) $591 \equiv 0 \pmod 3$ e) $50 \equiv 2 \pmod 4$

f) $53 \equiv 4 \pmod 7$ g) $-20 \equiv 1 \pmod 3$ h) $-44 \equiv 6 \pmod{10}$

Définition

Pour chaque $m \in \mathbb{N}$, on définit :

$$\mathbb{Z}_m = \{0, 1, 2, \ldots, m-1\}$$

En suivant le principe ci-dessus, l'addition et la multiplication de $\mathbb{Z}$ permet de mettre une structure d'anneau sur cet ensemble.

L'anneau $(\mathbb{Z}_m, +, \cdot)$ est appelé l'*anneau des restes de division modulo m*. Lorsqu'on calcule dans un tel anneau, on utilise le symbole $\equiv$ au lieu de $=$.

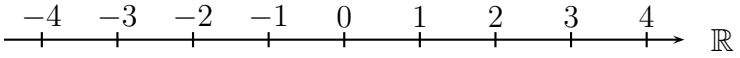
Utilités de tels anneaux

Ces anneaux apparaissent naturellement :

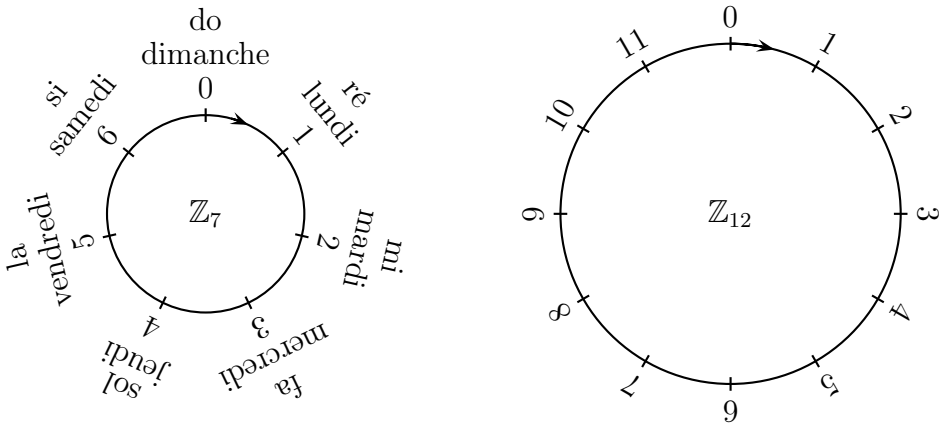
1. Les heures sont comptées modulo 24.
2. Les jours sont comptés modulo 7.
3. Les noms des notes naturelles (do, ré, mi, fa, sol, la, si) obéissent à une règle de calcul modulo 7.

Vision géométrique

Alors que les nombres entiers sont placés sur la droite réelle.



Les congruences modulo m sont représentés sur un cercle. Voici par exemple une représentation de $\mathbb{Z}_7$ et de $\mathbb{Z}_{12}$.



Remarque

On peut démontrer que : $\mathbb{Z}_m$ est un anneau intègre $\iff$ m est un nombre premier

Chapitre 3

Les équations diophantiennes

Les nombres réels sont très utiles, mais parfois on préfère résoudre des problèmes qui nécessitent des solutions à valeurs entières. Voici deux problèmes nécessitant l'utilisation d'outils spécifiquement élaborés pour résoudre des problèmes sur les entiers.

1. Un cinéma vend deux sortes de tickets : ceux à 12 CHF et ceux à 17 CHF.

   Un soir, la caissière constate qu'elle a encaissé 285 CHF, mais elle ne se souvient pas du nombre de billets de chaque sorte qu'elle a vendus.

   Est-il possible de le lui dire ? Et si le ticket le plus cher valait 18 CHF ?

2. On dispose de deux sabliers : un à 4 minutes, l'autre à 7 minutes. Comment faire pour déterminer un temps de 9 minutes ?

3.1 Calcul du pgcd, algorithme d'Euclide

Définition

Soit a et b deux nombres entiers.

On définit le *plus grand commun diviseur de a et b*, noté $\text{pgcd}(a, b)$, comme étant le plus grand nombre positif qui divise à la fois a et b.

Exemples

1. On a $\text{pgcd}(12, 14) = 2$.

   En effet, l'ensemble des diviseurs de 12 est $D_{12} = \{1, 2, 3, 4, 6, 12\}$ et l'ensemble des diviseurs de 14 est $D_{14} = \{1, 2, 7, 14\}$. L'ensemble des diviseurs commun à 12 et à 14 est donc $D_{12} \cap D_{14} = \{1, 2\}$. Ainsi, le plus grand commun diviseur est 2.

2. On a aussi $\text{pgcd}(2, 3) = 1$.

3. Ou encore $\text{pgcd}(7, -21) = 7$.

4. On a $\text{pgcd}(0, b) = b$ si $b \neq 0$, car 0 est divisible par tout nombre. On utilise la convention $\text{pgcd}(0, 0) = 0$ pour respecter la règle précédente (voir aussi le théorème de Bezout en page 23).

Définition. Deux nombres a et b tels que $\text{pgcd}(a, b) = 1$ sont dit *premiers entre-eux*.

Résultat

Soit a et b deux nombres entiers avec $b \neq 0$. En effectuant la division euclidienne de a par b, on obtient un quotient q et un reste r tels que $a = qb + r$ (et $0 \leqslant r < |b|$). Alors

$$\boxed{\text{pgcd}(a, b) = \text{pgcd}(b, r)}$$

Preuve

1. $\text{pgcd}(b, r) \leqslant \text{pgcd}(a, b)$.

   Pour montrer cela, il suffit de montrer que $\text{pgcd}(b, r)$ divise a et b. Ainsi, il sera bien plus petit ou égal au plus grand commun diviseur de a et de b, noté $\text{pgcd}(a, b)$.

   Or, il est évident que $\text{pgcd}(b, r)$ divise b et r (par définition). Ainsi il divise aussi qb et r, donc $\text{pgcd}(b, r)$ divise $a = qb + r$.

2. $\text{pgcd}(a, b) \leqslant \text{pgcd}(b, r)$.

   Pour montrer cela, il suffit de montrer que $\text{pgcd}(a, b)$ divise b et r. Ainsi, il sera bien plus petit ou égal au plus grand commun diviseur de b et de r, noté $\text{pgcd}(b, r)$.

   Or, il est évident que $\text{pgcd}(a, b)$ divise a et b (par définition). Ainsi il divise aussi a et qb, donc $\text{pgcd}(a, b)$ divise $r = a - qb$.

On a ainsi montré que $\text{pgcd}(b, r) \leqslant \text{pgcd}(a, b) \leqslant \text{pgcd}(b, r)$. Il est donc évident que $\text{pgcd}(a, b) = \text{pgcd}(b, r)$. □

3.1.1 L'algorithme d'Euclide

Soit a et b deux nombres entiers non nuls. Grâce au résultat précédent, on peut en effectuant des divisions euclidiennes successives calculer $\text{pgcd}(a, b)$.

$$\begin{array}{lllll}
\text{comme } b \neq 0, & a = bq_1 + r_1, & \text{pgcd}(a, b) = \text{pgcd}(b, r_1), & 0 \leqslant r_1 < |b| \\
\text{si } r_1 \neq 0, & b = r_1 q_2 + r_2, & \text{pgcd}(b, r_1) = \text{pgcd}(r_1, r_2), & 0 \leqslant r_2 < r_1 \\
\text{si } r_2 \neq 0, & r_1 = r_2 q_3 + r_3, & \text{pgcd}(r_1, r_2) = \text{pgcd}(r_2, r_3), & 0 \leqslant r_3 < r_2 \\
 & & \cdots & \\
\text{si } r_{n-1} \neq 0, & r_{n-2} = r_{n-1} q_n + r_n, & \text{pgcd}(r_{n-2}, r_{n-1}) = \text{pgcd}(r_{n-1}, r_n), & 0 = r_n < r_{n-1} \\
\text{si } r_n = 0, & & \text{pgcd}(r_{n-1}, r_n) = \text{pgcd}(r_{n-1}, 0) = r_{n-1} &
\end{array}$$

On a ainsi construit une suite d'égalité

$$\text{pgcd}(a, b) = \text{pgcd}(b, r_1) = \text{pgcd}(r_1, r_2) = \cdots = \text{pgcd}(r_{n-2}, r_{n-1}) = \text{pgcd}(r_{n-1}, r_n) = r_{n-1}$$

où le premier reste nul est r_n. Dans ce cas $\text{pgcd}(a, b)$ est égal au dernier reste non nul. Comme les restes forment une suite de nombres naturels strictement décroissante, il est certain qu'il y aura un premier reste nul (noté ici r_n).

Voici l'algorithme d'Euclide en JavaScript et en Python respectivement.

```
function euclide(a,b)
   {while ( b != 0 )
       {reste = a % b
        a = b
        b = reste
       }
   return a
   }
```

```
def euclide(a,b) :
    while ( b != 0 ) :
        reste = a % b
        a = b
        b = reste

    return a
```

3.2 Théorème de Bezout, algorithme d'Euclide étendu

3.2.1 Théorème de Bezout

Soit a et b deux nombres entiers.

Alors, il existe deux nombres entiers x et y tels que $ax + by = \text{pgcd}(a, b)$.

Preuve

L'algorithme d'Euclide étendu décrit ci-dessous permet de trouver les entiers x et y. □

3.2.2 L'algorithme d'Euclide étendu

Cet algorithme consiste à compléter l'algorithme d'Euclide.

Algorithme (la preuve est en annexe en page 30)

Il s'agit d'un tableau à quatre colonnes : la première colonne correspond à l'algorithme d'Euclide ; la dernière colonne est celle des quotients. Dans les deux colonnes au milieu, on place (de gauche à droite et de haut en bas) les nombres 1, 0, 0, 1. On peut choisir de remplir le tableau colonne par colonne en commençant par la première et la dernière colonne, on peut aussi remplir le tableau ligne par ligne. Les trois premières colonnes se construisent de manière similaire : on calcule le nombre suivant (en gris clair) à l'aide des deux nombres qui se trouvent juste en dessus et le quotient (en gris), comme montré ci-dessous. L'algorithme est terminé lorsqu'on a rempli la ligne du reste nul, notée (♢).

		a	b	
	a	1	0	quotients
	b	0	1	q_1
	$r_1 = a - bq_1$	$1 - 0q_1$	$0 - 1q_1$	q_2
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
	r_{i-1}	s_i	t_i	q_i
	r_i	u_i	v_i	q_{i+1}
	$r_{i+1} = r_{i-1} - r_iq_{i+1}$	$s_i - u_iq_{i+1}$	$t_i - v_iq_{i+1}$	q_{i+2}
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
(♡)	$r_{n-1} = \text{pgcd}(a, b)$	s_n	t_n	q_n
(♢)	$r_n = 0$	u_n	v_n	

On trouve à la ligne (♡) la combinaison de Bezout et un bonus à la ligne (♢).

$$(♡) : a \cdot s_n + b \cdot t_n = \text{pgcd}(a, b) \quad \text{et} \quad (♢) : a \cdot u_n + b \cdot v_n = 0$$

Exemple

On cherche la combinaison de Bezout pour $a = 28$ et $b = 6$.

		28	6	
	28	1	0	quotients
	6	0	1	4
	4	1	−4	1
(♡)	pgcd(28, 6) = 2	−1	5	2
(♢)	0	3	−14	

Ainsi, on a :

$$\begin{cases} \text{Bezout} & (\heartsuit): & 28\cdot(-1) & + & 6\cdot 5 & = & 2 \\ & (\diamondsuit): & 28\cdot 3 & + & 6\cdot(-14) & = & 0 \end{cases}$$

Remarque

À chaque ligne, on retrouve le terme de gauche en écrivant la combinaison avec les deux termes du milieu.

	a	b	
a	1	0	quotients
b	0	1	
...	...	...	
r	s	t	
...	...	...	

Autrement dit, quelque soit la ligne, on a :

$$r = a \cdot s + b \cdot t$$

Cette propriété, démontrée en page 32, permet de repérer les éventuelles erreurs de calcul.

3.2.3 **Lemme de Gauss** (généralisation du lemme d'Euclide)

Soit a et b deux nombres entiers.

Si c est un nombre tel que $\text{pgcd}(c, a) = 1$ et tel que c divise ab, alors c divise b.

Preuve

Par le théorème de Bezout, il existe deux nombres entiers x et y tels que

$$c \cdot x + a \cdot y = 1$$

En multipliant cette équation par b, on obtient :

$$\underbrace{c \cdot x \cdot b}_{\text{divisible par } c} + \underbrace{a \cdot y \cdot b}_{\text{divisible par } c\text{, car } c \text{ divise } ab} = b$$

Donc b est divisible par c. □

3.3 Les équations diophantiennes

Définition

Soit a, b et c trois nombres entiers. L'équation $ax+by=c$ est une *équation diophantienne* si les solutions cherchées x et y sont des nombres entiers.

Résultat d'existence d'une solution

Soit a et b deux nombres entiers. On a l'équivalence :

$$ax + by = c \text{ admet (au moins) une solution entière } \iff \text{pgcd}(a, b) \text{ divise } c$$

Preuve constructive

"$\Rightarrow$" Il est évident que $\text{pgcd}(a, b)$ divise ax et by, donc $\text{pgcd}(a, b)$ divise leur somme qui vaut c (car x et y sont solutions entières de l'équation $ax + by = c$).

"$\Leftarrow$" Par le théorème de Bezout, il existe deux nombres entiers m et n tels que :

$$am + bn = \text{pgcd}(a, b)$$

Ces deux nombres entiers m et n se trouvent grâce à l'algorithme d'Euclide étendu !

Par hypothèse, il existe $k \in \mathbb{Z}$ tel que $\text{pgcd}(a, b)k = c$ ($\Leftrightarrow k = \frac{c}{\text{pgcd}(a,b)}$). Ainsi en multipliant l'équation ci-dessus par k, on obtient :

$$a(mk) + b(nk) = \text{pgcd}(a, b)k = c$$

De ce fait, le couple $(x; y) = (mk; nk)$ est une solution de $ax + by = c$. □

Remarque importante

Avant de résoudre une équation diophantienne, on vérifie toujours si elle admet une solution en utilisant ce résultat d'existence. En effet, si l'équation n'admet pas de solution, alors le problème est clos. Alors que si elle possède une solution, il va falloir travailler pour toutes les trouver !

Recherche d'une solution particulière d'une équation diophantienne

Dans le cas où l'existence d'une solution est vérifiée, on peut commencer à chercher les solutions de l'équation diophantienne.

La méthode de recherche d'une solution particulière se trouve dans la preuve constructive du résultat d'existence d'une solution à l'équation diophantienne.

1. Grâce à l'algorithme d'Euclide étendu, on trouve une solution particulière $(m; n)$ de l'équation $ax + by = \text{pgcd}(a, b)$.
2. Pour trouver une solution particulière $(x_0; y_0)$ de l'équation $ax+by = c$, on multiplie m et n par $\frac{c}{\text{pgcd}(a,b)}$. Ainsi

$$(x_0; y_0) = \left(m \cdot \tfrac{c}{\text{pgcd}(a,b)}; n \cdot \tfrac{c}{\text{pgcd}(a,b)}\right)$$

Théorème de résolution d'une équation diophantienne

Soit l'équation diophantienne (ED) : $ax + by = c$ et $(x_0; y_0)$ une solution particulière. Soit aussi l'*équation homogène associée* (EH) : $ax + by = 0$. On a

1. Si $(x_h; y_h)$ est une solution de (EH), alors $(x_h + x_0; y_h + y_0)$ est une solution de (ED).
2. Si $(x; y)$ est une solution de (ED), alors $(x - x_0; y - y_0)$ est une solution de (EH).

Autrement dit, à travers la solution particulière $(x_0; y_0)$, à chaque solution de (ED) correspond une unique solution de (EH) et réciproquement.

Preuve

On suppose qu'on connaît une solution particulière $(x_0; y_0)$ de l'équation (ED).
On doit montrer :

1. Si $(x_h; y_h)$ est une solution de (EH), alors $(x; y) = (x_h + x_0; y_h + y_0)$ est une solution de (ED).

   Il suffit de vérifier (ED) pour $(x; y)$.

   $$ax + by = a(x_h + x_0) + b(y_h + y_0) = \underbrace{ax_h + by_h}_{=0 \text{ car } (x_h;y_h) \text{ est solution de } (EH)} + \overbrace{ax_0 + by_0}^{=c \text{ car } (x_0;y_0) \text{ est solution de } (ED)} = c$$

   Ainsi, $(x; y)$ est bien une solution de l'équation diophantienne (ED).

2. Si $(x; y)$ est une solution de (ED), alors $(x_h; y_h) = (x - x_0; y - y_0)$ est une solution de (EH).

   Il suffit de vérifier (EH) pour $(x_h; y_h)$.

   $$ax_h + by_h = a(x - x_0) + b(y - y_0) = \underbrace{ax + by}_{=c \text{ car } (x;y) \text{ est solution de } (ED)} - \big(\overbrace{ax_0 + by_0}^{=c \text{ car } (x_0;y_0) \text{ est solution de } (ED)}\big) = 0$$

   Ainsi, $(x_h; y_h)$ est bien une solution de l'équation homogène (EH). □

Solution générale de l'équation diophantienne

Lorsque $\text{pgcd}(a, b)$ divise c, les solutions de l'équation diophantienne $ax + by = c$ sont

$$\begin{cases} x = \boxed{x_0} \boxed{- \dfrac{b}{\text{pgcd}(a,b)} k} \\ y = \boxed{y_0} \boxed{+ \dfrac{a}{\text{pgcd}(a,b)} k} \end{cases}, \quad k \in \mathbb{Z}$$

Solution particulière (voir page précédente) ⇀ ↽ Solution générale de l'équation homogène (voir preuve page suivante)

Slogans

1. À chaque solution correspond un unique k (le même pour les deux équations).
2. À chaque nombre entier k correspond une unique solution.

Preuve

Le théorème de résolution permet d'énoncer la solution générale de l'équation diophantienne dès qu'on connaît une solution particulière et la solution générale de l'équation homogène. Ci-dessous, on démontre que la solution générale de l'équation homogène $ax + by = 0$ est bien celle précitée.

1. D'abord, on montre que les solutions entières de $ax + by = 0$ s'écrivent comme

$$x = -\frac{b}{\text{pgcd}(a,b)}\,k \quad \text{et} \quad y = \frac{a}{\text{pgcd}(a,b)}\,k \quad \text{avec} \quad k \in \mathbb{Z}$$

   Pour cela, on distingue :

   (a) Si $a \neq 0$ et $\text{pgcd}(a,b) = 1$.

   Dans ce cas, on a $-ax = by$, ainsi a divise by, mais comme $\text{pgcd}(a,b) = 1$, par le lemme de Gauss, on sait que a divise y (ou que y est un multiple de a).

   Par conséquent, $y = ak$ avec $k \in \mathbb{Z}$ et ainsi :

$$\left\{\begin{array}{l} ax + by = 0 \\ y = ak \end{array}\right. \overset{\text{subst.}}{\Longleftrightarrow} \left\{\begin{array}{l} ax + bak = 0 \\ y = ak \end{array}\right. \Longleftrightarrow \left\{\begin{array}{l} a(x + bk) = 0 \\ y = ak \end{array}\right.$$

$$\overset{a\neq 0}{\Longleftrightarrow} \left\{\begin{array}{l} x + bk = 0 \\ y = ak \end{array}\right. \Longleftrightarrow \left\{\begin{array}{l} x = -bk \\ y = ak \end{array}\right.$$

   On a donc les solutions désirées, puisque dans ce cas, on a $\text{pgcd}(a,b) = 1$.

   (b) Si $a \neq 0$ et $\text{pgcd}(a,b) \neq 1$.

   On se ramène au cas précédent en divisant l'équation $ax+by = 0$ par $\text{pgcd}(a,b)$.

$$ax + by = 0 \overset{:\text{pgcd}(a,b)}{\Longleftrightarrow} \frac{a}{\text{pgcd}(a,b)}\,x + \frac{b}{\text{pgcd}(a,b)}\,y = 0$$

   On se trouve bien dans le cas précédent car $\text{pgcd}\left(\frac{a}{\text{pgcd}(a,b)}, \frac{b}{\text{pgcd}(a,b)}\right) = 1$. Donc, il existe $k \in \mathbb{Z}$, tel que

$$x = -\frac{b}{\text{pgcd}(a,b)}\,k \quad \text{et} \quad y = \frac{a}{\text{pgcd}(a,b)}\,k \quad \text{avec} \quad k \in \mathbb{Z}$$

   (c) Dans le cas où $a = 0$, c'est b qui est non nul, et on effectue les raisonnements symétriques (en échangeant les rôles de a et b).

2. Il faut encore montrer que les valeurs

$$x = -\frac{b}{\text{pgcd}(a,b)}\,k \quad \text{et} \quad y = \frac{a}{\text{pgcd}(a,b)}\,k \quad \text{avec} \quad k \in \mathbb{Z}$$

   sont solutions de $ax + by = 0$ et ceci quelque soit la valeur de $k \in \mathbb{Z}$.

   C'est bien le cas, car

$$a \cdot \left(-\frac{b}{\text{pgcd}(a,b)}\,k\right) + b \cdot \left(\frac{a}{\text{pgcd}(a,b)}\,k\right) = -\frac{abk}{\text{pgcd}(a,b)} + \frac{abk}{\text{pgcd}(a,b)} = 0$$

□

Remarques

1. Puisque $\operatorname{pgcd}\left(\frac{a}{\operatorname{pgcd}(a,b)}, \frac{b}{\operatorname{pgcd}(a,b)}\right) = 1$, le pgcd des solutions de l'équation homogène est la valeur absolue de k.

   Par conséquent, si on a des solutions dont le pgcd vaut 1, alors ces solutions sont $x = -\frac{b}{\operatorname{pgcd}(a,b)}$ et $y = \frac{a}{\operatorname{pgcd}(a,b)}$, ou $x = \frac{b}{\operatorname{pgcd}(a,b)}$ et $y = -\frac{a}{\operatorname{pgcd}(a,b)}$ $(k = \pm 1)$.

2. Les deux dernières lignes de l'algorithme d'Euclide étendu sont très importantes.

		a	b	
	a	1	0	quotients
	b	0	1	q_1
	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$(\heartsuit)$	$\operatorname{pgcd}(a,b)$	m	n	q_n
$(\diamondsuit)$	0	$\pm\frac{b}{\operatorname{pgcd}(a,b)}$	$\mp\frac{a}{\operatorname{pgcd}(a,b)}$	

   À la ligne $(\heartsuit)$, on trouve une solution $(m; n)$ de l'équation $ax + by = \operatorname{pgcd}(a, b)$. Ainsi $(x_0; y_0) = (m \cdot \frac{c}{\operatorname{pgcd}(a,b)}; n \cdot \frac{c}{\operatorname{pgcd}(a,b)})$ est une solution particulière de l'équation diophantienne $ax + by = c$.

   À la ligne $(\diamondsuit)$, on trouve (au signe près) les coefficients de k de la solution générale de l'équation homogène $ax + by = 0$. Pour démontrer que c'est bien le cas, il suffit de combiner la remarque précédente avec la conséquence du bas de la page 33.

Exemple

On désire résoudre l'équation diophantienne $34x + 16y = 14$.

1. On commence par vérifier si l'équation admet au moins une solution.

   C'est bien le cas car $\operatorname{pgcd}(34, 16) = 2$ divise 14.

2. Pour trouver la solution générale, on va calculer les lignes $(\heartsuit)$ et $(\diamondsuit)$ de l'algorithme d'Euclide étendu.

		34	16	
	34	1	0	quotients
	16	0	1	2
$(\heartsuit)$	2	1	-2	8
$(\diamondsuit)$	0	-8	17	

   Ainsi $(1; -2)$ est une solution particulière de l'équation $34x + 16y = \operatorname{pgcd}(34, 16)$, puisque la ligne $(\heartsuit)$ dit que $34 \cdot 1 + 16 \cdot (-2) = 2$.

   Donc $(7; -14)$ est une solution particulière de $34x + 16y = 14$. En effet, on la trouve en multipliant par $7 = \frac{14}{\operatorname{pgcd}(34,16)}$ la solution de l'équation $34x + 16y = \operatorname{pgcd}(34, 16)$.

   En utilisant la ligne $(\diamondsuit)$, on peut directement donner la solution générale de l'équation diophantienne $34x + 16y = 14$, qui est

$$\left\{ \begin{array}{rcrcr} x & = & 7 & - & 8k \\ y & = & -14 & + & 17k \end{array} \right. , \quad k \in \mathbb{Z}$$

3.4 Annexe sur la relation entre les droites du plan et les équations diophantiennes

1. Dans le plan $\mathbb{R}^2$

Soit a, b, $c \in \mathbb{Z}$. L'équation $ax + by = c$ est l'équation d'une droite d dans le plan $\mathbb{R}^2$. Une solution particulière $P_0(x_0; y_0)$ est un point P_0 de cette droite. Dans le chapitre de géométrie du cours DF, il est démontré que le vecteur $\binom{-b}{a}$ est un vecteur directeur de cette droite. On donne ainsi une représentation paramétrique de la droite

$$d : \left\{ \begin{array}{rcrcl} x & = & x_0 & - & bk \\ y & = & y_0 & + & ak \end{array} \right. , \quad k \in \mathbb{R}$$

2. Dans le réseau $\mathbb{Z}^2$

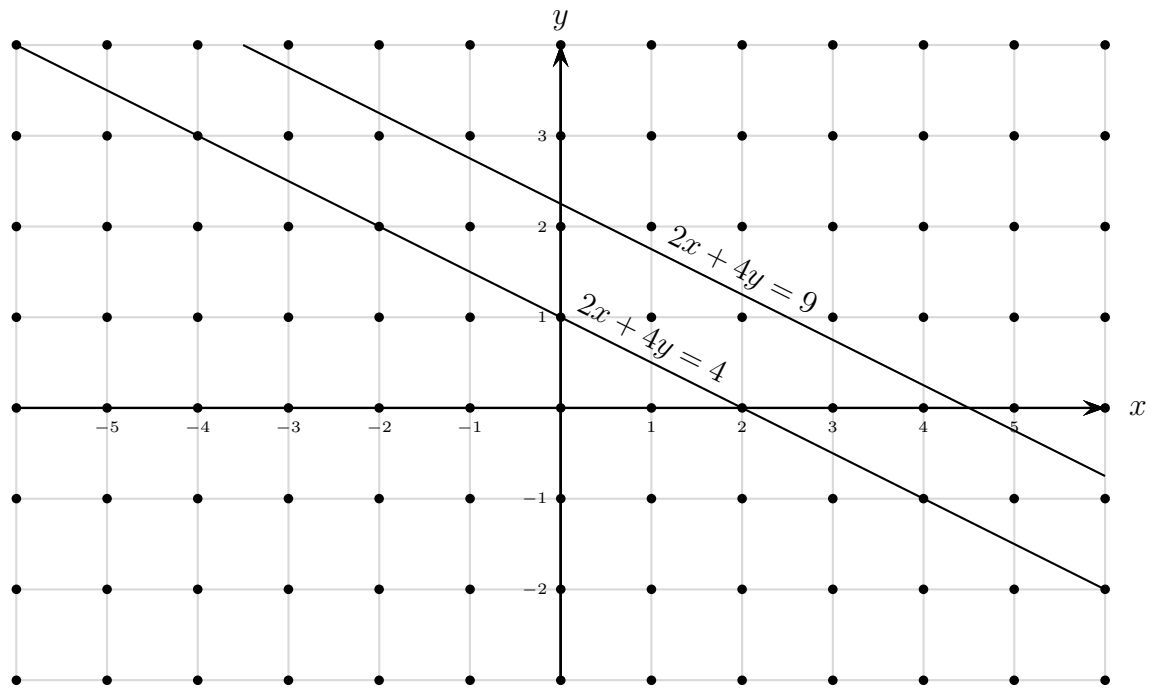
Soit a, b, $c \in \mathbb{Z}$. Le réseau $\mathbb{Z}^2$ est l'ensemble des points à coordonnées entières dans le plan $\mathbb{R}^2$. Lorsque $\text{pgcd}(a, b)$ divise c, on vient de voir que l'ensemble de solutions de l'équation diophantienne $ax + by = c$ est décrit par

$$\left\{ \begin{array}{rcrcl} x & = & x_0 & - & \frac{b}{\text{pgcd}(a,b)}k \\ y & = & y_0 & + & \frac{a}{\text{pgcd}(a,b)}k \end{array} \right. , \quad k \in \mathbb{Z}$$

En fait, les vecteurs $\binom{-b}{a}$ et $\begin{pmatrix} -\frac{b}{\text{pgcd}(a,b)} \\ \frac{a}{\text{pgcd}(a,b)} \end{pmatrix}$ sont parallèles, mais le deuxième est, au signe près, le vecteur parallèle à $\binom{-b}{a}$ à composantes entières le plus court.

Exemple

Ci-dessous, on voit la droite $d_1 : 2x + 4y = 9$, qui ne passe par aucun point du réseau $\mathbb{Z}^2$, puisque $\text{pgcd}(2, 4) = 2$ ne divise pas 9. On voit aussi la droite $d_2 : 2x + 4y = 4$, qui passe par une infinité de point du réseau $\mathbb{Z}^2$ car $\text{pgcd}(2, 4) = 2$ divise 4. Son vecteur directeur à composantes entières le plus court est, au signe près, $\binom{-2}{1}$.



3.5 Annexe sur l'algorithme d'Euclide étendu

Cet algorithme consiste à reproduire l'algorithme d'Euclide en n'oubliant pas les quotients de chaque étape.

Pour établir cet algorithme, la notation et la multiplication des matrices de taille 2 fois 2 sont essentielles.

Etape 1 : $\text{pgcd}(a,b) = \text{pgcd}(b,r_1)$. On effectue la division euclidienne $a = bq_1 + r_1$ avec $0 \leqslant r_1 < |b|$. Ainsi, on a $r_1 = a - bq_1$ et à l'aide de la notation matricielle, on peut écrire l'expression suivante.

$$\begin{pmatrix} b \\ r_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

Etape 2 : $\text{pgcd}(b,r_1) = \text{pgcd}(r_1,r_2)$. On effectue la division euclidienne $b = r_1q_2 + r_2$ avec $0 \leqslant r_2 < r_1$. Ainsi, on a $r_2 = b - r_1q_2$ et à l'aide de la notation matricielle, on peut écrire l'expression suivante.

$$\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_2 \end{pmatrix} \begin{pmatrix} b \\ r_1 \end{pmatrix}$$

Etape n : $\text{pgcd}(r_{n-2},r_{n-1}) = \text{pgcd}(r_{n-1},r_n)$. On effectue la division euclidienne $r_{n-2} = r_{n-1}q_n + r_n$ avec $0 = r_n < r_{n-1}$. Ainsi, on a $r_n = r_{n-2} - r_{n-1}q_n$ et à l'aide de la notation matricielle, on peut écrire l'expression suivante.

$$\begin{pmatrix} \text{pgcd}(a,b) \\ 0 \end{pmatrix} = \begin{pmatrix} r_{n-1} \\ r_n \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_n \end{pmatrix} \begin{pmatrix} r_{n-2} \\ r_{n-1} \end{pmatrix}$$

On retrouve la multiplication matricielle [1]

Grâce aux étapes 1 et 2, on peut exprimer r_2 à l'aide de a et b (rappelons que le but de l'algorithme est d'exprimer r_{n-1} (qui est égal au pgcd) en fonction de a et b (voir l'énoncé du théorème de Bezout)).
En effet, on a $r_1 = a - bq_1$ et $r_2 = b - r_1q_2$. Donc

$$r_2 = b - r_1q_2 = b - (a - bq_1)q_2 = b - aq_2 + bq_1q_2 = -q_2a + (1 + q_1q_2)b$$

Ce qui matriciellement donne l'expression suivante.

$$\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} 1 & -q_1 \\ -q_2 & 1 + q_1q_2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

Cette matrice s'obtient grâce à la multiplication matricielle suivante.

$$\begin{pmatrix} 0 & 1 \\ 1 & -q_2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -q_1 \end{pmatrix} = \begin{pmatrix} 1 & -q_1 \\ -q_2 & 1 + q_1q_2 \end{pmatrix}$$

On peut donc utiliser le produit matriciel suivant.

$$\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_2 \end{pmatrix} \begin{pmatrix} b \\ r_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -q_1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

1. Le lecteur avancé ne sera pas surpris de ce fait. En effet la multiplication matricielle correspond à la composition d'applications.

Les matrices produits

Maintenant que l'on a vu que des produits matriciels apparaissent, on va apporter une nouvelle notation. Pour chaque $i \in \{1, 2, \ldots, n\}$, on définit la matrice ci-dessous qui est en fait le produit des i premières matrices ayant le quotient comme coefficient.

$$M_i = \begin{pmatrix} s_i & t_i \\ u_i & v_i \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_i \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & -q_2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -q_1 \end{pmatrix}$$

Réécrivons nos étapes sous cette notation. Voici l'étape 1.

$$\begin{pmatrix} b \\ r_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = M_1 \begin{pmatrix} a \\ b \end{pmatrix}$$

Voici l'étape 2.

$$\begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_2 \end{pmatrix} \begin{pmatrix} b \\ r_1 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & -q_2 \end{pmatrix} M_1}_{M_2} \begin{pmatrix} a \\ b \end{pmatrix} = M_2 \begin{pmatrix} a \\ b \end{pmatrix}$$

Voici l'étape $i+1$

$$\begin{pmatrix} r_i \\ r_{i+1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_{i+1} \end{pmatrix} \begin{pmatrix} r_{i-1} \\ r_i \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & -q_{i+1} \end{pmatrix} M_i}_{M_{i+1}} \begin{pmatrix} a \\ b \end{pmatrix} = M_{i+1} \begin{pmatrix} a \\ b \end{pmatrix}$$

Et voici l'étape n (la dernière).

$$\begin{pmatrix} \text{pgcd}(a,b) \\ 0 \end{pmatrix} = \begin{pmatrix} r_{n-1} \\ r_n \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ 1 & -q_n \end{pmatrix} \cdots \begin{pmatrix} 0 & 1 \\ 1 & -q_2 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -q_1 \end{pmatrix}}_{M_n} \begin{pmatrix} a \\ b \end{pmatrix} = M_n \begin{pmatrix} a \\ b \end{pmatrix}$$

Ainsi, à la dernière étape, on voit la combinaison voulue dans le théorème de Bezout.

$$\begin{pmatrix} \text{pgcd}(a,b) \\ 0 \end{pmatrix} = \begin{pmatrix} r_{n-1} \\ r_n \end{pmatrix} = M_n \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} s_n & t_n \\ u_n & v_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} s_n \cdot a & + & t_n \cdot b \\ u_n \cdot a & + & v_n \cdot b \end{pmatrix} \quad (\bigstar)$$

Procédure itérative

Comme on vient de le voir, il faut trouver les coefficients de la matrice M_n pour trouver la combinaison voulue dans le théorème de Bezout.

La méthode la plus simple pour calculer M_n est itérative (penser à une démonstration par récurrence (aussi appelée démonstration par induction)). On connaît la matrice M_1.

$$M_1 = \begin{pmatrix} s_1 & t_1 \\ u_1 & v_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -q_1 \end{pmatrix} \quad \text{puisque} \quad \begin{pmatrix} b \\ r_1 \end{pmatrix} = M_1 \begin{pmatrix} a \\ b \end{pmatrix}$$

On peut aussi considérer une étape 0 qui fait intervenir une matrice M_0 (qui est l'identité car il s'agit de l'élément neutre de la multiplication).

$$M_0 = \begin{pmatrix} s_0 & t_0 \\ u_0 & v_0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{puisque} \quad \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

Si on connaît la i-ième matrice M_i, on peut trouver la matrice M_{i+1}. En effet, en se basant sur l'étape $i+1$ vue ci-dessus, on voit que :

$$\underbrace{\begin{pmatrix} s_{i+1} & t_{i+1} \\ u_{i+1} & v_{i+1} \end{pmatrix}}_{M_{i+1}} = \begin{pmatrix} 0 & 1 \\ 1 & -q_{i+1} \end{pmatrix} \underbrace{\begin{pmatrix} s_i & t_i \\ u_i & v_i \end{pmatrix}}_{M_i} = \begin{pmatrix} u_i & v_i \\ s_i - u_i q_{i+1} & t_i - v_i q_{i+1} \end{pmatrix}$$

On constate que la première ligne de M_{i+1} est égale à la deuxième ligne de M_i. Il se passe le même phénomène avec les vecteurs issus de l'algorithme d'Euclide.

$$\begin{array}{ccc} \text{Etape } i & \rightsquigarrow & \text{Etape } i+1 \\ M_i = \begin{pmatrix} s_i & t_i \\ u_i & v_i \end{pmatrix} & \rightsquigarrow & \begin{pmatrix} u_i & v_i \\ s_i - u_i q_{i+1} & t_i - v_i q_{i+1} \end{pmatrix} = M_{i+1} \\ \underbrace{\begin{pmatrix} r_{i-1} \\ r_i \end{pmatrix}}_{\text{vecteur de l'étape } i} & \rightsquigarrow & \underbrace{\begin{pmatrix} r_i \\ r_{i+1} \end{pmatrix}}_{\text{vecteur de l'étape } i+1} \end{array}$$

Algorithme

Dans cet algorithme, on place les vecteurs dans la première colonne, les matrices M_i dans les deux colonnes centrales et dans la dernière colonne, on écrit les quotients.

	a	b	
a	1	0	quotients
b	0	1	q_1
$\cdots$	$\cdots$	$\cdots$	$\cdots$
r_{i-1}	s_i	t_i	q_i
r_i	u_i	v_i	q_{i+1}
$r_{i+1} = r_{i-1} - r_i q_{i+1}$	$s_i - u_i q_{i+1}$	$t_i - v_i q_{i+1}$	q_{i+2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$r_{n-1} = \text{pgcd}(a,b)$	s_n	t_n	q_n
$r_n = 0$	u_n	v_n	

On trouve à l'avant-dernière ligne la combinaison de Bezout cherchée et un bonus à la dernière ligne (voir formule (★)) : $\text{pgcd}(a,b) = s_n a + t_n b$ et $0 = u_n a + v_n b$.

Remarque

À chaque ligne, on retrouve le terme de gauche en écrivant la combinaison avec les deux termes du milieu.

Quelque soit la ligne, on a $r = s \cdot a + t \cdot b$. Cette propriété permet de repérer une éventuelle erreur de calcul.

Cette propriété est issue des matrices M_i précédentes.
En effet, à chaque étape i, on a bien

$$\begin{pmatrix} r_{i-1} \\ r_i \end{pmatrix} = M_i \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} s_i & t_i \\ u_i & v_i \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} s_i \cdot a & + & t_i \cdot b \\ u_i \cdot a & + & v_i \cdot b \end{pmatrix}$$

	a	b	
a	1	0	quotients
b	0	1	
$\cdots$	$\cdots$	$\cdots$	
r	s	t	
$\cdots$	$\cdots$	$\cdots$	

Proposition 1

Pour tout $i \in \mathbb{N}$, les coefficients de $M_i = \begin{pmatrix} s_i & t_i \\ u_i & v_i \end{pmatrix}$ satisfont la propriété $s_i v_i - t_i u_i = \pm 1$.

Preuve par récurrence

1. Ancrage pour $i = 0$:

   Les coefficients de $M_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ satisfont la propriété qui est : $1 \cdot 1 - 0 \cdot 0 = 1$.

2. Pas de récurrence simple :

   on suppose que c'est vrai pour i et on montre que c'est vrai pour $i + 1$.

   L'algorithme d'Euclide étendu dit que :

   $$M_{i+1} = \begin{pmatrix} s_{i+1} & t_{i+1} \\ u_{i+1} & v_{i+1} \end{pmatrix} = \begin{pmatrix} u_i & v_i \\ s_i - u_i q_{i+1} & t_i - v_i q_{i+1} \end{pmatrix}$$

   où s_i, t_i, u_i et v_i sont les coefficients de la matrice M_i.

   La propriété peut se simplifier ainsi :

   $$\begin{aligned} s_{i+1}v_{i+1} - t_{i+1}u_{i+1} &= u_i(t_i - v_i q_{i+1}) - v_i(s_i - u_i q_{i+1}) \\ &= u_i t_i - v_i s_i = -(s_i v_i - t_i u_i) \overset{\text{HR}}{=} -(\pm 1) = \mp 1 \end{aligned}$$ □

Proposition 2

Soit a et b deux nombres entiers. S'il existe deux nombres entiers x et y tels que

$$ax + by = \pm 1$$

Alors a et b sont premiers entre-eux (c'est-à-dire que $\text{pgcd}(a, b) = 1$).

Preuve

Soit d un diviseur positif de a et de b. Alors d divise ax et by, donc d divise $ax + by$. Comme $ax + by = \pm 1$, on sait donc que d divise ± 1. Or, le seul diviseur positif de ± 1 est 1, donc $d = 1$.

Par conséquent, le seul diviseur positif commun à a et à b est 1. Cela signifie que a et b sont premiers entre-eux. □

Conséquence des propositions 1 et 2

Les nombres qui sont inscrits à chaque ligne dans les colonnes centrales de l'algorithme d'Euclide étendu sont premiers entre-eux !

	a	b	
a	1	0	quotients
b	0	1	q_1
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$r_{n-1} = \text{pgcd}(a, b)$	s_n	t_n	q_n
$r_n = 0$	u_n	v_n	

Chapitre 4

Systèmes de restes chinois

4.1 Un exemple de problème

Trois pilotes d'avion aimeraient à l'occasion dîner ensemble à Paris. Ils se concertent un dimanche par SMS et constatent que :

1. André se rendra à Paris le mardi suivant et y retournera tous les 5 jours.
2. Bernard se rendra à Paris le mercredi suivant et y retournera tous les 8 jours.
3. Cloé se rendra à Paris le jeudi suivant et y retournera tous les 13 jours.

Quand est-ce qu'ils pourront se retrouver pour dîner ?

4.2 Le ppcm

Définition

Soit a et b deux nombres entiers.

On définit le *plus petit commun multiple de a et b*, noté $\text{ppcm}(a, b)$, comme étant le plus petit nombre positif qui est multiple à la fois de a et de b.

Exemples

1. On a $\text{ppcm}(12, 14) = 84$.

   En effet, l'ensemble des multiples positifs (ou nul) de 12 est
   $$M_{12} = \{0, 12, 24, 36, 48, 60, 72, 84, 96, \ldots\}$$
   et l'ensemble des multiples positifs (ou nul) de 14 est
   $$M_{14} = \{0, 14, 28, 42, 56, 70, 84, 98, \ldots\}$$
   L'ensemble des multiples positifs (ou nul) commun à 12 et à 14 est donc $M_{12} \cap M_{14} = \{0, 84, 168, 252, \ldots\}$. Ainsi, le plus petit commun multiple est 84.

2. On a aussi $\text{ppcm}(2, 3) = 6$.

3. Ou encore $\text{ppcm}(7, -21) = 21$.

Le cas particulier du zéro

Lorsqu'un des deux termes est nul (ou les deux), on est obligé d'admettre la valeur 0 pour le ppcm. Autrement dit, on a $\text{ppcm}(0, b) = 0$ pour tout $b \in \mathbb{Z}$.

Résultat

Soit a et b deux entiers. Alors $\text{ppcm}(a,b) \cdot \text{pgcd}(a,b) = |ab|$.

Preuve

On va montrer que $\frac{|ab|}{\text{pgcd}(a,b)} = \text{ppcm}(a,b)$. On a :

$$\frac{|ab|}{\text{pgcd}(a,b)} = \frac{\pm|a|}{\text{pgcd}(a,b)} \cdot b = \frac{\pm|b|}{\text{pgcd}(a,b)} \cdot a \quad \text{où} \quad \frac{\pm|a|}{\text{pgcd}(a,b)} \text{ et } \frac{\pm|b|}{\text{pgcd}(a,b)} \in \mathbb{Z}$$

on constate ainsi que $\frac{|ab|}{\text{pgcd}(a,b)}$ est un multiple de a et de b. C'est le plus petit possible, puisqu'on ne peut pas diviser a et b par un nombre plus grand que $\text{pgcd}(a,b)$. □

4.3 Résolution de systèmes de restes chinois

Le théorème des restes chinois

Soit a_1 et a_2 deux nombres entiers. Soit m_1 et m_2 deux nombres naturels.

Le système suivant possède une solution si et seulement si $a_1 \equiv a_2 \pmod{\text{pgcd}(m_1,m_2)}$.

$$\begin{cases} x \equiv a_1 \pmod{m_1} \\ x \equiv a_2 \pmod{m_2} \end{cases}$$

De plus, si une solution existe, elle est unique modulo $\text{ppcm}(m_1,m_2)$.

Preuve

Il existe k_1 et $k_2 \in \mathbb{Z}$ tels qu'on a les équivalences :

$$\begin{cases} x \equiv a_1 \pmod{m_1} \\ x \equiv a_2 \pmod{m_2} \end{cases} \iff \begin{cases} x = a_1 + k_1 m_1 \\ x = a_2 + k_2 m_2 \end{cases}$$

$$\overset{\text{subst.}}{\iff} \begin{cases} x = a_1 + k_1 m_1 \\ a_1 + k_1 m_1 = a_2 + k_2 m_2 \end{cases} \iff \begin{cases} x = a_1 + k_1 m_1 \\ k_1 m_1 - k_2 m_2 = a_2 - a_1 \end{cases}$$

Ainsi le système de restes chinois admet une solution si et seulement si l'équation diophantienne $k_1 m_1 - k_2 m_2 = a_2 - a_1$ (d'inconnues k_1 et k_2) admet une solution. Or dans le résultat d'existence, on a vu que c'est le cas si et seulement si $\text{pgcd}(m_1,m_2)$ divise $a_2 - a_1$. Autrement dit $a_1 \equiv a_2 \pmod{\text{pgcd}(m_1,m_2)}$.

Pour l'unicité, prenons deux solutions x_1 et x_2 du système de restes chinois et montrons qu'elles sont égales modulo $\text{ppcm}(m_1,m_2)$.

Puisque $x_1 \equiv a_1$ et $x_2 \equiv a_1$ modulo m_1, on a $x_1 \equiv x_2 \pmod{m_1}$. De même, on a $x_1 \equiv x_2 \pmod{m_2}$. Ainsi, on a $x_1 - x_2 = k_1 m_1$ avec $k_1 \in \mathbb{Z}$ et $x_1 - x_2 = k_2 m_2$ avec $k_2 \in \mathbb{Z}$. On obtient ainsi une équation diophantienne $k_1 m_1 - k_2 m_2 = 0$. Les solutions de cette équation homogène sont (voir pages 26 et 27) :

$$k_1 = \frac{m_2}{\text{pgcd}(m_1,m_2)}\, k \quad \text{et} \quad k_2 = \frac{m_1}{\text{pgcd}(m_1,m_2)}\, k \quad \text{avec} \quad k \in \mathbb{Z}$$

Donc

$$x_1 - x_2 = \frac{m_1 m_2}{\text{pgcd}(m_1,m_2)}\, k = \text{ppcm}(m_1,m_2)k$$

C'est-à-dire que $x_1 \equiv x_2 \pmod{\text{ppcm}(m_1,m_2)}$. □

Pour la résolution

Lorsqu'on veut résoudre un système chinois à deux équations, on suit le principe de la démonstration en utilisant l'équivalence établie dans la preuve ci-dessus.

$$\left\{\begin{array}{rcll} x & \equiv & a_1 & (\text{mod } m_1) \\ x & \equiv & a_2 & (\text{mod } m_2) \end{array}\right. \iff \left\{\begin{array}{l} x = a_1 + k_1 m_1 \\ k_1 m_1 - k_2 m_2 = a_2 - a_1 \ \textit{«équation diophantienne»} \end{array}\right.$$

Il faut ainsi chercher k_1 (et k_2) en résolvant l'équation diophantienne à l'aide de l'algorithme d'Euclide étendu. On aura ainsi l'équivalence suivante (démontrée dans la preuve ci-dessus) où k_1 est la solution trouvée :

$$\left\{\begin{array}{rcll} x & \equiv & a_1 & (\text{mod } m_1) \\ x & \equiv & a_2 & (\text{mod } m_2) \end{array}\right. \iff x \equiv a_1 + k_1 m_1 \pmod{\text{ppcm}(m_1, m_2)}$$

Chapitre 5

Les bases de la cryptographie

Ce cours se considère bien modeste par rapport à l'immense travail de Didier Müller sur son site `https://www.apprendre-en-ligne.net/crypto/` avec une partie cryptologie très complète, incluant deux livres écrits par lui-même. Ce cours se base aussi sur la quatrième édition du livre de Friedrich L. Bauer intitulé «DECRYPTED SECRETS Methods and Maxims of Cryptology» des éditions Springer, ainsi que sur le cahier «Cryptologie» de Nicolas Martignoni édité par la Commission Romande de Mathématique (disponible à partir de `https://www.crm-editions.com/`). Une autre bonne lecture est le livre de Simon Singh appelé «Histoire des codes secrets» des éditions Le Livre de Poche.

5.1 Introduction au principe de cryptographie

Le but de la cryptographie est de cacher le contenu d'un message. Il y a pour cela différentes possibilités. Voici deux méthodes utilisées (parmi tant d'autres) :
- Cacher le message (dans une image, par exemple) : stéganographie.
- Rendre le message incompréhensible en le transformant en cryptogramme : chiffrement.

Deux exemples de chiffrement

1. Le carré de Polybe est la clé qui a permis de créer le premier système de chiffrement polygraphique connu.

↗	1	2	3	4	5
1	a	b	c	d	e
2	f	g	h	ij	k
3	l	m	n	o	p
4	q	r	s	t	u
5	v	w	x	y	z

   Ainsi, le mot SECRET sera codé par 43 15 13 42 15 44. Pour coder et décoder, il faut que celui qui envoie le message et que celui qui le reçoit aient tous les deux la même clé (ici, le carré de Polybe).

2. La machine Enigma était la clé du cryptage allemand durant la deuxième guerre mondiale. Le film *U571*[1] s'est inspiré du fait qu'une machine Enigma a été dérobée aux Allemands durant l'assaut d'un de leur sous-marin. Le film *The Imitation Game*[2] présente l'histoire de Alan Turing, l'homme qui a cassé le code de *Enigma*.

1. `https://fr.wikipedia.org/wiki/U-571_(film)`
2. `https://en.wikipedia.org/wiki/The_Imitation_Game`

5.2 Chiffrements par substitution monoalphabétique

Commençons par une notation : le *message clair* qu'on va crypter est écrit en minuscules dans un alphabet d'origine. Le *message crypté* est écrit en majuscules dans un alphabet de chiffrement ; il est coutume de mettre une espace tous les cinq caractères.

Dans ce cours, nous allons utiliser notre propre alphabet : l'alphabet latin[3], mais on pourrait utiliser une multitudes d'autres alphabets, même imaginaires[4]. Par exemple, pour crypter le message "J'aime les mathématiques", on utilise "jaimelesmathematiques" comme message clair. Ensuite, on choisit un autre alphabet qui peut être une simple permutation de notre alphabet latin. La correspondance entre les deux alphabets est la clé privée : il servira à encrypter et à décrypter.

Alphabet d'origine	a b c d e f g h i j k l m n o p q r s t u v w x y z
Alphabet de chiffrement	Z Y X W V U T S R Q P O N M L K J I H G F E D C B A

Ainsi "jaimelesmathematiques" est crypté par "QZRNV OVHNZ GSVNZ GRJFV H" en remplaçant les lettres de l'alphabet d'origine par les lettres correspondantes dans l'alphabet de chiffrement. Le décryptage s'obtient en effectuant la correspondance inverse.

Exemples de tels chiffrements

1. Le chiffrement *Atbash*[5] utilise l'alphabet présenté ci-dessus (à l'origine, il était utilisé avec l'alphabet Hébreu en lui faisant correspondre le même alphabet mais lu à l'envers).

2. Le carré de Polybe est un exemple de chiffrement par substitution monoalphabétique.

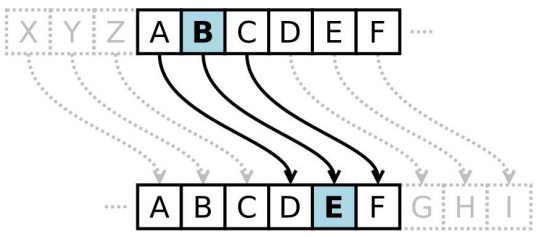
Alphabet d'origine	a	b	c	d	e	f	g	h	ij	k	...	y	z
Alphabet de chiffrement	11	12	13	14	15	21	22	23	24	25	...	54	55

3. Le chiffrement par décalage[6] aussi connu comme le *chiffre de César*.

   Pour ce chiffrement, Jules César utilisait un alphabet obtenu par un décalage cyclique de trois vers la droite pour ses correspondances secrètes.

Alphabet d'origine	a b c d e f g h i j k l m n o p q r s t u v w x y z
Alphabet de chiffrement	D E F G H I J K L M N O P Q R S T U V W X Y Z A B C

   L'expression décalage à droite est illustrée par l'image[7] suivante.



   On peut imaginer d'autres décalages, ce qui fait un total de 25 chiffres de César (sans compter le 26e, associé à un décalage de 0).

3. https://fr.wikipedia.org/wiki/Alphabet_latin
4. https://fr.wikipedia.org/wiki/Alphabet
5. https://fr.wikipedia.org/wiki/Atbash
6. https://fr.wikipedia.org/wiki/Chiffrement_par_d%C3%A9calage
7. https://commons.wikimedia.org/wiki/File:Caesar3.svg, libre de droit

5.3 Chiffrements par substitution polyalphabétique

Comme on le verra dans la section sur la cryptanalyse, les chiffrements par substitution monoalphabétique sont très sensibles à une attaque par analyse de fréquences. C'est pour éviter cette attaque que l'idée d'utiliser plusieurs alphabets est née. La manière d'indiquer quel alphabet de chiffrement on utilise à quel moment est donnée par ce qu'on appelle une *clé de chiffrement*. Le *chiffre de Vigenère* est un des chiffrement par substitution polyalphabétique les plus connus. Il associe à chaque caractère de la clé un chiffre de César avec un décalage différent (la clé doit être composée de caractères différents).

Le *chiffre de Vigenère* utilise le tableau suivant pour encrypter et décrypter.

	Lettre en clair																									
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
Lettre de la clé	Lettres chiffrées (au croisement de la colonne Lettre en clair et de la ligne Lettre de la clé)																									
A	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
C	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
D	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
E	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
F	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
G	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
H	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
I	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
J	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
K	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
L	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
M	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
N	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
O	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
P	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Q	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
R	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
S	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
T	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
U	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
V	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
W	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
X	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
Y	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Z	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

Exemple

Cet exemple est tiré à la fois du site web de Didier Müller et de son livre «Les 9 couronnes». On y encrypte un message qui contient 3 noms de thé avec le mot clé KILO.

clé	K	I	L	O	K	I	L	O	K	I	L	O	K	I	L	O	K	I	L	O	K	I	L	O	K
message en clair	t	h	e	r	u	s	s	e	t	h	e	j	a	s	m	i	n	t	h	e	c	h	i	n	e
message crypté	D	P	P	F	E	A	D	S	D	P	P	X	K	A	X	W	X	B	S	S	M	P	T	B	O

Le lecteur vérifiera qu'il peut chiffrer et déchiffrer le message grâce au tableau.

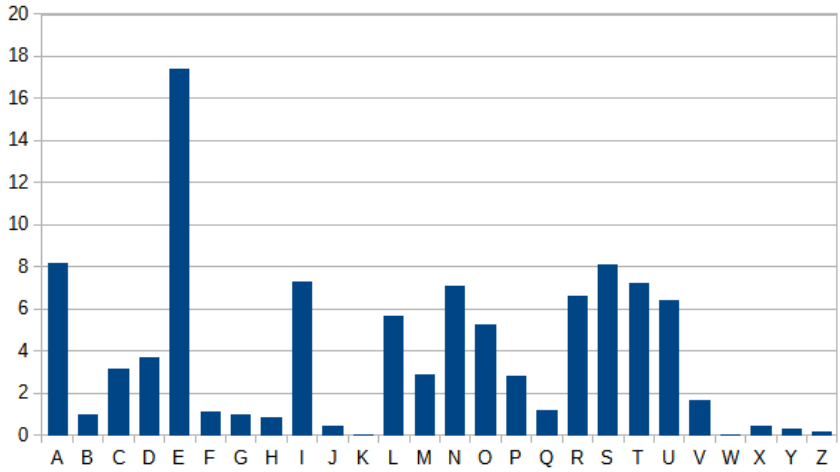
Incassable Si un message est chiffré avec une clé de même longueur, il est incassable.

5.4 Cryptanalyse des chiffrements par substitution

Cryptanalyse des chiffrements par substitution monoalphabétique

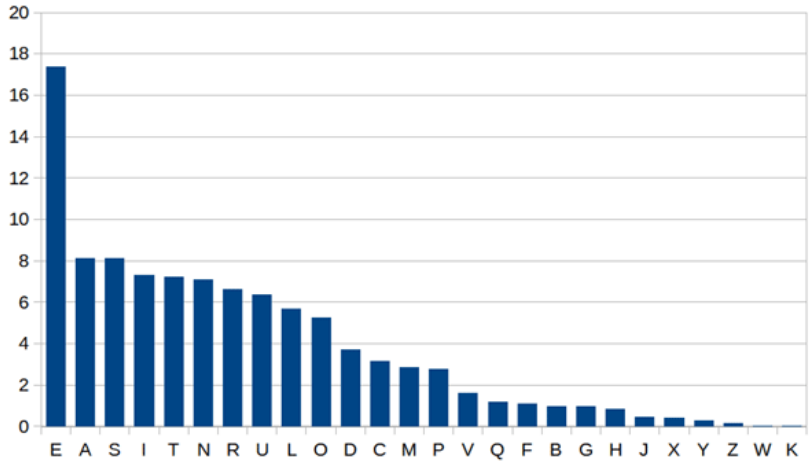
Le français (comme toute langue) est construit de sorte que les lettres qui composent les mots suivent une certaine loi : la lettre e est la plus fréquente, alors que k est la moins fréquente.

Didier Müller [8] a lui-même calculé ces fréquences à partir d'ouvrages divers (sur l'ordonnée on lit les fréquences en pourcentage).



Cet histogramme aide à cryptanalyser les chiffres à substitution polyalphabétique.

Ces fréquences peuvent aussi être classées selon un tri décroissant.



Cet histogramme aide à cryptanalyser les chiffres à substitution monoalphabétique.

Le principe de l'attaque par analyse de fréquences est de mettre en correspondance les fréquences ci-dessus avec les fréquences du texte à décrypter. Malheureusement, sur un texte relativement court, l'association par fréquences ne permet pas de trouver immédiatement le texte en clair à partir du texte crypté. En règle générale, cela permet d'identifier la lettre e, car c'est de loin la plus fréquente.

Toutefois, certain auteurs s'amusent avec les fréquences des lettres comme Georges Perec dans *La Disparition* qui ne contient aucunement la lettre e.

8. `https://www.apprendre-en-ligne.net/crypto/stat/francais.html`

C'est là qu'interviennent les bigrammes (aussi appelés digrammes), les trigrammes, etc.

bigrammes	bigrammes (doublets)	trigrammes	tétragrammes	pentagrammes	hexagrammes
es	ss	ent	ment	ement	dansle
le	ee	les	elle	aient	lement
en	ll	ait	quel	etait	quelle
re	tt	que	emen	dansl	quelqu
de	nn	ede	tion	comme	uelque
nt	mm	des	dans	avait	dansla
te	rr	lle	ient	ation	endant
ai	ff	res	esde	cette	taient
et	pp	ant	dela	elles	encore
er	cc	tre	omme	uelle	ementd
on	aa	eme	etai	ansle	edansl
ou	uu	ere	pour	equel	ements
se	dd	ese	tait	autre	ecomme
it	ii	our	aien	quele	aitpas
el	oo	sde	eque	edans	ations
an	bb	ela	ille	entde	pendan
la	gg	ien	plus	lemen	ansles
qu	xx	nte	ente	tions	queles
ne	zz	ele	vait	grand	vaient
ur	kk	men	avai	toute	tement
me	hh	qui	edes	quell	ecette
ie	qq	eur	sles	edela	autres
is	ww	ais	mais	squel	grande
em	yy	une	tout	uelqu	tionde
ec	vv	est	sque	quelq	etaien
ed	jj	par	comm	epour	tdansl

Dans le tableau ci-dessus, on trouve de haut en bas, les n-grammes les plus fréquents. Cela permet d'affiner l'analyse des fréquences en utilisant l'intuition qu'on a de la langue.

Ce code QR mène directement à la partie cryptanalyse des chiffrements par substitution monoalphabétique de `www.vive-les-maths.net`.

À partir de cette page HTML, on pourra utiliser l'analyse des fréquences et de la connaissances des n-grammes les plus fréquents pour casser les chiffrements proposés sur les différents exemples (un de ces exemples est tiré du fameux livre *La disparition*).

Cette page HTML mènera à la cryptanalyse des chiffrements à subtitution polyalphabétique qui permettra aussi aussi de décrypter un chiffre de César en réglant la longueur de la clé de chiffrement sur $k = 1$ et en effectuant le bon décalage.

Cryptanalyse des chiffrements par substitution polyalphabétique

Pour le chiffre de Vigenère, le point clé (!) de la cryptanalyse est de trouver la longueur de la clé de chiffrement. Une fois cette longueur déterminée, il suffit pour chaque lettre de la clé de trouver le bon décalage, car Vigenère utilise le chiffre de César, en utilisant l'analyse des fréquences.

En revenant à l'exemple de la page 41, on montre que la clé de longueur 4 permet de montrer les 4 parties encryptées avec chaque alphabet. On lit ainsi le message crypté de haut en bas, en commençant par la gauche.

Encrypté avec l'alphabet du K	D	E	D	K	X	M	O
Encrypté avec l'alphabet du I	P	A	P	A	B	P	
Encrypté avec l'alphabet du L	P	D	P	X	S	T	
Encrypté avec l'alphabet du O	F	S	X	W	S	B	

Avec Vigenère, on peut casser le cryptage séparément pour chaque ligne avec un décalage de César, même si on n'a aucune idée de la clé exacte ; il faut juste connaître sa longueur ! Une fois le code cassé, la clé apparaitra.

Pour la recherche de la longueur de la clé, il y a plusieurs moyens. Le premier moyen est basé sur les études des répétitions de suites de caractères dans le message crypté ; le deuxième moyen est basé sur les indices de coïncidence.

La méthode de Kasiski

Examinons plus attentivement l'exemple de la page 41.

clé	K	I	L	O	K	I	L	O	K	I	L	O	K	I	L	O	K	I	L	O	K	I	L	O	K
message en clair	t	h	e	r	u	s	s	e	t	h	e	j	a	s	m	i	n	t	h	e	c	h	i	n	e
message crypté	D	P	P	F	E	A	D	S	D	P	P	X	K	A	X	W	X	B	S	S	M	P	T	B	O

En 1863, Friedrich Wilhelm Kasiski a remarqué qu'une répétition *peut* se produire parce que, fortuitement, la clé se répète pour exactement les mêmes caractères du message en clair. Ainsi l'alphabet utilisé pour chacun de ces caractères sera exactement le même, et par conséquent une répétition se produit aussi dans le message crypté.

L'écart entre deux répétitions est donc un multiple de la longueur du mot clé, autrement dit la longueur de la clé est un diviseur de cet écart pour autant que la répétition n'est pas due au hasard mais bien à une répétition de la clé. Ici l'écart vaut 8. La clé étant de longueur 4, on voit bien que 4 divise 8.

clé	K	I	L	O	K	I	L	O	K	I	L	O	K	I	L	O	K	I	L	O	K	I	L	O	K
message en clair	t	h	e	r	u	s	s	e	t	h	e	j	a	s	m	i	n	t	h	p	f	h	i	n	e
message crypté	D	P	P	F	E	A	D	S	D	P	P	X	K	A	X	W	X	B	S	D	P	P	T	B	O

En modifiant légèrement le message clair, on peut faire en sorte qu'une autre répétition de DPP se produise de manière aléatoire avec un écart de 11. Pourtant, 4 (la longueur du mot clé) ne divise pas 11 (cet écart).

En pratique, on peut donc chercher toutes les répétitions sous forme de bigrammes, trigrammes, etc. et déterminer la liste des diviseurs des écarts trouvés. Les diviseurs revenant le plus souvent vont probablement donner la longueur de la clé. À l'époque, Kasiski cherchait à la main quelques écarts ; aujourd'hui, on peut utiliser un ordinateur pour trouver tous les écarts !

La méthode de Friedman

En 1925, William F. Friedman, utilisa un indice de coïncidence qu'il appela κ (kappa). Afin de définir cet indice de coïncidence, posons quelques notations.

On note $X = (x_i)_{1\leqslant i\leqslant n}$ la suite représentant un message de longueur n. De même, on note $X' = (x'_i)_{1\leqslant i\leqslant n}$ la suite représentant un autre message de même longueur. On définit l'indice κ ainsi

$$\kappa(X, X') = \sum_{i=1}^{n} \frac{\delta(x_i, x'_i)}{n} \quad \text{où} \quad \delta(x_i, x'_i) = \begin{cases} 1 & \text{si } x_i = x'_i \\ 0 & \text{sinon} \end{cases}$$

Théorème

Soit une langue S avec un alphabet $L = (a_k)_{1\leqslant k\leqslant m}$ dont les m lettres sont utilisées avec des probabilités p_i. Alors la variable aléatoire K (kappa majuscule) qui à un couple de messages X et X' associe le nombre $\kappa(X, X')$ est d'espérance $\sum_{k=1}^{m} p_k^2$.

Notation

Ce théorème montre qu'à chaque langue S, il correspond un unique indice de coïncidence, défini par

$$\mathrm{IC}_S = \sum_{k=1}^{m} p_k^2$$

Certains auteurs notent cet indice κ_S, ce qui fait penser qu'il est lié à $\kappa(X, X')$. Toutefois la *remarque importante* de la page 48 indique que ce nombre n'est pas lié qu'à $\kappa(X, X')$.

Remarque pour les novices en variables aléatoires

Une analogie est le lancé d'un dé bien équilibré à 6 faces. Calculer $\kappa(X, X')$ est analogue à lancer le dé (et obtenir un nombre de 1 à 6). L'espérance est une valeur moyenne *théorique* : en moyenne $\kappa(X, X')$ sera proche de IC_S, tout comme les résultats du dé seront, en moyenne, proche de $3.5 = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6)$.

Valeurs théoriques de certains indices de coïncidence

À partir des fréquences mesurées par Didier Müller, on a l'estimation

$$\mathrm{IC}_{\text{français}} \cong 0.0785$$

Cette estimation dépend des textes choisis[9] pour le dénombrement.

Pour une langue S à m lettres avec une même probabilité d'apparition, on a

$$\mathrm{IC}_{\text{random}} = \sum_{k=1}^{m} p_k^2 = \sum_{k=1}^{m} \left(\frac{1}{m}\right)^2 = m \cdot \frac{1}{m^2} = \frac{1}{m}$$

Pour le français *aléatoire*, on a $\mathrm{IC}_{\text{random}} = \frac{1}{26} \cong 0.0385$.

9. Les sources de Didier Müller : `https://apprendre-en-ligne.net/crypto/stat/francais.html`

Preuve du théorème

On utilise les variables aléatoires suivante ($i \in \{1, \ldots, n\}$).

$$\begin{array}{rccc} K: & \Omega & \longrightarrow & \mathbb{R} \\ & (X;X') & \longmapsto & \kappa(X,X') = \displaystyle\sum_{i=1}^{n} \frac{\delta(x_i, x_i')}{n} \end{array} \quad \text{et} \quad \begin{array}{rccc} K_i: & \Omega & \longrightarrow & \mathbb{R} \\ & (X;X') & \longmapsto & \delta(x_i, x_i') \end{array}$$

Pour calculer l'espérance de K, il suffit de calculer l'espérance de chaque K_i.

$$\begin{aligned} E(K_i) &= \sum_{x\in\mathbb{R}} x \cdot \mathbb{P}(K_i = x) = 0 \cdot \mathbb{P}(K_i = 0) + 1 \cdot \mathbb{P}(K_i = 1) = \mathbb{P}(x_i = x_i') \\ &= \mathbb{P}(x_i = x_i' = a_1) + \mathbb{P}(x_i = x_i' = a_2) + \cdots + \mathbb{P}(x_i = x_i' = a_m) \\ &= \mathbb{P}\left(\boxed{a_1 \mid a_1}\right) + \mathbb{P}\left(\boxed{a_2 \mid a_2}\right) + \cdots + \mathbb{P}\left(\boxed{a_m \mid a_m}\right) \\ &= p_1 \cdot p_1 + p_2 \cdot p_2 + \cdots + p_m \cdot p_m \end{aligned}$$

Ainsi, en utilisant la notation précédente, on a $E(K_i) = \mathrm{IC}_S$.

Comme $K = \frac{1}{n}\sum_{i=1}^{n} K_i$, par linéarité de l'espérance, on obtient

$$E(K) = \frac{1}{n}\sum_{i=1}^{n} E(K_i) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}_S = \frac{1}{n} \cdot n \cdot \mathrm{IC}_S = \mathrm{IC}_S$$

C'est la conclusion attendue. □

Retour à la méthode de Friedman

Friedman a remarqué que si la clé est de longueur k, alors en décalant le message crypté de k lettres vers la droite afin d'obtenir un nouveau message noté $X^{(k)}$ les alphabets seraient encore en correspondances (il y a quelques incohérences lorsque k n'est pas un diviseur de la longueur n du message, comme on le voit dans l'exemple de la page 41). Et ainsi, l'indice κ devrait être proche de la langue d'origine. C'est-à-dire $\kappa(X^{(k)}, X) \cong \mathrm{IC}_S$.

Alors que si on décale le message de ρ lettres où ρ n'est pas un multiple de k, les alphabets ne seront pas en correspondance, et $\kappa(X^{(\rho)}, X)$ devrait être a priori plus proche de $\mathrm{IC}_{\text{random}}$ que de IC_S.

Ainsi, en regardant $\kappa(X^{(\rho)}), X)$ pour quelques valeurs de ρ, on doit pouvoir détecter la longueur de la clé.

Méthode de Kasiski ou méthode de Friedman ?

Selon Friedrich L. Bauer dans *Decrypted Secrets*, page 343, la méthode de Kasiski à cause de sa partie aléatoire est moins digne de confiance que la méthode de Friedman. Néanmoins, grâce aux ordinateurs et au ratissage complet de toutes les répétitions (bigrammes, trigrammes, etc.), les deux méthodes paraissent complémentaires quand on utilise les pages HTML sur `www.vive-les-maths.net`.

La méthode de Kullback

En 1935, Solomon Kullback, utilisa des nombres qu'il appella χ (chi) et ψ (psi).

On reprend les notations précédentes, et on note n_k le nombre de fois que la lettre a_k apparait dans X, de même, on note n'_k le nombre de fois que la lettre a_k apparait dans X'. On définit les nombres χ et ψ ainsi

$$\chi(X,X') = \sum_{k=1}^{m} \frac{n_k \cdot n'_k}{n^2} \quad \text{et} \quad \psi(X) = \chi(X,X) = \sum_{k=1}^{m} \frac{n_k^2}{n^2}$$

Lien avec les probabilités

Il se trouve que $\chi(X,X')$ est exactement la probabilité de tirer la même lettre dans le message X que dans le message X'.

$$\mathbb{P}\left(\boxed{a_1}\boxed{a_1}\right) + \mathbb{P}\left(\boxed{a_2}\boxed{a_2}\right) + \cdots + \mathbb{P}\left(\boxed{a_m}\boxed{a_m}\right) = \frac{n_1}{n}\cdot\frac{n'_1}{n} + \frac{n_2}{n}\cdot\frac{n'_2}{n} + \cdots + \frac{n_m}{n}\cdot\frac{n'_m}{n}$$

De même, $\psi(X)$ est exactement la probabilité de tirer deux fois la même lettre dans le message X si on effectue un tirage avec remise.

$$\mathbb{P}\left(\boxed{a_1}\boxed{a_1}\right) + \mathbb{P}\left(\boxed{a_2}\boxed{a_2}\right) + \cdots + \mathbb{P}\left(\boxed{a_m}\boxed{a_m}\right) = \frac{n_1}{n}\cdot\frac{n_1}{n} + \frac{n_2}{n}\cdot\frac{n_2}{n} + \cdots + \frac{n_m}{n}\cdot\frac{n_m}{n}$$

Théorème kappa-chi et sa conséquence évidente

On a $\dfrac{1}{n}\displaystyle\sum_{\rho=0}^{n-1} \kappa(X^{(\rho)},X') = \chi(X,X')$ et son cas particulier $\dfrac{1}{n}\displaystyle\sum_{\rho=0}^{n-1} \kappa(X^{(\rho)},X) = \psi(X)$.

Preuve

On note $g_{k,j} = \begin{cases} 1 & \text{si } x_j = a_k \\ 0 & \text{sinon} \end{cases}$ et $g'_{k,i} = \begin{cases} 1 & \text{si } x'_i = a_k \\ 0 & \text{sinon} \end{cases}$.

On retrouve ainsi $n_k = \displaystyle\sum_{j=1}^{n} g_{k,j}$ et $n'_k = \displaystyle\sum_{i=1}^{n} g'_{k,i}$.

Cela permet aussi de généraliser les $\delta(x_i,x'_i)$ de la définition de $\kappa(X,X')$ et cela sera très utile pour la démonstration. En effet

$$\delta(x_j,x'_i) = \sum_{k=1}^{m} g_{k,j}\cdot g'_{k,i}$$

Le seul terme de cette somme qui vaut 1 arrive lorsque $x_j = a_k = x'_i$.

On a donc la généralisation voulue.

$$\delta(x_j,x'_i) = \begin{cases} 1 & \text{si } x_j = x'_i \\ 0 & \text{sinon} \end{cases}$$

On a

$$\frac{1}{n}\sum_{\rho=0}^{n-1} \kappa(X^{(\rho)},X') = \frac{1}{n}\sum_{\rho=0}^{n-1}\sum_{i=1}^{n} \frac{\delta(x_{i-\rho-1 \pmod n)+1}, x'_i)}{n}$$

les indices $i-\rho-1 \pmod n)+1$ parcourent tous les nombres entre 1 et n

$$= \frac{1}{n}\cdot\frac{1}{n}\cdot\sum_{j=1}^{n}\sum_{i=1}^{n}\delta(x_j,x'_i) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\sum_{k=1}^{m} g_{k,j}\cdot g'_{k,i} = \frac{1}{n^2}\sum_{k=1}^{m}\sum_{j=1}^{n}\sum_{i=1}^{n} g_{k,j}\cdot g'_{k,i}$$

$$= \frac{1}{n^2}\sum_{k=1}^{m}\left(\sum_{j=1}^{n} g_{k,j}\right)\cdot\left(\sum_{i=1}^{n} g'_{k,i}\right) = \frac{1}{n^2}\sum_{k=1}^{m} n_k\cdot n'_k = \chi(X,X')$$

□

Sur le chemin de phi

En constatant que $\kappa(X^{(0)}, X)$ vaut toujours 1 (car $\kappa(X^{(0)}, X) = \kappa(X, X) = 1$), Kullback a eu l'idée d'étudier la somme du théorème kappa-chi en partant de $\rho = 1$ au lieu de $\rho = 0$, il trouva le théorème kappa-phi et son résultat permit de poser la définition

$$\varphi(X) = \sum_{k=1}^{m} \frac{n_k \cdot (n_k - 1)}{n \cdot (n-1)}$$

Lien avec les probabilités

Il se trouve que $\varphi(X)$ est exactement la probabilité de tirer deux fois la même lettre dans le message X si on effectue un tirage sans remise.

$$\mathbb{P}\left(\boxed{a_1}\boxed{a_1}\right) + \cdots + \mathbb{P}\left(\boxed{a_m}\boxed{a_m}\right) = \frac{n_1}{n} \cdot \frac{n_1 - 1}{n-1} + \cdots + \frac{n_m}{n} \cdot \frac{n_m - 1}{n-1}$$

Théorème kappa-phi (preuve en exercice)

On a la relation $\displaystyle \frac{1}{n-1} \sum_{\rho=1}^{n-1} \kappa(X^{(\rho)}, X) = \sum_{k=1}^{m} \frac{n_k \cdot (n_k - 1)}{n \cdot (n-1)}$.

Avantages de phi sur psi

L'indice $\varphi(X)$ présente un avantage dans les calculs à la main, les termes de la somme sont nuls lorsque $n_i = 0$ (lorsque la lettre a_i n'apparait pas), mais aussi lorsque $n_i = 1$ (lorsque la lettre a_i n'apparait qu'une fois).

Selon Friedrich L. Bauer dans *Decrypted Secrets*, page 327, il y eu une autre raison pour laquelle les spécialistes du domaine travaillèrent de manière prédominante avec φ plutôt que ψ : Solomon Kullback, arguments à l'appui, le plus important étant probablement le fait que la moyenne des kappa est biaisée par le fait que $\kappa(X^{(0)}, X) = 1$, a proposé d'utiliser φ plutôt que de travailler avec χ et ψ.

Sur `www.vive-les-maths.net`, on voit clairement que φ met plus facilement la bonne longueur de la clé en évidence que ψ.

Théorème (preuve en exercice)

On a la relation $n \cdot \psi(X) = (n-1) \cdot \varphi(X) + 1$, et ainsi $\varphi(X) \leqslant \psi(X)$.

Indices de coïncidences

Le lien avec les probabilités (tirer aléatoirement deux fois de suite la même lettre est une sacrée coïncidence) le montre bien : χ, ψ et φ sont aussi des indices de coïncidences, mais historiquement c'est plutôt κ qui est le plus susceptible d'avoir droit à ce nom [10].

Remarque importante

Selon Friedrich L. Bauer dans *Decrypted Secrets*, pages 322, 325 et 328, tous ces indices de coïncidence ont la même espérance (pour φ, ce résultat est asymptotique). Ils permettent tous d'estimer $\mathrm{IC_S}$!

10. source : `https://de.wikipedia.org/wiki/Koinzidenzindex`

Retour à la méthode de Kullback

Pour un entier k donné, on définit X_1 à X_k comme les coupures de la chaîne X.

Par exemple, voici le cryptogramme du chapitre 6 de *Les 9 couronnes* de Didier Müller :

```
CRBQN NBREM YEHQE GIUTP OAVUQ NNQRR AQMIR VNVFE NAIEM CUYPI YIHIY FLYMQ
TUFTZ WDTYA EFINE FIEZW DTFBG YQEEU RRNQF ENBAZ WUTCB VVLQT IRJBC DSAOA
PMMMI FLRKM NNLNQ CVULX EFBYA CKTRV MNNZO AVGDU KSGWG TYIEH ZAPYB TZMYE
URDRT MOHAE IZMIN JEEMY ELZIR ZMUFF EHLQM YQRNY GELJA VAQNW LRRNM UXOAV
BULMX VBQDQ OFJRA GIMBN NBFEH AAABO OHQIA CQZXL NPIYA JMEYM DLYCA PBQUL
```

Dans cet exemple, on choisit $k = 7$ (essai avec 7 alphabets). Voici les 7 coupures de X :

$$\begin{aligned}
X_1 &= \texttt{CREARVIILTEEGRAVJALNETOSETDEEIERARAVJNAINEA}\\
X_2 &= \texttt{REGVRNEYYZFZYNZVBPRQFRAGHZRIERHNVRVBRNAAPYP}\\
X_3 &= \texttt{BMIUAVMIMWIWQQWLCMKCBVVWZMTZMZLYANBQABBCIMB}\\
X_4 &= \texttt{QYUQQFCHQDNDEFUQDMMVYMGGAYMMYMQGQMUDGFOQYDQ}\\
X_5 &= \texttt{NETNMEUITTETEETTSMNUANDTPEOIEUMENULQIEOZALU}\\
X_6 &= \texttt{NHPNINYYUYFFUNCIAINLCNUYYUHNLFYLWXMOMHHXJYL}\\
X_7 &= \texttt{BQOQRAPFFAIBRBBROFLXKZKIBRAJZFQJLOXFBAQLMC}
\end{aligned}$$

La méthode de Kullback consiste à calculer $\varphi(X_1)$ à $\varphi(X_k)$. Si k est la longueur de la clé (ou au moins un multiple de cette longueur), alors tous ces $\varphi(X_i)$ devraient être proches de $\text{IC}_{\text{français}}$, sinon, ils seront plus proches de $\text{IC}_{\text{random}}$. Il ne reste plus qu'à faire les calculs et à comparer les valeurs obtenues.

Il est aussi possible de calculer $\psi(X_1)$ à $\psi(X_k)$ avec les mêmes envies de conclusion.

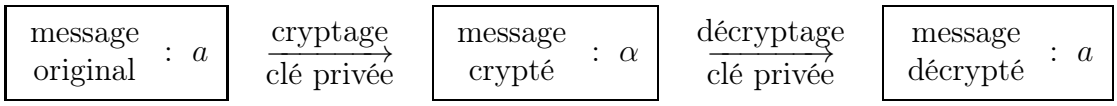
Ce code QR mène directement à la partie cryptanalyse des chiffrements par substitution polyalphabétique de `www.vive-les-maths.net`.

À partir de cette page HTML, on pourra s'orienter sur la méthode de Kasiski et les indices de coïncidence pour déterminer la longueur de la clé de chiffrement, puis utiliser l'analyse des fréquences de la page principale pour casser les chiffrements proposés sur les différents exemples (un de ces exemples est tiré du fameux livre *La disparition*).

5.5 Cryptages à clé privée et cryptages à clé publique

Les systèmes de cryptographie vus précédemment utilisent tous une clé unique qui appartient au codeur et au décodeur. Si une tierce personne parvient à s'emparer de la clé, elle serait en mesure d'intercepter et de décoder les messages, ou même d'usurper l'identité d'une des deux autres personnes.

Ces méthodes sont dites à clé privée. Elles sont symétriques, car celui qui reçoit le message utilise la même clé que celui qui l'envoie.



On voit qu'en utilisant la même clé, celui qui reçoit le message peut envoyer une réponse !

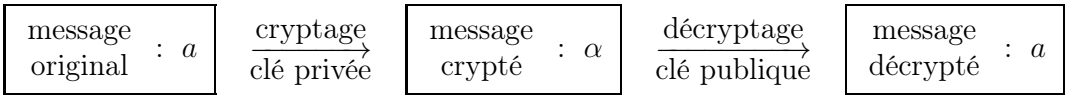
Vers la fin des années 1960, on a découvert des systèmes cryptographiques à clé publique. Ce système utilise deux clés : l'une privée, l'autre publique. La clé publique peut être connue de tout le monde, tandis que la clé privée n'est connue que d'une personne (donc elle est plus facile à protéger que dans les méthodes à clé privée vue précédemment).

Deux utilités des systèmes à clé publique

1. Tout le monde peut envoyer un message qui ne pourra être lu que par celui qui a la clé privée.

2. Tout le monde peut recevoir un message qui n'a pu être écrit que d'une seule personne (authentification de l'auteur, signature électronique).

Bien évidemment, on peut combiner ces deux utilités pour avoir un message qui ne peut être lu que par une seule personne et qui n'a pu être écrit que par une unique personne !

5.6 Le système de cryptographie RSA

Le plus célèbre et le premier des systèmes de cryptographie à clé publique est le système RSA (Ronald Rivest, Adi Shamir et Leonard Adleman). Entre autres, ce système est à la base des méthodes de paiements par Internet !

5.6.1 Mise en place

1. On choisit deux nombres premiers distincts p et q suffisamment grands. On calcule le produit $n = pq$. Généralement, on utilise des nombres premiers d'environ 300 chiffres (en 1024 bits, on forme des chaînes de longueur $2^{1024} \cong 1.79 \cdot 10^{308}$).
2. On choisit un nombre e premier à $\varphi(n)$, c'est-à-dire que $\text{pgcd}(e, \varphi(n)) = 1$.
3. On cherche un nombre $d \in \mathbb{N}$ qui correspond à l'inverse de e modulo $\varphi(n)$. Autrement dit, on cherche $d \in \mathbb{N}$ tel que $d \cdot e \equiv 1 \pmod{\varphi(n)}$.
   Le nombre d existe, car e est premier à $\varphi(n)$.

La clé privée

La clé privée est composée des nombres p, q et d.

La clé publique

Les nombres n et e sont mis à disposition dans un annuaire. Les nombres p, q et d sont gardés secrets.

5.6.2 Sûreté du système RSA

Les informations données dans la clé publique ne permettent pas de retrouver la clé privée, car il est actuellement impossible de trouver p et q si on connaît le nombre n en un temps raisonnable (pour autant que n soit suffisamment grand). Or, pour trouver d, il faut connaître $\varphi(n)$ qui ne peut être connu qu'à l'aide de p et de q.

En 1991, les laboratoires RSA ont créé le «RSA factoring challenge». Ils ont publié des grands nombres n qui étaient le produit de deux nombres premiers et ont offert des prix à ceux qui arrivaient à les factoriser. En 2001, le challenge a été étendu à des nombres de 576 bits à 2048 bits pour des récompenses allant de 10 000$ à 200 000$.

En 2007, le challenge fut arrêté. Pour plus d'informations, le lecteur consultera le site suivant.

`https://en.wikipedia.org/wiki/RSA_Factoring_Challenge`

5.6.3 Théorème RSA

Soit p et q deux nombres premiers distincts et $n = pq$. Si e est un nombre premier à $\varphi(n)$ et si d est son inverse modulo $\varphi(n)$, alors pour tout entier a ($a < n$), on a :

$$(a^e)^d \equiv (a^d)^e \equiv a \pmod{n}$$

Preuve

Puisque p et q sont des nombres premiers distincts, on sait que $n = \text{ppcm}(p, q)$. Ainsi, pour montrer que $a^{de} \equiv a \pmod{pq}$, il suffit de montrer que a^{de} est solution du système chinois suivant

$$\begin{cases} x \equiv a \pmod{p} \\ x \equiv a \pmod{q} \end{cases}$$

En effet, dans ce cas a^{de} serait solution du système, au même titre que a. Par unicité de la solution modulo pq, on saurait que $a^{de} \equiv a \pmod{pq}$.

On ne va démontrer que $a^{de} \equiv a \pmod{p}$, car pour l'autre, on reprend les mêmes arguments en échangeant les rôles de p et de q.

Par hypothèse, on sait que $de \equiv 1 \pmod{\varphi(n)}$. Ainsi, il existe $k \in \mathbb{Z}$ tel que

$$de = 1 + k\,\varphi(n) = 1 + k(p-1)(q-1)$$

Donc, comme $a^{p-1} \equiv 1 \pmod{p}$ grâce au théorème de Fermat, on obtient

$$a^{de} = a^{1+k(p-1)(q-1)} = a^1 \cdot a^{k(p-1)(q-1)} = a \cdot (a^{p-1})^{k(q-1)} \equiv a \cdot 1 \equiv a \pmod{p}$$

□

5.6.4 Méthode de codage et de décodage

Le processus de codage et de décodage se déroule en plusieurs étapes.

1. Préparation du message à encoder.

   Le message doit être éventuellement décomposé en blocs. À chaque bloc, on va associer un nombre strictement compris entre 1 et n. Ce sont ces nombres que l'on va encoder à l'aide du système RSA.

   La façon dont on associe des nombres à chaque bloc est très variable. Cela peut suivre une démarche assez simple (comme on le verra en exercice) ou un procédé bien plus complexe où un autre cryptage pourrait être utilisé.

2. Encodage avec RSA

   Le message est maintenant une suite de nombres strictement compris entre 1 et n. Pour coder ce message on élève chacun des nombres le composant à la puissance d ou à la puissance e selon si on veut encoder avec la clé privée ou publique.

3. Décodage avec RSA

   Le message est maintenant une suite de nombres strictement compris entre 1 et n. Pour décoder ce message on élève chacun des nombres le composant à la puissance e ou à la puissance d selon si on veut décoder avec la clé privée ou publique.

   Bien sûr, si le message a été encodé avec la clé privée, il faut le décoder avec la clé publique, et vice-versa.

   On retombe bien sur les nombres qu'on avait avant l'encodage, puisque le théorème nous dit que $(a^e)^d \equiv (a^d)^e \equiv a \pmod{n}$.

4. Reconstitution du message original.

   Il faut effectuer la démarche inverse de celle effectuée lors de la préparation du message à encoder.

Remarque banale mais importante

Lors de la préparation du message, il ne faut pas associer chaque lettre à un nombre (en code `ASCII`, entre 1 et 255), car une simple analyse de fréquences permettra de casser le code (bien sûr, si le message est suffisamment long pour qu'une telle analyse soit pertinente).

Chapitre 6

Résolution numérique d'équations

But

Utiliser des méthodes numériques (programmes informatiques) pour trouver les zéros d'une fonction f continue, c'est-à-dire les nombres x tels que $f(x) = 0$.

On va pour cela découvrir deux méthodes parmi de nombreuses méthodes existant sur le marché.

6.1 Méthode de la bissection

Cette méthode utilise le théorème de Bolzano.

Théorème de Bolzano

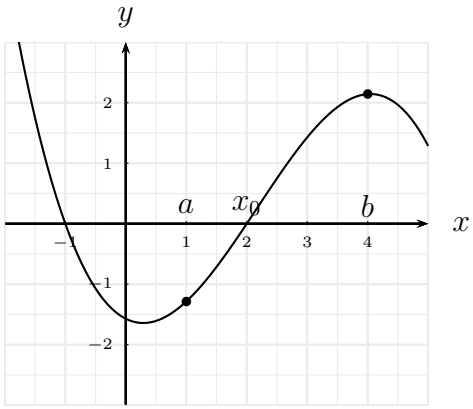
Soit a et b deux nombres réels tels que $a < b$.
Soit f une fonction réelle continue sur un intervalle $[a, b]$ satisfaisant la condition suivante qui est équivalente à dire que $f(a)$ et $f(b)$ sont de signes opposés.

$$f(a) \cdot f(b) < 0$$

Alors, il existe (au moins un) $x_0 \in]a, b[$ tel que $f(x_0) = 0$.

Illustration

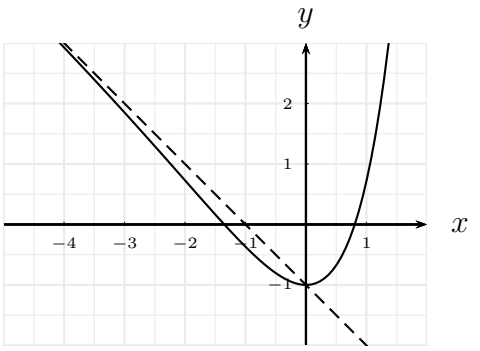
La fonction suivante est continue sur $\mathbb{R}$ et satisfait $f(1) \cdot f(4) < 0$. Elle admet bien un zéro dans l'intervalle $]1, 4[$.



La fin d'un rêve

Les mathématiciens ont toujours cherchés des formules explicites pour trouver les zéros des équations. Ils furent heureux d'en trouver pour résoudre les équations polynomiales de degré 1, 2 (Viète), 3 (Cardan) et 4 (Ferrari). Néanmoins, Galois a montré qu'il n'y avait aucune formule explicite permettant de résoudre celles de degré 5 ou plus.

Il existe aussi des fonctions dont les zéros ne peuvent pas être exprimé par une formule explicite. Par exemple, c'est le cas pour la fonction d'expression $f(x) = xe^x - (x+1)$ dont on voit le graphe ci-contre.



On a $f(-2) \cong 0.729329$, $f(0) = -1$. Donc, par le théorème de Bolzano, on a un zéro dans $]-2, 0[$. De même, comme $f(0) = -1$ et $f(1) \cong 0.718282$, il y a un zéro dans $]0, 1[$.

Approche méthodologique

Pour trouver les zéros d'une fonction, on va développer des méthodes itératives, c'est-à-dire construire des suites $(x_n)_{n\geqslant 1}$ qui vont converger vers un zéro, noté x_0.
En d'autres termes :

$$x_0 = \lim_{n\to+\infty} x_n$$

6.1.1 La méthode de la bissection et son algorithme

Soit a, $b \in \mathbb{R}$ tels que $a < b$. Soit f une fonction réelle continue sur $[a, b]$ qui change de signe entre a et b, c'est-à-dire qui satisfait :

$$f(a) \cdot f(b) < 0$$

Le théorème de Bolzano nous permet d'affirmer que f admet un zéro entre a et b. Voici un algorithme permettant de construire une suite $(x_n)_{n\geqslant 1}$ qui converge vers ce zéro.

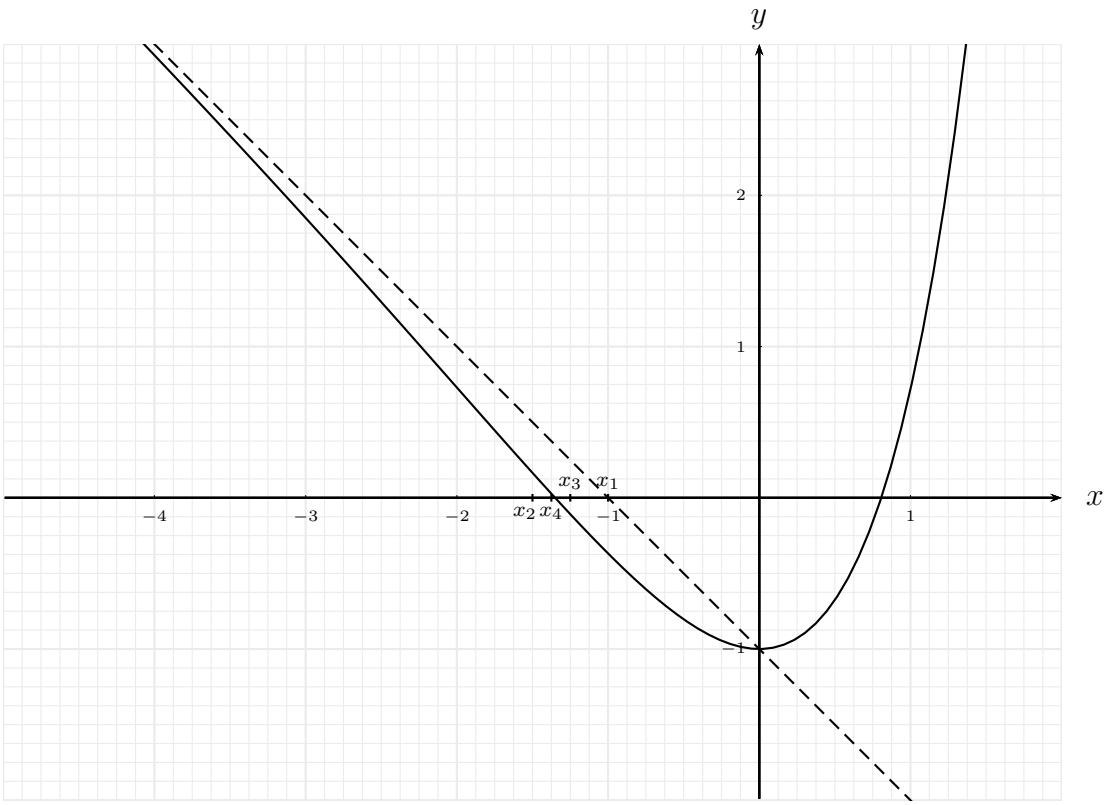
Principe À chaque itération, on coupe l'intervalle contenant un zéro en deux parties égales et on choisit celle où la fonction s'annule.

Algorithme On pose $\alpha_1 = a$, $\beta_1 = b$ et $x_1 = \frac{\alpha_1+\beta_1}{2}$ (x_1 est le milieu de l'intervalle $[a, b] = [\alpha_1, \beta_1]$). Puis, on distingue les trois cas suivants :

1. $f(x_1) = 0$ (ou $f(\alpha_1)f(x_1) = 0$). Ainsi, x_1 est un zéro. On peut arrêter de chercher.
2. $f(\alpha_1)f(x_1) < 0$. Ainsi, par Bolzano, on sait qu'un zéro se trouve entre α_1 et x_1. On choisit la partie gauche de l'intervalle coupé en deux en x_1. Le nouvel intervalle est $[\alpha_2, \beta_2]$ où $\alpha_2 = \alpha_1$ et $\beta_2 = x_1$.
3. $f(\alpha_1)f(x_1) > 0$. Cela signifie que le signe de $f(\alpha_1)$ est le même que celui de $f(x_1)$. Par conséquent, il y a un changement de signe entre x_1 et β_1. Par Bolzano, un zéro se trouve donc dans la partie droite de l'intervalle coupé en deux en x_1. Le nouvel intervalle est $[\alpha_2, \beta_2]$ où $\alpha_2 = x_1$ et $\beta_2 = \beta_1$.

Puis, on recommence avec l'intervalle $[\alpha_2, \beta_2]$ que l'on coupe en deux en $x_2 = \frac{\alpha_2+\beta_2}{2}$. Etc...

Interprétation graphique Appliquons cet algorithme à $f(x) = xe^x - (x+1)$ pour trouver quelques décimales du zéro se trouvant dans l'intervalle $[-2, 0]$. On prend $a = -2$ et $b = 0$. On a bien $f(a)f(b) < 0$.



On coupe l'intervalle $[\alpha_1, \beta_1] = [-2, 0]$ en $x_1 = -1$. On a $f(-2)f(-1) < 0$, on choisit donc la moitié de gauche qui est l'intervalle $[\alpha_2, \beta_2] = [-2, -1]$.

On coupe l'intervalle $[\alpha_2, \beta_2] = [-2, -1]$ en $x_2 = -1.5$. On a $f(-2)f(-1.5) > 0$, on choisit donc la moitié de droite qui est l'intervalle $[\alpha_3, \beta_3] = [-1.5, -1]$.

On coupe l'intervalle $[\alpha_3, \beta_3] = [-1.5, -1]$ en $x_3 = -1.25$. On a $f(-1.5)f(-1.25) < 0$, on choisit donc la moitié de gauche qui est l'intervalle $[\alpha_4, \beta_4] = [-1.5, -1.25]$.

On coupe $[\alpha_4, \beta_4] = [-1.5, -1.25]$ en $x_4 = -1.375$. On a $f(-1.5)f(-1.375) > 0$, on choisit donc la moitié de droite qui est l'intervalle $[\alpha_5, \beta_5] = [-1.375, -1.25]$.

On a ainsi $x_5 = -1.3125$.
En continuant, l'algorithme va choisir les moitiés de la manière suivante : gauche ; gauche ; droite ; droite ; gauche ; gauche ; droite ; droite ; gauche ; gauche ; droite ; droite ; droite ; droite ; droite ; droite ; gauche ; gauche ; gauche ; gauche ; droite ; droite ; droite ; gauche. Ce qui nous donnera $x_{30} = -1.34997649$.

Remarque

Cet algorithme ne permet de trouver qu'un zéro dans l'intervalle de départ $[a, b]$. Si on veut trouver un zéro précis, il faut s'arranger pour choisir a et b de telle manière que seul un zéro se trouve dans cet intervalle (ce que l'on peut facilement faire en regardant le graphe de la fonction).

6.1.2 Critère d'arrêt de l'algorithme

Puisqu'on calcule les éléments d'une suite $(x_n)_{n\geqslant 1}$ un par un, il faut décider d'un critère d'arrêt qui respecte une précision désirée (à moins que, par un coup de chance extraordinaire, l'algorithme s'arrête car il existe i tel que $f(x_i) = 0$).

Définition

Si $(x_n)_{n\geqslant 1}$ est une suite qui converge vers x_0. L'*erreur (absolue) au pas n* est définie par :

$$e_n = |x_n - x_0|$$

Théorème

Si on effectue la méthode de la bissection sur une fonction f continue sur l'intervalle $[a, b]$ (où $f(a)f(b) < 0$). Alors :

$$e_n < \frac{b-a}{2^n} \qquad \text{pour tout } n \geqslant 1$$

Preuve

La longueur de l'intervalle de départ $[a, b]$ est égale à $b - a$. Donc l'erreur au pas 1 est forcément plus petite que la moitié de cet intervalle. Autrement dit :

$$e_1 < \frac{b-a}{2}$$

Comme, à chaque itération de l'algorithme, on divise la longueur de l'intervalle (dans lequel le zéro cherché se trouve) par 2, la formule du théorème devient évidente (il faudrait la démontrer par récurrence pour être pédant). Précisons tout de même que dans le cas où il existe i tel que $f(x_i) = 0$, alors l'erreur au pas i vaut zéro. Ce qui reste compatible avec la formule. □

6.2 La méthode du point fixe

Définition

Soit $g : D \to A$ une fonction et $x_0 \in D$. On dit que x_0 est *un point fixe de* g si $g(x_0) = x_0$.

Autrement dit, les points fixes de g sont les solutions de l'équation $g(x) = x$.

Remarque fondamentale

Chercher un zéro x_0 d'une fonction f (c'est-à-dire résoudre l'équation $f(x) = 0$) revient à chercher un point fixe x_0 d'une fonction g bien choisie (c'est-à-dire résoudre l'équation $g(x) = x$). Pour que la fonction g soit bien choisie, il faut que :

$$f(x) = 0 \iff g(x) = x$$

La méthode du point fixe ne fonctionnera que si la fonction g est continue autour du point fixe cherché (qui est le zéro de f).

Exemples de fonction g bien choisie

Reprenons la fonction $f(x) = xe^x - (x+1)$. On a différents choix de fonctions g possible.

1. Premier choix possible : $g_1(x) = xe^x - 1$. En effet, on a :
$$f(x) = 0 \iff xe^x - (x+1) = 0 \iff \underbrace{xe^x - 1}_{g_1(x)} = x$$
2. Deuxième choix possible : $g_2(x) = (x+1)e^{-x}$. En effet, on a :
$$f(x) = 0 \iff xe^x - (x+1) = 0 \iff xe^x = x+1 \iff x = \underbrace{(x+1)e^{-x}}_{g_2(x)}$$

6.2.1 La méthode du point fixe et son algorithme

On construit de manière itérative une suite $(x_n)_{n\geqslant 1}$ qui converge vers un zéro x_0 de f.

1. On transforme l'équation $f(x) = 0$ en $g(x) = x$ avec g continue (au voisinage de x_0).
2. On choisit x_1 moralement proche de x_0.
3. On calcule successivement les éléments de la suite $(x_n)_{n\geqslant 1}$ à l'aide de la relation de récurrence $x_{n+1} = g(x_n)$ pour tout $n \geqslant 1$.

La suite a donc l'allure suivante.

$$\Big(\ x_1\ ,\ \underbrace{g(x_1)}_{x_2=g(x_1)}\ ,\ \underbrace{g(g(x_1))}_{x_3=g(x_2)}\ ,\ \underbrace{g(g(g(x_1)))}_{x_3=g(x_2)}\ ,\ \dots\ \Big)$$

Théorème de convergence

Si la suite $(x_n)_{n\geqslant 1}$ définie précédemment converge, alors elle converge vers un point fixe x_0 de g qui sera un zéro de f (voir remarque fondamentale).

Démonstration

On suppose que la suite $(x_n)_{n\geqslant 1}$ converge vers un nombre appelé x_0. Il faut montrer que x_0 est un point fixe de la fonction g (qui est supposée continue au voisinage de x_0).

Or, dire que x_n tend vers x_0 lorsque n tend vers $+\infty$ est équivalent à dire que la distance entre x_0 et x_n tend vers 0 quand n tend vers $+\infty$. En d'autres termes :

$$x_n \text{ converge vers } x_0 \iff |x_0 - x_n| \overset{n\to+\infty}{\longrightarrow} 0$$

En utilisant l'inégalité triangulaire ($|x+y| \leqslant |x| + |y|$), on montre que :

$$\begin{aligned} |x_0 - g(x_0)| &= |x_0 - x_{n+1} + x_{n+1} - g(x_0)| \\ &\leqslant |x_0 - x_{n+1}| + |x_{n+1} - g(x_0)| \\ &= \underbrace{|x_0 - x_{n+1}|}_{\overset{n\to+\infty}{\longrightarrow} 0} + \underbrace{|g(x_n) - g(x_0)|}_{\overset{n\to+\infty}{\longrightarrow} 0 \text{ car } g \text{ est continue}} \end{aligned}$$

Donc $|x_0 - g(x_0)| = 0$, c'est-à-dire que $x_0 - g(x_0) = 0$ ou encore que $g(x_0) = x_0$. □

Exemple

On reprend la fonction $f(x) = xe^x - (x+1)$ avec les deux choix de g précédemment vus.

1. Avec $g_1(x) = xe^x - 1$.

   Si on prend $x_1 = -5$, on a $x_2 = g_1(x_1) \cong -1.03369$, puis $x_3 = g_1(x_2) \cong -1.36768$, $x_4 \cong -1.34834$, $x_5 \cong -1.35012$, …, $x_{10} \cong -1.34998$, …, $x_{20} \cong -1.34998$.
   On trouve ainsi le zéro de gauche.

   Si on prend $x_1 = 0.8$, on a $x_2 = g_1(x_1) \cong 0.78043$, puis $x_3 = g_1(x_2) \cong 0.70323$, $x_4 \cong 0.42071$, $x_5 \cong -0.35924$, …, $x_{10} \cong -1.34997$, …, $x_{20} \cong -1.34998$.
   C'est de nouveau le zéro de gauche !

   Si on prend $x_1 = 0.85$, on a $x_2 = g_1(x_1) \cong 0.98870$, puis $x_3 = g_1(x_2) \cong 1.65737$, $x_4 \cong 7.69367$, $x_5 \cong 166882.1$, $x_6 > 10^{499}$.
   On voit que la suite diverge et on ne trouve pas le zéro de droite.

2. Avec $g_2(x) = (x+1)e^{-x}$.

   Si on prend $x_1 = 0.8$, on trouve $x_2 = g_2(x_1) \cong 0.80879$, puis $x_3 = g_2(x_2) \cong 0.80563$, $x_4 \cong 0.80677$, $x_5 \cong 0.80636$, …, $x_{10} \cong 0.80647$, …, $x_{20} \cong 0.80647$.
   C'est le zéro de droite.

   Si on prend $x_1 = -1.3$, on trouve $x_{20} \cong 0.80647$. C'est de nouveau le zéro de droite !

   Si on prend $x_1 = -1.4$, on trouve $x_6 < -10^{499}$. On voit que la suite diverge et on ne trouve pas le zéro de gauche.

Moralité la fonction g_1 permet de trouver le zéro de gauche, mais pas celui de droite, tandis que la fonction g_2 permet de trouver le zéro de droite, mais pas celui de gauche.

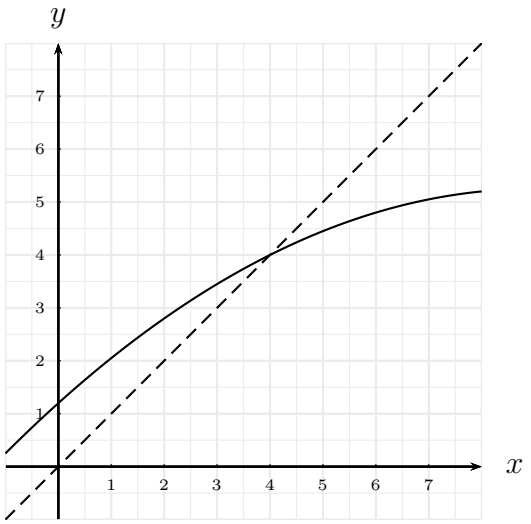
Théorème de sélection

Soit $g : \mathbb{R} \to \mathbb{R}$ une fonction dont la dérivée est continue. Supposons que g admet un point fixe x_0 qui satisfait $|g'(x_0)| < 1$.

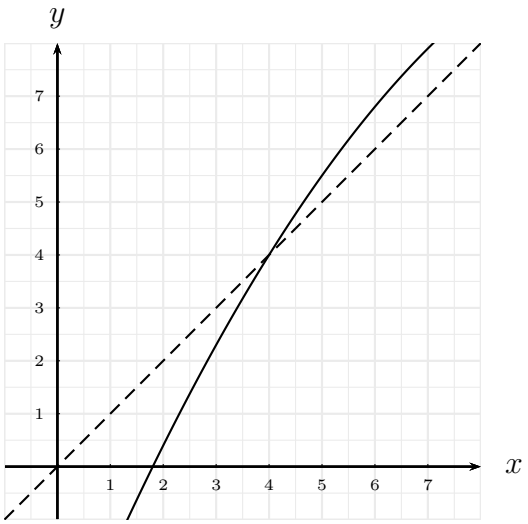
Alors, si on choisit x_1 suffisamment proche de x_0, la suite des approximations successives $x_{n+1} = g(x_n)$ converge vers x_0.

Interprétation graphique

cas $g'(x_0) < 1$



cas $g'(x_0) > 1$



Moralité Le choix de la fonction g est très important !

Définition

Soit I un intervalle dans $\mathbb{R}$. Une fonction $g : I \to I$ est dite *contractante* si pour tout x, y dans I, il existe un nombre $k \in [0, 1[$ (indépendant de x et de y) qui satisfait :

$$\underbrace{|g(x) - g(y)|}_{\text{distance entre } g(x) \text{ et } g(y)} \leqslant k \underbrace{|x - y|}_{\text{distance entre } x \text{ et } y}$$

Moralement : Cela signifie que la fonction resserre tous les points de I.

Théorème du point fixe de Banach [1]

Soit I un intervalle fermé de $\mathbb{R}$ et $g : I \to I$ une fonction contractante. Alors g admet un unique point fixe dans l'intervalle I.

Remarque La démonstration fait appel aux suites de Cauchy.

Proposition

Soit I un intervalle fermé et $g : I \to I$ une fonction contractante. Alors pour n'importe quel point $x_1 \in I$, la suite des approximations successives $x_{n+1} = g(x_n)$ converge (donc converge vers un point fixe de g (voir théorème de convergence)).

1. Ce théorème se généralise aux espaces $\mathbb{R}^n$ et permet ainsi d'appliquer le même théorème à la construction des fractales par MCRM. Dans ce contexte, le théorème est rebaptisé : théorème de Banach-Hausdorff.

Preuve de la proposition

Par le théorème du point fixe de Banach, on sait que la fonction g admet un unique point fixe dans I, que l'on note x_0. Soit x_1 un point quelconque de l'intervalle I et montrons que la suite $(x_n)_{n\geqslant 1}$, définie par $x_{n+1} = g(x_n)$ pour tout $n \geqslant 1$, converge vers ce point fixe x_0.

C'est le cas, car l'erreur [2] au pas n, donnée par $e_n = |x_n - x_0|$, diminue de la même façon à chaque itération de l'algorithme du point fixe. En effet :

$$e_{n+1} = |x_{n+1} - x_0| = |g(x_n) - g(x_0)| \overset{g \text{ contractante}}{\leqslant} k|x_n - x_0| = k\, e_n$$

On a ainsi la relation $e_{n+1} \leqslant k^n e_1$ avec $k \in [0, 1[$ (le nombre k provient de la définition de fonction contractante).

Donc, lorsque $n \to +\infty$, on a $k^n \to 0$ (car $0 \leqslant k < 1$) et ainsi $e_{n+1} \to 0$. Comme l'erreur tend vers 0, la suite converge vers x_0. □

Idée de preuve du théorème de sélection

Soit I un intervalle fermé de $\mathbb{R}$ et $g : I \to I$ une fonction dont la dérivée est continue et satisfait la condition ★ : $|g'(x)| \leqslant k < 1$ pour tout $x \in I$.

Alors, on peut montrer (grâce au théorème des accroissements finis) que g est contractante. Ainsi, grâce à la proposition précédente, la suite des approximations successives $x_{n+1} = g(x_n)$ converge vers un point fixe $x_0 \in I$.

Le lecteur attentif aura remarqué que dans le théorème de sélection, on ne parle ni d'un intervalle fermé noté I, ni de la condition ★.

En effet, c'est pour s'assurer l'existence d'un tel intervalle fermé I que l'on se doit de choisir x_1 suffisamment proche de x_0 ($x_0 \in I$). De même, la condition ★ est automatiquement satisfaite pour des x proches de x_0 si on a $|g'(x_0)| < 1$.

6.2.2 La méthode de Newton-Raphson

Il s'agit d'une des méthodes les plus utilisées pour trouver les zéros d'une fonction. La théorie de la méthode du point fixe est reprise telle quelle. L'astuce de Newton-Raphson est d'avoir réussi à trouver une fonction g qui fonctionne toujours !

On suppose que la fonction f dont on cherche les zéros satisfait les conditions : f'' est continue et le zéro cherché x_0 est simple (c'est-à-dire que $f'(x_0) \neq 0$). Remarquons, que dans la pratique, la méthode fonctionne même si le zéro cherché n'est pas simple. On a :

$$f(x) = 0 \iff x - \frac{f(x)}{f'(x)} = x$$

On prend donc $g(x) = x - \dfrac{f(x)}{f'(x)}$. On a :

$$g'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2}$$

Comme $g'(x_0) = 0$ (car $f(x_0) = 0$), le théorème de sélection s'applique et ainsi pour x_1 suffisamment proche de x_0, la suite des approximations successives $x_{n+1} = g(x_n)$ converge vers x_0.

2. Il s'agit de la même définition que celle se trouvant dans la méthode de la bissection.

Interprétation graphique : la méthode de la sécante

L'idée de Newton et de Raphson est de construire une suite $(x_i)_{i\geqslant 1}$, à partir d'un nombre x_1 moralement proche de x_0, de manière à ce que x_{i+1} soit le zéro de la tangente à la fonction f en x_i.

L'équation de la tangente en x_i est

$$y = f(x_i) + f'(x_i)(x - x_i)$$

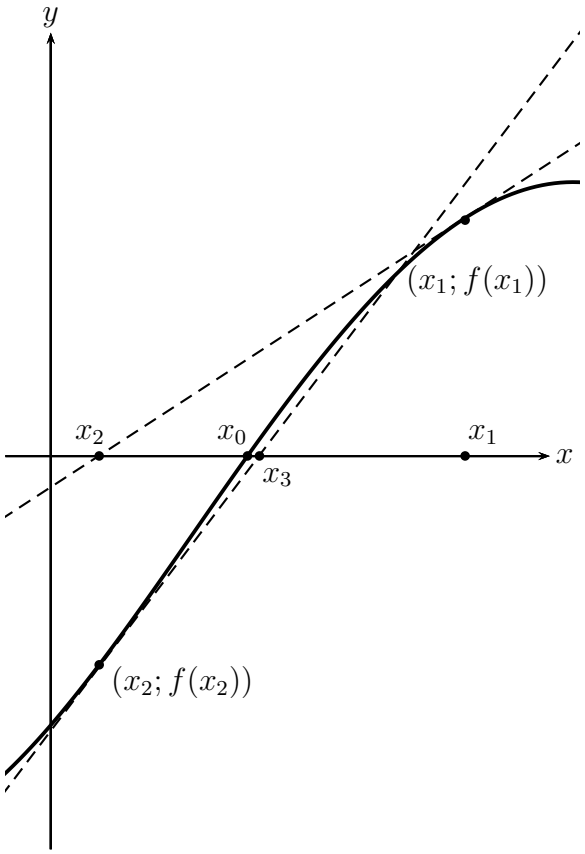
Si la pente de la tangente n'est pas nulle (c'est-à-dire si $f'(x_i) \neq 0$), on peut calculer son zéro x_{i+1}.

$$0 = f(x_i) + f'(x_i)(x_{i+1} - x_i)$$

$$\iff x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Ainsi, on reconnaît un algorithme équivalent à la méthode du point fixe pour

$$g(x) = x - \frac{f(x)}{f'(x)}$$



6.2.3 Critère d'arrêt pour la méthode du point fixe

Lorsque la fonction g satisfait les hypothèses du théorème de sélection, on a vu (dans la preuve de la proposition) que l'erreur au pas n, donnée par $e_n = |x_n - x_0|$, diminue à chaque itération de l'algorithme du point fixe (ou de Newton-Raphson qui est un cas particulier de la méthode du point fixe). En effet, dans la preuve de cette proposition, on trouvait la ligne suivante :

$$e_{n+1} = |x_{n+1} - x_0| \leqslant k|x_n - x_0| = k\, e_n \quad \text{avec} \quad k \in [0, 1[$$

Par conséquent, l'erreur entre deux termes successifs de la suite des approximations successives $x_{n+1} = g(x_n)$ devient de plus en plus petite. En effet, cette erreur s'exprime de la manière suivante grâce à l'inégalité triangulaire :

$$|x_{n+1} - x_n| = |x_{n+1} - x_0 + x_0 - x_n| \leqslant |x_{n+1} - x_0| + |x_0 - x_n| \leqslant e_{n+1} + e_n \leqslant ke_n + e_n < 2e_n$$

Donc, comme $e_n \to 0$ de manière strictement décroissante, on a $|x_{n+1} - x_n| \to 0$ de manière strictement décroissante.

Par conséquent, on choisit le critère d'arrêt suivant pour l'algorithme. Dès que la différence entre deux éléments consécutifs de la suite des itérés est plus petite qu'une certaine tolérance (fixée à l'avance), on stoppe l'algorithme.

Chapitre 7

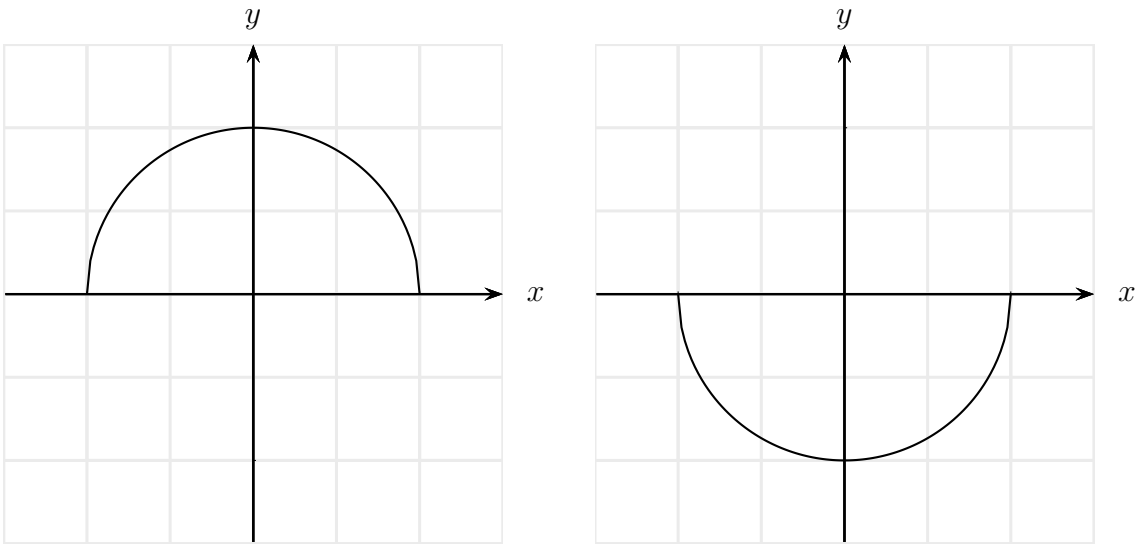
Courbes paramétrées

7.1 Introduction

Les limitations des fonctions

Lorsqu'on désire dessiner des courbes dans le plan, on a envie d'avoir la possibilité de dessiner des courbes qui ne respectent pas le test de la droite verticale cher aux fonctions. C'est-à-dire, une courbe qui peut couper une droite verticale plusieurs fois (ce qu'une fonction ne doit jamais faire).

Par exemple, le cercle trigonométrique ne peut pas être décrit comme étant le graphe d'une fonction. On aurait besoin de deux fonctions au minimum.



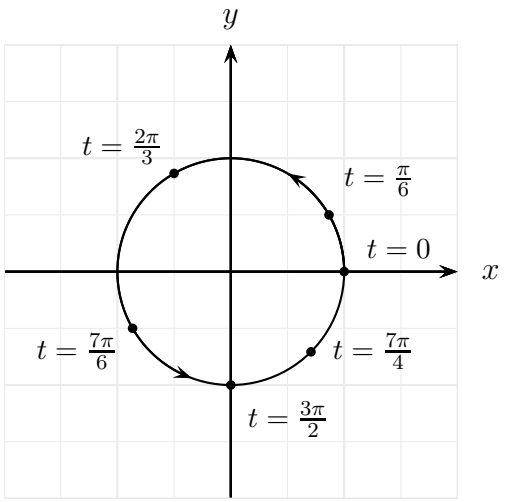
$f(x) = \sqrt{1-x^2}$ $\qquad$ $f(x) = -\sqrt{1-x^2}$

avec $x \in [-1, 1]$

Courbes paramétrées

On sait que le cercle trigonométrique est l'ensemble des points du plan qui se trouvent à distance 1 de l'origine du plan. On sait aussi qu'un point du cercle a les coordonnées $(\cos(t); \sin(t))$.

Ainsi pour dessiner le cercle trigonométrique, on dessine les points $(\cos(t); \sin(t))$ en faisant varier t dans $[0, 2\pi[$. On obtient une fonction f dont la variable t vit dans le domaine de définition $[0, 2\pi[$ et dont le domaine d'arrivée est le plan $\mathbb{R}^2$ (le domaine image est le cercle trigonométrique).



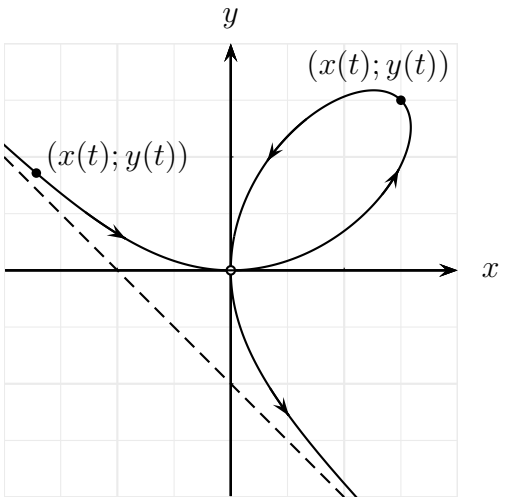
$$f : [0, 2\pi[\to \mathbb{R}^2;\ t \mapsto (\cos(t); \sin(t))$$

Définitions

On considère deux fonctions réelles

$$x : D_x \to \mathbb{R}$$
et
$$y : D_y \to \mathbb{R}$$

La fonction ci-dessous est appelée *fonction paramétrée*. Son domaine de définition est évidemment $D_f = D_x \cap D_y$.



$$f : D_f \to \mathbb{R}^2;\ t \mapsto (x(t); y(t))$$

La *courbe paramétrée* associée est l'ensemble des points du plan suivant.

$$\mathcal{C}_f = \{(x(t); y(t)) : t \in D_f\}$$

Cas particuliers

1. **Les droites sont des courbes paramétrées.**

   Une droite est une courbe paramétrée dont les fonctions x et y sont données par les équations paramétriques de cette droite.

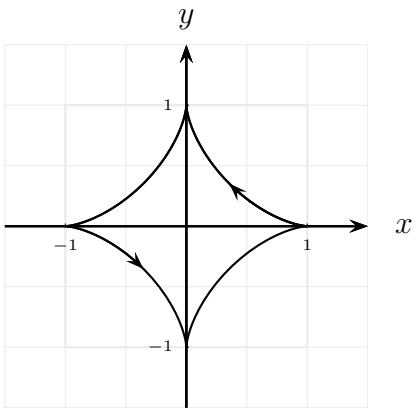
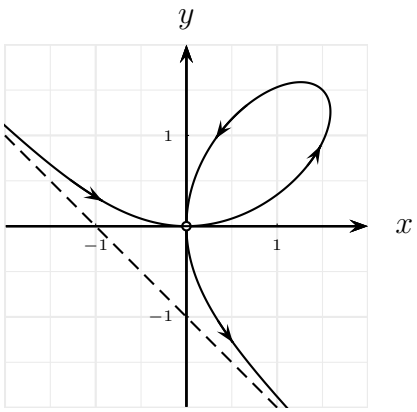
$$d : \begin{cases} x = x_0 + \lambda d_1 \\ y = y_0 + \lambda d_2 \end{cases},\ \lambda \in \mathbb{R} \quad \leftrightsquigarrow \quad \begin{array}{rcl} f : \mathbb{R} & \to & \mathbb{R}^2 \\ t & \mapsto & (x(t); y(t)) \end{array} \text{ où } \begin{cases} x(t) = x_0 + t d_1 \\ y(t) = y_0 + t d_2 \end{cases}$$

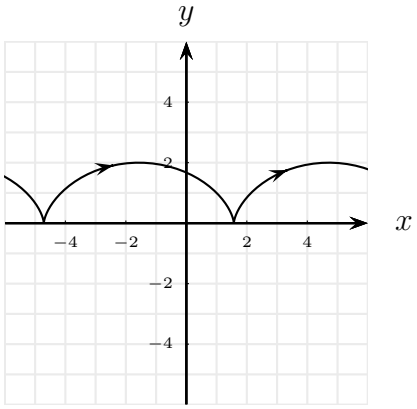
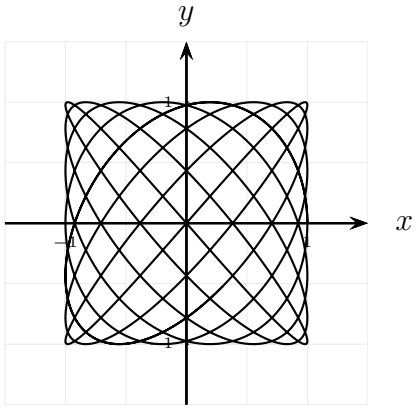
2. **Les fonctions réelles sont des courbes paramétrées.**

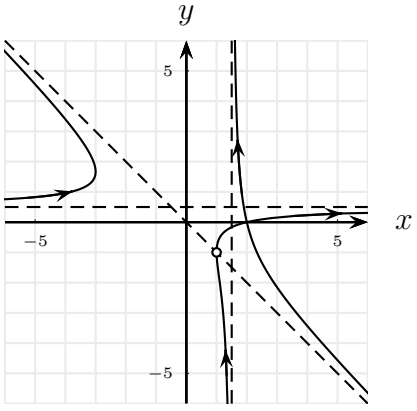
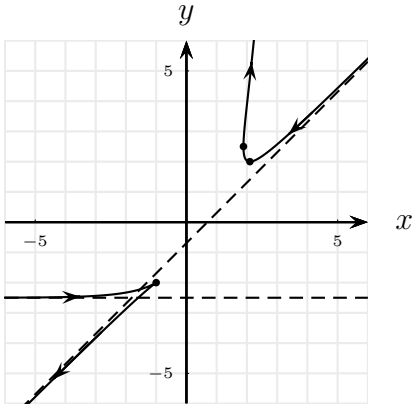
   Soit $f : D \to \mathbb{R}$ une fonction réelle (c'est-à-dire D, $A \subset \mathbb{R}$). Alors f est aussi une fonction paramétrée.

$$\begin{array}{rcl} f : D & \to & \mathbb{R} \\ x & \mapsto & f(x) \end{array} \quad \leftrightsquigarrow \quad \begin{array}{rcl} f : D & \to & \mathbb{R}^2 \\ t & \mapsto & (x(t); y(t)) \end{array} \text{ où } \begin{cases} x(t) = t \\ y(t) = f(t) \end{cases}$$

Exemples de courbes paramétrées

Astroïde	Folium de Descartes
$\{(\cos^3(t);\sin^3(t)) : t \in [0,2\pi[\,\}$	$\{(\frac{3t}{1+t^3};\frac{3t^2}{1+t^3}) : t \in \mathbb{R}\setminus\{-1\}\}$
	

Cycloïde	Une courbe de Lissajou
$\{(t+\cos(t);1-\sin(t)) : t \in \mathbb{R}\}$	$\{(\sin(7t+\frac{\pi}{2});\sin(8t)) : t \in \mathbb{R}\}$
	

Courbe de la page 66	Courbe de la page 67
$\{(\frac{t^2+t+1}{t(t+1)};\frac{t^2+t-1}{t(1-t)}) : t \in \mathbb{R}\setminus\{-1,0,1\}\}$	$\{(\frac{\ln(t+2)t+1}{t};\frac{t^2+1}{t}) : t \in \,]-2,+\infty[\setminus\{0\}\}$
	

7.2 Asymptotes obliques et verticales

Asymptotes verticales

On constate l'existence d'une asymptote verticale d'équation $x = x_0$ lorsqu'il existe $t_0 \in \mathbb{R} \cup \{\pm\infty\}$ tel que[1]

$$\lim_{t\to t_0} x(t) = x_0 \quad \text{et} \quad \lim_{t\to t_0} y(t) = \pm\infty$$

Asymptotes horizontales

On constate l'existence d'une asymptote horizontale d'équation $y = y_0$ lorsqu'il existe $t_0 \in \mathbb{R} \cup \{\pm\infty\}$ tel que[1]

$$\lim_{t\to t_0} x(t) = \pm\infty \quad \text{et} \quad \lim_{t\to t_0} y(t) = y_0$$

Asymptotes obliques

On constate que si la courbe paramétrée admet une asymptote oblique, alors il existe $t_0 \in \mathbb{R} \cup \{\pm\infty\}$ tel que[1]

$$\lim_{t\to t_0} x(t) = \pm\infty \quad \text{et} \quad \lim_{t\to t_0} y(t) = \pm\infty$$

Attention : il est possible qu'il existe un tel t_0 sans que la courbe admette une asymptote oblique.

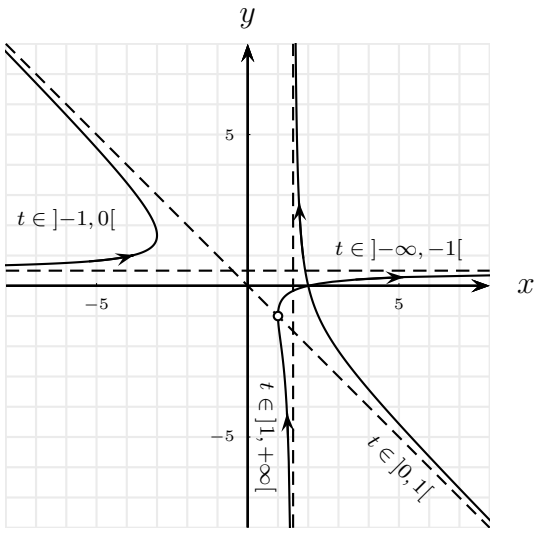
Théorème Si la fonction f admet une asymptote oblique d'équation $d(x) = mx + h$, alors

$$\boxed{m = \lim_{t\to t_0} \frac{y(t)}{x(t)}} \quad \text{et} \quad \boxed{h = \lim_{t\to t_0} \big(y(t) - mx(t)\big)}$$

Exemple

La courbe paramétrée ci-contre est décrite par la fonction paramétrée suivante.

$$\begin{array}{rccl} f: & \mathbb{R}\setminus\{-1,0,1\} & \to & \mathbb{R}^2 \\ & t & \mapsto & \left(\frac{t^2+t+1}{t(t+1)}; \frac{t^2+t-1}{t(1-t)}\right) \end{array}$$



On a

1. une asymptote horizontale en $t = -1$. Son équation est $y = \frac{1}{2}$,
2. une asymptote oblique en $t = 0$. Son équation est $y = -x$.
3. une asymptote verticale en $t = 1$. Son équation est $x = \frac{3}{2}$,

1. On utilise la vision d'Alexandrov de la droite réelle (voir cours DF, page 169).

7.3 Pente en un point et points particuliers

La pente de la tangente à une courbe paramétrée en un point

Comme pour la dérivée, on calcule la pente moyenne entre les points correspondant à t et à $t+\Delta t$ et on fait tendre Δt vers 0.

Notons $m(t)$ la pente de la tangente à la courbe paramétrée $\mathcal{C}$ en t. On a

$$\begin{aligned} m(t) &= \lim_{\Delta t\to 0} \frac{y(t+\Delta t)-y(t)}{x(t+\Delta t)-x(t)} = \lim_{\Delta t\to 0} \frac{y(t+\Delta t)-y(t)}{\Delta t}\cdot\frac{\Delta t}{x(t+\Delta t)-x(t)} \\ &= \lim_{\Delta t\to 0} \frac{y(t+\Delta t)-y(t)}{\Delta t}\cdot \lim_{\Delta t\to 0}\frac{\Delta t}{x(t+\Delta t)-x(t)} = y'(t)\cdot\frac{1}{x'(t)} \end{aligned}$$

Donc

$$\boxed{m(t)=\frac{y'(t)}{x'(t)}}$$

Point à tangente horizontale

Un point de la courbe $(x(t_0),y(t_0))$ est à tangente horizontale si $x'(t_0)\neq 0$ et $y'(t_0)=0$.

Point à tangente verticale

Un point de la courbe $(x(t_0),y(t_0))$ est à tangente verticale si $x'(t_0)=0$ et $y'(t_0)\neq 0$.

Point singulier

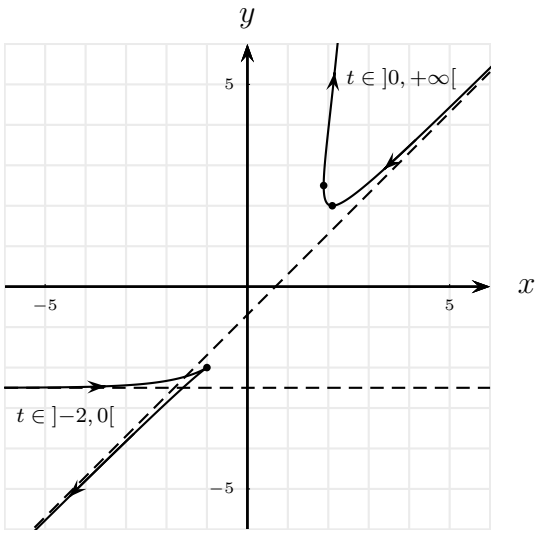
Un point de la courbe $(x(t_0),y(t_0))$ est appelé *point singulier* si $x'(t_0)=0$ et $y'(t_0)=0$.

Important : il faut toujours calculer $\lim\limits_{t\to t_0} m(t)$ pour un point singulier afin de pouvoir le dessiner correctement.

Exemple

La courbe paramétrée ci-contre est décrite par la fonction paramétrée suivante.

$$\begin{array}{rccl} f: &]-2,+\infty[\setminus\{0\} & \to & \mathbb{R}^2 \\ & t & \mapsto & (\frac{\ln(t+2)t+1}{t};\frac{t^2+1}{t}) \end{array}$$



On a

1. un point singulier en $t=-1$. Ce point est $(-1,-2)$. On a $\lim\limits_{t\to -1} m(t)=\frac{2}{3}$.
2. un point non singulier à tangente horizontale lorsque $t=1$. Ce point est $(\ln(3)+1;2)$.
3. un point non singulier à tangente verticale en $t=2$. Ce point est $(\ln(4)+\frac{1}{2};\frac{5}{2})$,

7.4 Étude de fonction paramétrique

Marche à suivre

Étudions la fonction $f : D \to \mathbb{R}^2;\ t \mapsto (\frac{1}{t(t+2)}; \frac{e^{-t}}{t})$.

1. **Calcul des limites au bord du domaine de définition et des asymptotes**
   Le domaine de définition est $D = D_x \cap D_y$. Ici $D = \mathbb{R} \setminus \{-2, 0\}$.
   (a) Pour $t \to -\infty$, on a une asymptote verticale d'équation $x = 0$.
   (b) Pour $t \to -2$, on a une asymptote horizontale d'équation $y = -\frac{e^2}{2}$.
   (c) Pour $t \to 0$, on a une asymptote oblique d'équation $y = 2x - \frac{1}{2}$.
   (d) Pour $t \to +\infty$, on a $(x(t), y(t)) \to (0, 0)$. Il y a un trou en $(0; 0)$.

2. **Dérivées et tableau de variation**

   On factorise les dérivées $x'(t)$ et $y'(t)$ pour le tableau de variation.

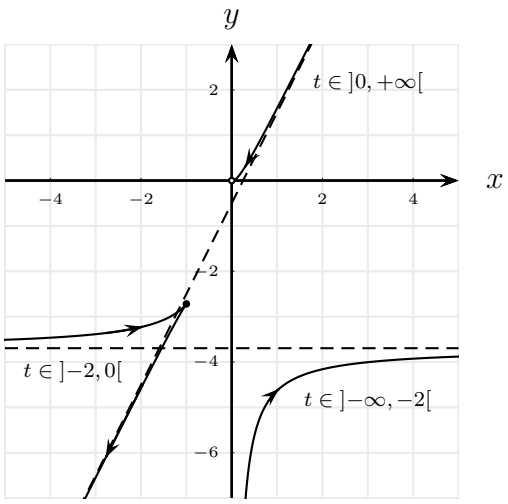
$$x'(t) = \frac{-2(t+1)}{t^2(t+2)^2} \quad \text{et} \quad y'(t) = \frac{-(t+1)}{t^2 e^t} \quad \text{donc} \quad m(t) = \frac{y'(t)}{x'(t)} = \frac{(t+2)^2}{2e^t}$$

   Contrairement aux études de fonctions, ici on fait le tableau de signes uniquement pour les dérivées. On remplit le reste en se basant sur le point 1.

t	$-\infty$		-2		-1		0		$+\infty$
$x'(t)$	0	$+$	↯	$+$	0	$-$	↯	$-$	0
$y'(t)$	$+\infty$	$+$	$+$	$+$	0	$-$	↯	$-$	0
$x(t)$	0	$\to$	$\pm\infty$	$\to$	-1	$\leftarrow$	$\pm\infty$	$\leftarrow$	0
$y(t)$	$\pm\infty$	$\uparrow$	$-\frac{e^2}{2}$	$\uparrow$	$-e$	$\downarrow$	$\pm\infty$	$\downarrow$	0
comportement	A.V.	↗	A.H.	↗	point sing.	↙	A.O.	↙	trou

3. **Points particuliers**
   Les points particuliers sont :
   (a) Pour $t = -1$, on a le point singulier $(-1; -e)$. On a $\lim\limits_{t \to -1} m(t) = \frac{e}{2}$.
   (b) Pour $t \to +\infty$, on a le point singulier $(0; 0)$. On a $\lim\limits_{t \to +\infty} m(t) = 0$.

4. **Graphe**



Chapitre 8

Fractales

L'introduction ci-dessous est inspirée du livre «Introducing Fractal Geometry» de Nigel Lesmoir-Gordon, Will Rood et Ralph Edney.

8.1 Introduction

La plupart des formes de la nature sont dynamiques, elles se distinguent de la géométrie fixe et statique de l'Homme dans la mesure où elles se développent et évoluent dans le temps. Ces structures en développement sont apparemment dictées par le chaos, comme par exemples les turbulences (prévisions météorologiques, simulation de courants marins, fumée de cigarettes), la forme d'un éclair, la structure d'un arbre, d'une fougère, de nos poumons, de notre système sanguin, le cours d'une action à la bourse, les mouvements browniens, les feux de forêts, la structure des flocons de neige. Des paysages imaginaires peuvent aussi être créés à l'aide de fractales.

Les fractales sont des objets mathématiques très variés tous construits à partir d'un processus itératif. Elles sont utilisées à des fins de simulations pour tenter de comprendre et de faire des prévisions à propos des structures en développement citées ci-dessus.

Dans un avenir proche ces modèles pourraient permettre de réduire les risques de crises cardiaques, de détecter un cancer (comme les cancers du sein) ou la fin de la période d'incubation du virus du SIDA [1]. Des modèles basés sur les fractales sont déjà utilisés pour soigner les os fragiles. La géométrie fractale est utilisée efficacement pour trouver des objets créés par l'Homme à partir de photos prises depuis les satellites (comme des sous-marins). Les tremblements de terre possèdent une signature fractale, tout comme les épidémies. Les images peuvent aussi être compressées en utilisant des fractales.

Le chou romanesco est un exemple classique de structure fractale se trouvant dans la nature.

1. Beaucoup de malades du SIDA restent séropositifs une dizaine d'années avant que le virus se réveille.

8.2 Fractalisation dans le plan

Pour créer des fractales tels que l'ensemble de Cantor, la courbe de Von Koch, les fractales de Sierpinski, on a besoin d'effectuer des transformations dans le plan.

	Transformations affines du plan	
	Transformations linéaires	Transformations non linéaires
Similitudes	Les rotations Les homothéties Les symétries	Les translations
	Les projections Les étirements	

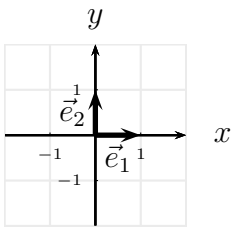
Les transformations linéaires sont effectuées grâces à des matrices[2].

8.2.1 Les applications affines et les matrices

On dit que $f : \mathbb{R}^2 \to \mathbb{R}^2$ est une transformation *linéaire* si :

1. $f(\vec{v}_1 + \vec{v}_2) = f(\vec{v}_1) + f(\vec{v}_2)$ pour tout vecteurs $\vec{v}_1$, $\vec{v}_2$ dans $\mathbb{R}^2$.
2. $f(\lambda\vec{v}) = \lambda f(\vec{v})$ pour tout vecteur $\vec{v} \in \mathbb{R}^2$ et tout $\lambda \in \mathbb{R}$.

Dans le reste du cours, on utilisera la base canonique $\{\vec{e}_1, \vec{e}_2\}$.



Ces vecteurs de base nous permettent de décrire chaque vecteur du plan comme unique combinaison linéaire de $\vec{e}_1$ et $\vec{e}_2$. Par exemple, le vecteur $\overrightarrow{OP}$ reliant l'origine $O(0;0)$ au point $P(x;y)$ est décrit de la manière suivante.

$$\overrightarrow{OP} = x\vec{e}_1 + y\vec{e}_2 \overset{\text{notation}}{=} \begin{pmatrix} x \\ y \end{pmatrix}$$

Considérons maintenant le vecteur $\vec{v} = \lambda_1\vec{e}_1 + \lambda_2\vec{e}_2$, aussi noté $\vec{v} = \binom{\lambda_1}{\lambda_2}$, et regardons comment une transformation linéaire f agit sur ce vecteur. Par linéarité on a :

$$f(\vec{v}) = \lambda_1 f(\vec{e}_1) + \lambda_2 f(\vec{e}_2)$$

Ainsi, il suffit de connaître $f(\vec{e}_1)$ et $f(\vec{e}_2)$ pour pouvoir connaître l'image de n'importe quel vecteur par la fonction f. Or $f(\vec{e}_1)$ et $f(\vec{e}_2)$ sont des vecteurs qui s'écrivent aussi dans la base canonique $\{\vec{e}_1, \vec{e}_2\}$. Disons que

$$f(\vec{e}_1) = a_{1,1}\vec{e}_1 + a_{2,1}\vec{e}_2 \overset{\text{notation}}{=} \begin{pmatrix} a_{1,1} \\ a_{2,1} \end{pmatrix} \quad \text{et} \quad f(\vec{e}_2) = a_{1,2}\vec{e}_1 + a_{2,2}\vec{e}_2 \overset{\text{notation}}{=} \begin{pmatrix} a_{1,2} \\ a_{2,2} \end{pmatrix}$$

Regardons comment on peut écrire le vecteur $f(\vec{v})$ dans la base canonique.

$$f(\vec{v}) = \lambda_1 f(\vec{e}_1) + \lambda_2 f(\vec{e}_2) \overset{\text{notation}}{=} \lambda_1 \begin{pmatrix} a_{1,1} \\ a_{2,1} \end{pmatrix} + \lambda_2 \begin{pmatrix} a_{1,2} \\ a_{2,2} \end{pmatrix} = \begin{pmatrix} a_{1,1}\lambda_1 + a_{1,2}\lambda_2 \\ a_{2,1}\lambda_1 + a_{2,2}\lambda_2 \end{pmatrix}$$

2. Le programme de troisième année reviendra en détail sur le sujet.

Ainsi, un vecteur est décrit par deux nombres et une application linéaire par quatre nombres. Une idée géniale a été d'utiliser une notation matricielle pour décrire les transformations linéaires.

Notation vectorielle	$\vec{v}$	f	$f(\vec{v})$
Notation matricielle	$\vec{v} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$	$A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix}$	$A\vec{v} = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$

Cela permet de définir la multiplication matrice-vecteur.

$$f(\vec{v}) \overset{\text{notation}}{=} A\vec{v} = \left(\begin{array}{|cc|}\hline a_{1,1} & a_{1,2} \\ \hline a_{2,1} & a_{2,2} \\ \hline \end{array}\right) \left(\begin{array}{|c|}\hline \lambda_1 \\ \hline \lambda_2 \\ \hline \end{array}\right) = \left(\begin{array}{|c|}\hline a_{1,1}\lambda_1 + a_{1,2}\lambda_2 \\ \hline a_{2,1}\lambda_1 + a_{2,2}\lambda_2 \\ \hline \end{array}\right)$$

On remarque qu'avec cette notation, on a

$$A = \left(\begin{array}{|c|c|}\hline a_{1,1} & a_{1,2} \\ \hline a_{2,1} & a_{2,2} \\ \hline \end{array}\right) = \left(\begin{array}{|c|c|}\hline f(\vec{e}_1) & f(\vec{e}_2) \\ \hline \end{array}\right)$$

Cela nous permet d'énoncer la *règle pour la construction de la matrice A* associée à la transformation f :

LES COLONNES DE LA MATRICE SONT LES IMAGES DES VECTEURS DE BASE

8.2.2 Addition de transformations linéaires

Soit f et g deux transformations linéaires du plan. Notons A et B les matrices associées respectivement à f et à g.

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \quad \text{et} \quad B = \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix}$$

La matrice associée à la transformation $f + g$ est donnée par l'*addition des matrices A* et B :

$$A + B = \begin{pmatrix} a_{1,1} + b_{1,1} & a_{1,2} + b_{1,2} \\ a_{2,1} + b_{2,1} & a_{2,2} + b_{2,2} \end{pmatrix}$$

Preuve

En effet, comme les colonnes de la matrice sont les images des vecteurs de base, il suffit de calculer les images de $\vec{e}_1$ et de $\vec{e}_2$ par l'application $f + g$. Grâce à la règle pour la construction des matrices A et B, on a :

$$f(\vec{e}_i) \overset{\text{notation}}{=} A\vec{e}_i = \begin{pmatrix} a_{1,i} \\ a_{2,i} \end{pmatrix} \quad \text{et} \quad g(\vec{e}_i) \overset{\text{notation}}{=} B\vec{e}_i = \begin{pmatrix} b_{1,i} \\ b_{2,i} \end{pmatrix}$$

Par conséquent, l'image du i-ième vecteur de base est :

$$(f+g)(\vec{e}_i) = f(\vec{e}_i) + g(\vec{e}_i) \overset{\text{notation}}{=} \begin{pmatrix} a_{1,i} \\ a_{2,i} \end{pmatrix} + \begin{pmatrix} b_{1,i} \\ b_{2,i} \end{pmatrix} = \begin{pmatrix} a_{1,i} + b_{1,i} \\ a_{2,i} + b_{2,i} \end{pmatrix}$$

On reconnaît ainsi les colonnes de $A + B$. □

8.2.3 Composition de transformations linéaires

Soit f et g deux transformations linéaires du plan. Notons A et B les matrices associées respectivement à f et à g.

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \quad \text{et} \quad B = \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix}$$

La matrice associée à la transformation $f \circ g$ est donnée par la *multiplication des matrices* A et B :

$$AB = \left(\begin{array}{|cc|} \hline a_{1,1} & a_{1,2} \\ \hline a_{2,1} & a_{2,2} \\ \hline \end{array}\right) \left(\begin{array}{|c|c|} \hline b_{1,1} & b_{1,2} \\ \hline b_{2,1} & b_{2,2} \\ \hline \end{array}\right) = \left(\begin{array}{|c|c|} \hline a_{1,1}b_{1,1} + a_{1,2}b_{2,1} & a_{1,1}b_{1,2} + a_{1,2}b_{2,2} \\ \hline a_{2,1}b_{1,1} + a_{2,2}b_{2,1} & a_{2,1}b_{1,2} + a_{2,2}b_{2,2} \\ \hline \end{array}\right)$$

Preuve

En effet, comme les colonnes de la matrice sont les images des vecteurs de base, il suffit de calculer les images de $\vec{e}_1$ et de $\vec{e}_2$ par l'application $f \circ g$. Par hypothèse, on a

$$(f \circ g)(\vec{e}_i) = f\big(g(\vec{e}_i)\big) \overset{\text{notation}}{=} A\big(B\vec{e}_i\big)$$

Grâce à la règle de construction des matrices, on constate que $B\vec{e}_i$ est la i-ième colonne de B. Cela permet de continuer le calcul, puisqu'on sait multiplier une matrice et un vecteur.

$$A\big(B\vec{e}_i\big) = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \begin{pmatrix} b_{1,i} \\ b_{2,i} \end{pmatrix} = \begin{pmatrix} a_{1,1}b_{1,i} + a_{1,2}b_{2,i} \\ a_{2,1}b_{1,i} + a_{2,2}b_{2,i} \end{pmatrix}$$

On reconnaît ainsi les colonnes de AB. □

8.2.4 Exemples de matrices

On utilise la règle pour la construction de la matrice A associée à la transformation f désirée.

LES COLONNES DE LA MATRICE SONT LES IMAGES DES VECTEURS DE BASE

Matrice identité

Voici la matrice associée à la transformation qui ne fait rien.

$$\mathbb{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Matrice de rotation d'un quart de tour

Voici la matrice R associée à la rotation d'angle $\frac{\pi}{2}$.

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

Matrice associée à une homothétie de facteur 2

Voici la matrice H associée à une homothétie de facteur 2.

$$H = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

Matrices de symétrie

Voici la matrice S_x associée à la symétrie selon l'axe des x.

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

Voici la matrice S_y associée à la symétrie selon l'axe des y.

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

Matrice d'étirement

On peut considérer un étirement d'un facteur 2 selon l'axe des x et d'un facteur 3 selon l'axe des y. Voici sa matrice associée.

$$E = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$$

Matrices de projection

On peut effectuer une projection orthogonale sur l'axe des x.

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

On peut aussi projeter tout le plan sur l'origine.

$$\mathbb{O} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

8.3 Création de fractales

Quelques fractales célèbres comme les napperons de Sierpinski, le flocon de von Koch, la fougère de Barnsley sont obtenus en utilisant des machines à copies réduites multiples (MCRM). Ces machines consistent à prendre une image et à la transformer en un collage de plusieurs images obtenues à l'aide de transformations affines contractantes [3] de l'image précédente. Pour simplifier une telle étape sera appelée *fractalisation*.

Mathématiquement, si A est une image (un sous-ensemble du plan), sa fractalisation sera notée $W(A)$. Si on utilise n transformations affines contractantes, notées w_1, ..., w_n, alors on a

$$W(A) = w_1(A) \cup w_2(A) \cup \cdots \cup w_n(A)$$

Si A_0 est l'image de départ, $A_1 = W(A_0)$ sera sa première fractalisation, $A_2 = W(A_1)$ sera sa deuxième fractalisation. Ainsi de suite, $A_k = W(A_{k-1})$ sera sa k-ième fractalisation.

Dans A_1 on retrouve n copies de A_0. Dans A_2 on retrouve n copies de A_1, donc n^2 copies de A_0. Ainsi, on voit que A_k contient n^k copies de A_0.

Théorème du point fixe de Banach-Hausdorff

Si, dans $\mathbb{R}^n$, on a n transformations affines contractantes[3] w_1, ..., w_n et l'opérateur de Hutchinson

$$W(A) = w_1(A) \cup w_2(A) \cup \cdots \cup w_n(A) \quad \text{avec} \quad A \subset \mathbb{R}^n$$

Alors, il existe une seule image (compacte [4], non vide) qui est solution de l'équation

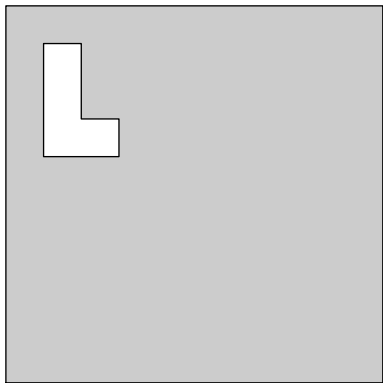
$$W(X) = X$$

De plus cette image est la limite des fractalisations de n'importe quel ensemble borné non vide dans $\mathbb{R}^n$. Pour cette raison, cette image est notée $A_{+\infty}$ et appelée l'attracteur associé à la machine à copies réduites multiples.

8.3.1 Description d'une MCRM

Pour décrire une MCRM, on utilise des modèles : il s'agit d'une image non symétrique qui montre les transformations affines contractantes utilisée à chaque fractalisation.

Voici le modèle qui sera utilisé dans ce cours.

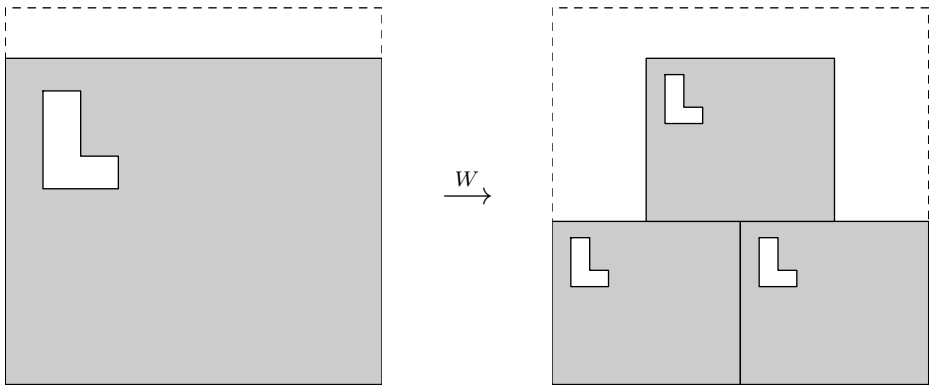


3. Une fonction est dite contractante si la distance entre deux points quelconques diminue lorsque l'on applique la fonction.

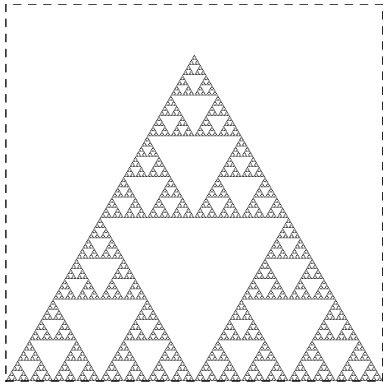
4. Dans $\mathbb{R}^n$, les parties compactes sont les parties fermées et bornées. Une partie est fermée si toute suite convergente contenue dans la partie converge vers un point de la partie (par exemple, l'intervalle $]0,1]$ n'est pas fermé car la suite $(\frac{1}{n})_{n\geqslant 1}$ converge vers 0 et que $0 \notin]0,1]$).

Le napperon de Sierpinski

Voici la MCRM qui permet d'obtenir le napperon de Sierpinski. Afin d'obtenir une image finale inscrite dans un triangle équilatéral, on donne des dimensions légèrement différentes au modèle (bien que seul les applications affines contractantes comptent, elles sont plus facilement discernables avec ce modèle).

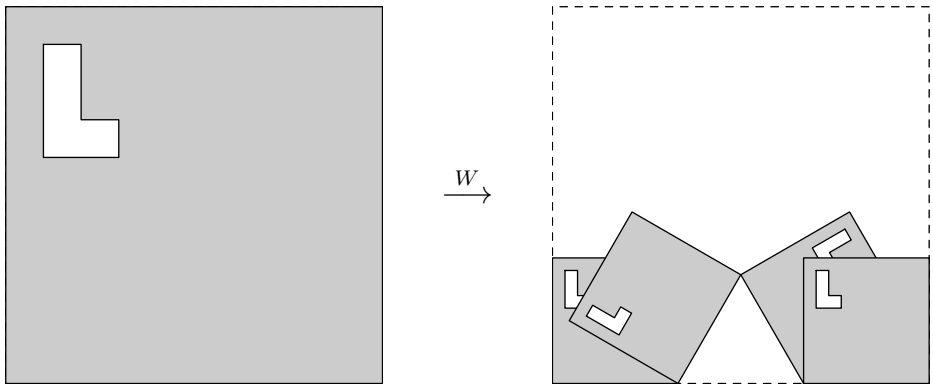


Voici l'attracteur d'une telle MCRM.

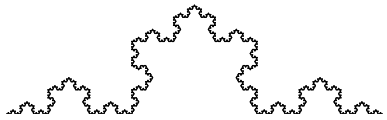


La courbe de von Koch

Voici la MCRM qui permet d'obtenir la courbe de von Koch associée à un angle de 60°.



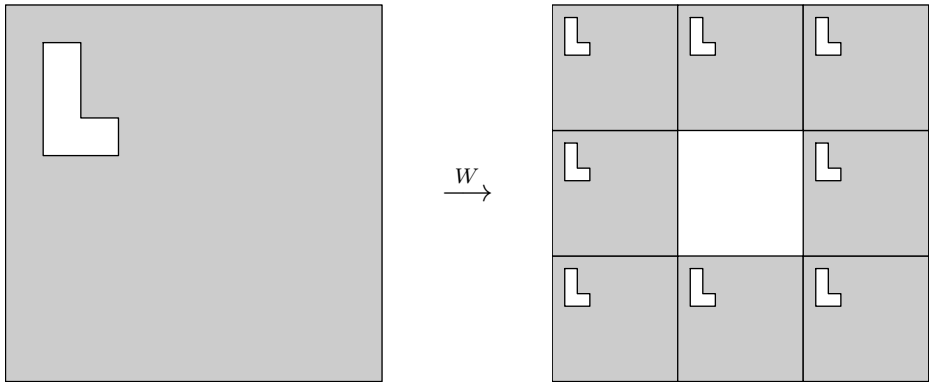
Voici l'attracteur d'une telle MCRM.



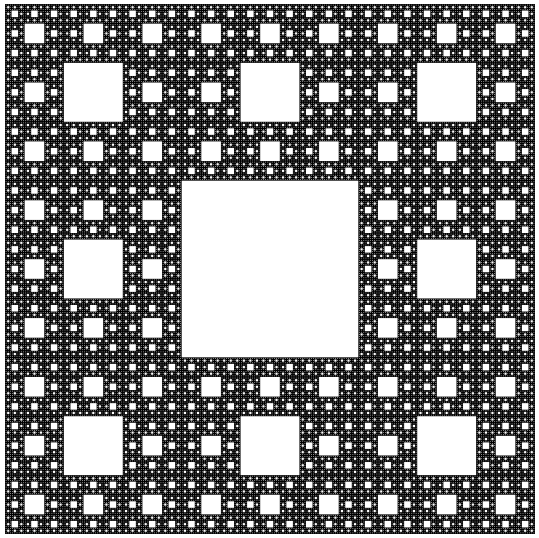
Cet attracteur est une courbe continue (pas une fonction !) qui n'est dérivable en aucun point !!!

Le tapis de Sierpinski

Voici la MCRM qui permet d'obtenir le tapis de Sierpinski.

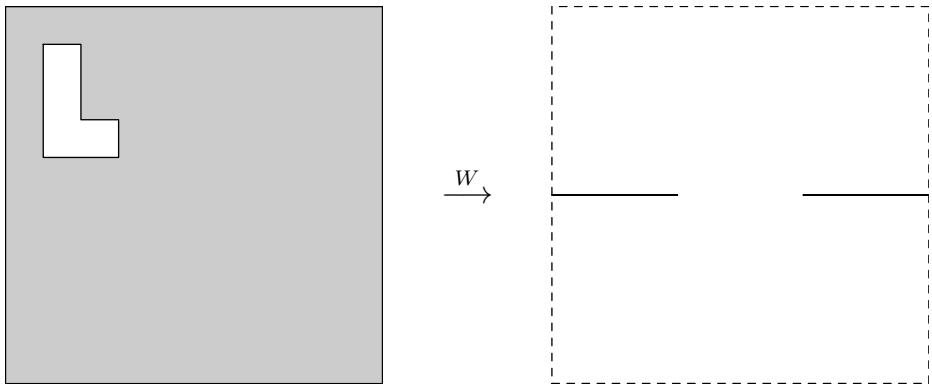


Voici l'attracteur d'une telle MCRM.



L'ensemble de Cantor

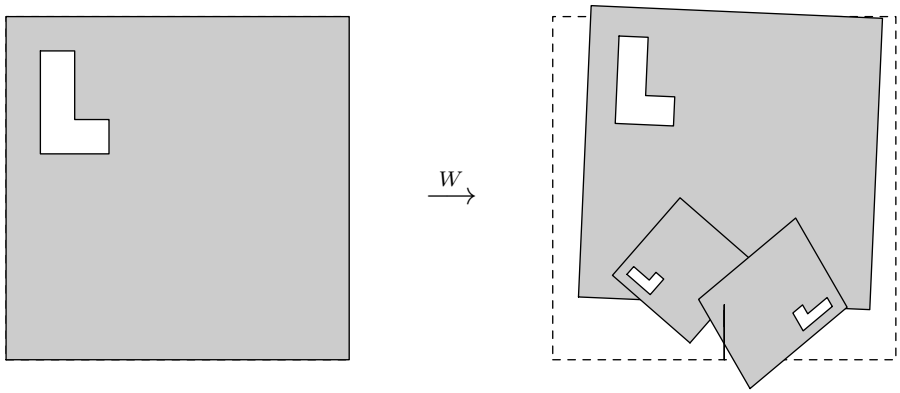
L'ensemble de Cantor s'obtient en enlevant à la ligne $[0, 1]$ son tiers médian, puis à chaque ligne restante on enlève le tiers médian, et ainsi de suite... Voici la MCRM qui permet d'obtenir cette fractale.



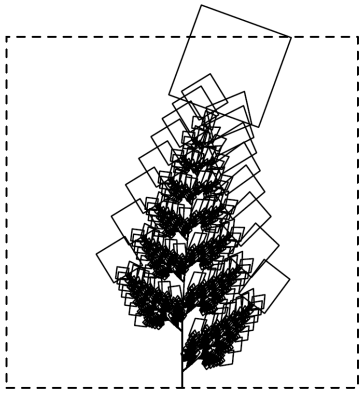
On peut voir l'ensemble de Cantor dans le tapis de Sierpinski (prendre l'intersection du tapis de Sierpinski avec la droite horizonzale passant par le milieu du carré (d'équation $y = \frac{1}{2}$). On peut aussi la voir comme la droite verticale passant par le milieu du carré, ou encore comme l'une ou l'autre des diagonales.

La fougère de Barnsley

Voici la MCRM qui permet d'obtenir la fougère de Barnsley.



L'attracteur d'une telle MCRM est informatiquement pénible à obtenir à l'aide de la MCRM. Voici la huitième fractalisation qui, sur un Pentium 4 cadencé à 3.2 GHz, a nécessité plus de 72 secondes. Cette fractalisation à été obtenue en prenant un carré vide comme image de départ.



Cherchons combien de temps cela prendrait-il pour avoir une image de taille de 1000 pixels par 1000 pixels (les photos numériques de haute qualité ont plus de pixels que cela) avec une fougère en haute résolution. Le plus grand côté du grand modèle à la première fractalisation est le 85% du côté correspondant sur le modèle initial. Le nombre N de fractalisations nécessaire satisfait donc l'équation suivante (puisqu'il faudrait que la taille du grand modèle soit de 1 pixel carré afin d'avoir une image haute définition).

$$1000 \cdot 0.85^N \cong 1 \iff 0.85^N \cong 0.001 \iff N \cong \log_{0.85}(0.001) \cong 42.50$$

Il faudrait ainsi un minimum de 43 fractalisations. Si on note M le nombre de rectangles qu'il faut dessiner (et dont il faut calculer les coordonnées), on a

$$M = 1 + 4 + 4^2 + 4^3 + \cdots + 4^N = \frac{4^{N+1} - 1}{3}$$

Pour $N = 43$, on a $M \cong 1.03 \cdot 10^{26}$ rectangles. En se basant sur le fait que le pentium ci-dessus à pris 72 secondes pour dessiner 87'381 rectangles ($M = 8$) et en supposant que le temps nécessaire est proportionnel, on aurait besoin plus de $8.50 \cdot 10^{22}$ secondes, ce qui fait plus de $9.8380 \cdot 10^{17}$ jours. En se basant sur le fait qu'une année astronomique prends environ $365,2422$ jours, le calcul prendrait plus de $2.69 \cdot 10^{15}$ années. Ce qui est un temps plus grand que l'âge de l'Univers !

8.4 Le jeu du Chaos

8.4.1 Une surprise

On prend un triangle isocèle et l'on numérote les sommets. Considérons le jeu de hasard suivant. On prend un point du plan et on choisit aléatoirement (en lançant un dé par exemple) un nombre parmi 1, 2 ou 3. Le point suivant sera au milieu du segment dont les sommets sont le point précédent et le sommet du triangle associé au choix aléatoire.

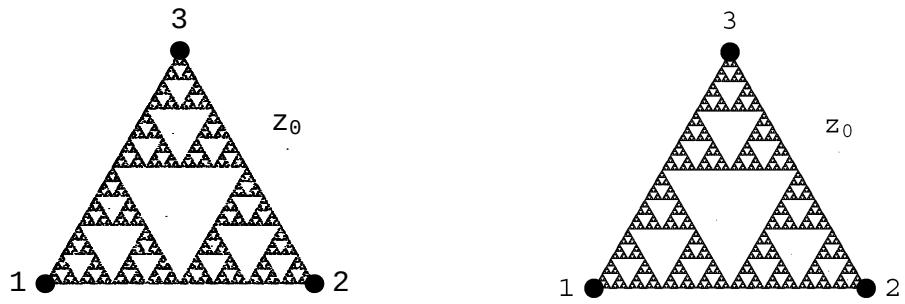
Voici deux exemples où le point de départ z_0 est le même et où on a joué 6 fois.

On peut se convaincre aisément qu'une fois qu'un point arrive dans le triangle il n'en sort plus. Mis à part cette remarque, on a l'impression que les points peuvent se déplacer n'importe où et qu'il n'y a pas d'intérêt à étudier ce jeu de manière plus attentive.

Voici deux exemples où le point de départ z_0 est le même et où on a joué 50 fois. Pour un meilleur aspect seul les 10 premiers points ont été reliés.

Maintenant on constate que le milieu du triangle contient relativement moins de points. Ainsi, il se passe peut-être quelque chose d'intéressant.

Voici ce qui se passe si on joue 10′000 fois (à gauche) et 100′000 fois (à droite).



Ohh... On voit apparaître une fractale : le napperon de Sierpinski !

8.4.2 Le jeu du chaos et les attracteurs des MCRM

Rappelons qu'une MCRM est composée de n transformations affines contractantes, notées w_1, ..., w_n et que l'attracteur est obtenu en itérant l'opérateur de Hutchinson suivant sur une image bornée quelconque.

$$W(A) = w_1(A) \cup w_2(A) \cup \cdots \cup w_n(A) \quad \text{avec} \quad A \subset \mathbb{R}^n$$

Le jeu du chaos associé consiste à choisir un point du plan et à lui appliquer itérativement une seule transformation affine contractante choisie au hasard parmi les n possibles. Si le point choisi est un point de l'attracteur, alors tous les points suivants seront aussi dans l'attracteur. Mieux : pour chaque point de l'attracteur, il y a une probabilité non nulle d'avoir un point de cette suite d'itérations qui sera autant proche que l'on veut du point de l'attracteur. En langage mathématique cela se traduit par le théorème suivant.

Théorème

Si, dans $\mathbb{R}^n$, on a n transformations affines contractantes w_1, ..., w_n et des nombres réels positifs p_1, ..., p_n tels que $\sum_{i=1}^{n} p_i = 1$ (ce sont les probabilités de choisir les transformations correspondantes).

Notons $(s_i)_{i\geqslant 1}$ la suite de nombres aléatoires choisis entre 1 et n avec les probabilités associées ci-dessus.

Soit z_0 un point de l'attracteur de la MCRM associée aux transformations, notée $A_{+\infty}$ (on peut prendre n'importe quel point fixe d'une des transformations (car ce point est forcément dans l'attracteur)).

On considère la suite aléatoire $z = (z_i)_{i\in\mathbb{N}}$ telle que $z_k = w_{s_k}(z_{k-1})$ pour tout $k \geqslant 1$.

Alors

1. Tous les points de la suite z sont dans l'attracteur $A_{+\infty}$.
2. Cette suite remplit presque sûrement [5] de manière dense [6] l'attracteur $A_{+\infty}$.

Remarque

Si le point z_0 n'est pas dans l'attracteur, on a tout de même une excellente approximation, en effet plus on avance dans la suite plus on se trouve dans une fractalisation proche de l'attracteur.

En effet, on applique le théorème du point fixe de Banach-Hausdorff avec la partie $A_0 = \{z_0\}$. Soulignons le fait que le i-ième élément de la suite z se trouve dans la i-ième fractalisation A_i.

Idée de la preuve

Le point 1 provient de l'invariance de l'attracteur par l'opérateur de Hutchinson, c'est-à-dire

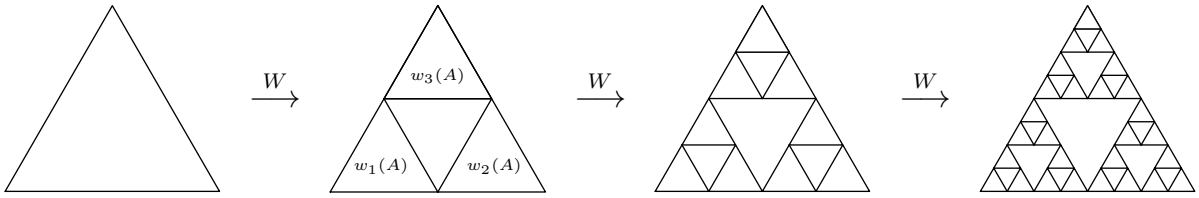
$$W(A_{+\infty}) = A_{+\infty}$$

5. Cela signifie que la probabilité pour que cela ne soit pas le cas est nulle.
6. Un ensemble est dit dense dans un autre si tout point de l'autre ensemble admet un point arbitrairement proche dans le premier ensemble.

Rappelons que cet opérateur est défini comme suit.

$$W(A) = w_1(A) \cup w_2(A) \cup \cdots \cup w_n(A) \quad \text{avec} \quad A \subset \mathbb{R}^n$$

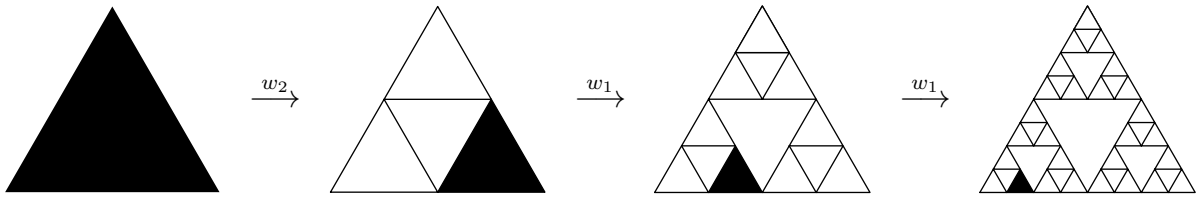
Le point 2 est plus délicat à montrer. On va regarder ce qu'il se passe sur le napperon de Sierpinski et on va remplacer la difficulté de la démonstration due à la densité en pensant à ce qu'il se passe lors du dessin (résolution de l'image).
Voici les premières fractalisations du napperon de Sierpinski (en prenant des triangles vides comme image de base).



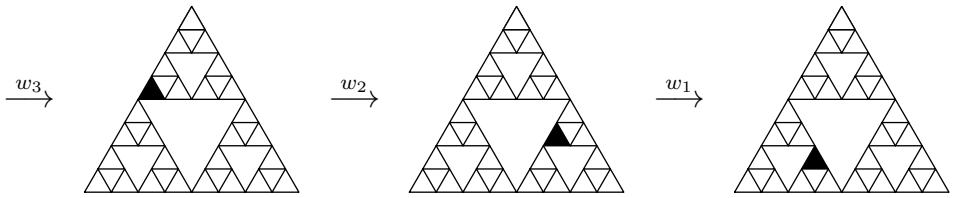
Imaginons que la résolution du dernier dessin soit suffisante (si ce n'est pas le cas, alors il suffit de continuer un peu la fractalisation).

On doit tirer au hasard une suite de nombres entre 1 et 3 (puisqu'il y a exactement trois transformations affines contractantes pour le napperon de Sierpinski).

Imaginons que l'on ait $s_1 = 2$, $s_2 = 1$, $s_3 = 1$, $s_4 = 3$, $s_5 = 2$ et $s_6 = 1$ pour les six premiers termes de la suite aléatoire. Prenons le coin en bas à gauche pour z_0. Ainsi les quatre premiers termes de la suite z sont z_0, $w_2(z_0)$, $w_1(w_2(z_0))$, $w_1(w_1(w_2(z_0)))$ et $w_3(w_1(w_1(w_2(z_0))))$. Ci-dessous, on noircit les triangles de la fractalisation dans lequel se trouve les éléments de la suite z (dans le cas où le point se trouve sur un coin, on noircit le triangle dont le coin est en bas à gauche).



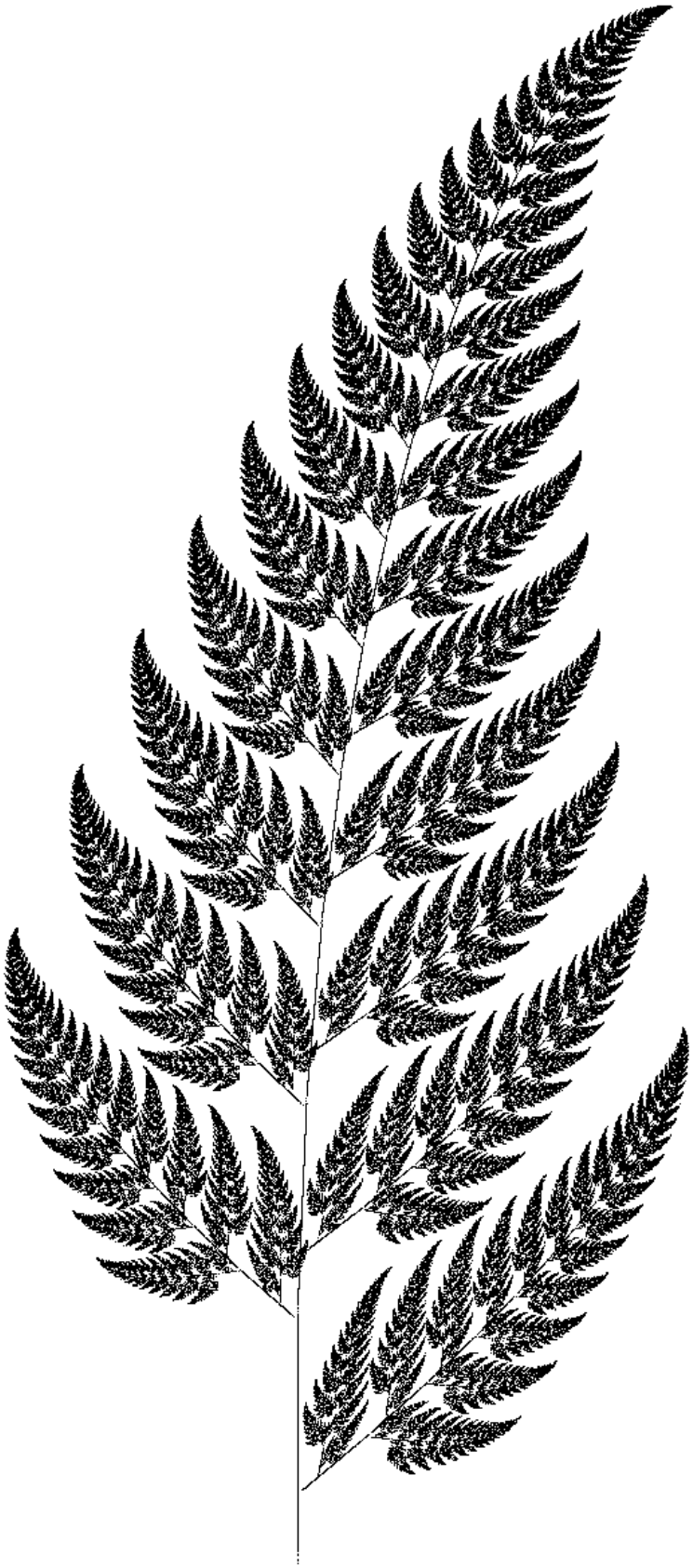
Comme on a supposé avoir atteint la résolution minimale, on va imaginer la fractalisation suivante, mais seulement noircir le triangle dans la résolution minimale.



Maintenant qu'on a vu ce qu'il se passe sur la résolution minimale, il faut démontrer que tous les triangles de cette fractalisation seront remplit lorsque l'on avance le long de la suite aléatoire. Or, dans notre exemple il y a 27 triangles (puisqu'on s'est arrêté à la troisième fractalisation et qu'il y a 3 transformations). Le premier triangle noirci ci-dessus dans la résolution minimale correspond aux valeurs $(2, 1, 1)$ de la suite, le triangle noirci dans l'image suivante correspond aux valeurs $(1, 1, 3)$, le suivant correspond à $(1, 3, 2)$, le suivant correspond à $(3, 2, 1)$. Ainsi on voit que si dans la suite aléatoire, toutes les combinaisons apparaissent, alors tous les triangles seront noircis (il y a bien 27 triplets de nombres choisis entre 1 et 3). Or en choisissant à chaque étape un nombre au hasard parmi 1, 2 ou 3, on a une probabilité non nulle de trouver tous les 27 triplets cités ci-dessus.

La fougère de Barnsley

Grâce au jeu du chaos, on peut dessiner la fougère de Barnsley dans un temps beaucoup plus réaliste.



8.5 Dimension d'ensembles auto-semblables

Définitions

1. Une *similitude* est une transformation affine qui n'est composée que de rotations, d'homothéties, de symétries et de translation.
2. Une partie est *semblable* à une autre partie s'il existe une similitude qui transforme la partie en l'autre partie.
3. Un ensemble est *auto-semblable* s'il peut être partitionné en morceaux arbitrairement petits qui sont tous semblables à l'ensemble lui-même.

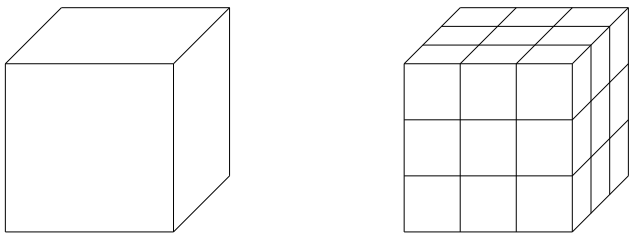
Remarques

1. Il existe des ensembles auto-semblables qui ne sont pas des fractales, tels qu'un segment, un carré et un cube.
2. Le napperon et le tapis de Sierpinski, l'ensemble de Cantor et la courbe de von Koch sont auto-semblables. La fougère de Barnsley n'est pas auto-semblable (il y a des étirements dans les transformations affines).

Dimension d'un segment, d'un carré et d'un cube

Un segment, un carré et un cube sont des ensembles auto-semblables puisqu'on peut les partitionner en utilisant des réductions.

Voici une partition sur le cube associée à un facteur de réduction de $\frac{1}{3}$.



Le tableau suivant indique combien de morceaux correspondent à une partition associée à un facteur de réduction donné.

Objet	nombre de morceaux	facteur de réduction	dimension de l'objet
segment	3	$1/3$	
segment	9	$1/9$	1
segment	n	$1/n$	
carré	9	$1/3$	
carré	81	$1/9$	2
carré	n^2	$1/n$	
cube	27	$1/3$	
cube	729	$1/9$	3
cube	n^3	$1/n$	

Si on note n pour le nombre de morceaux, r pour le facteur de réduction et D pour la dimension, on remarque la relation suivante.

$$n = \frac{1}{r^D} \iff \ln(n) = D \ln\left(\frac{1}{r}\right) \iff D = \frac{\ln(n)}{\ln(1/r)}$$

8.5.1 Dimension des fractales auto-semblables

Théorème

Pour tout ensemble auto-semblable, le nombre $D = \frac{\ln(n)}{\ln(1/r)}$ est le même quelque soit la partition en n morceaux pour un (unique) facteur de réduction r.

Ce nombre D est appelé la *dimension auto-semblable*.

Dimension de la courbe de von Koch

nombre de morceaux	facteur de réduction
4	1/3
16	1/9
4^k	$1/3^k$

Regardons ce qu'il se passe si on reprend la formule précédente.

$$D = \frac{\ln(n)}{\ln(1/r)} = \frac{\ln(4^k)}{\ln(3^k)} = \frac{k\ln(4)}{k\ln(3)} = \frac{\ln(4)}{\ln(3)} \cong 1.2619$$

Dimension du napperon de Sierpinski

nombre de morceaux	facteur de réduction
3	1/2
9	1/4
3^k	$1/2^k$

Regardons ce qu'il se passe si on reprend la formule précédente.

$$D = \frac{\ln(n)}{\ln(1/r)} = \frac{\ln(3^k)}{\ln(2^k)} = \frac{k\ln(3)}{k\ln(2)} = \frac{\ln(3)}{\ln(2)} \cong 1.5850$$

Dimension de l'ensemble de Cantor

nombre de morceaux	facteur de réduction
2	1/3
4	1/9
2^k	$1/3^k$

Regardons ce qu'il se passe si on reprend la formule précédente.

$$D = \frac{\ln(n)}{\ln(1/r)} = \frac{\ln(2^k)}{\ln(3^k)} = \frac{k\ln(2)}{k\ln(3)} = \frac{\ln(2)}{\ln(3)} \cong 0.6309$$

Chapitre 9

Codes correcteurs d'erreurs

9.1 Introduction : Le sport-toto

Le sport-toto est un jeu (de hasard !) où il faut deviner le score des matchs de football.

Imaginons un sport-toto à 4 matchs. Pour chaque match, on note 1 pour une victoire de l'équipe qui joue à domicile, 2 pour une victoire de l'équipe invitée et x pour un match nul.

On remplit une grille pour les 4 matchs. Par exemple on peut jouer la colonne suivante.

1er match	1
2e match	2
3e match	x
4e match	1

On peut se poser les questions suivantes.

1. Combien de grilles doit-on remplir pour être sûr de gagner (une grille a les 4 bons résultats) ?
2. Combien de grilles doit-on remplir pour être sûr d'avoir 3 matchs sur 4 avec les bons résultats ?

Les réponses sont les suivantes.

1. Cette réponse est facile, il y a $\boxed{}$ possibilités et une seule combinaison est gagnante. Il faut donc jouer $\boxed{}$ grilles.
2. Ici, c'est plus subtil, mais possible. Pour cela, on peut jouer les 9 colonnes suivantes.

x	x	x	1	1	1	2	2	2
x	1	2	x	1	2	x	1	2
x	1	2	1	2	x	2	x	1
x	1	2	2	x	1	1	2	x

   En effet, pour chaque grille parmi toutes celles possibles (que l'on supposera être le résultat des 4 matchs), si on regarde combien de points on réalise avec chacune des 9 colonnes ci-dessus (un point pour un match juste), on verra que l'on a toujours une colonne ci-dessus qui livre 3 ou 4 points. C'est bien sûr une démonstration longue et ennuyeuse. Il y a une démonstration plus rapide.

Preuve

Créons un peu de vocabulaire. Disons que la *sphère d'influence centrée en une grille* est l'ensemble des colonnes qui ont au plus une différence avec cette grille. Il est facile de constater que chaque sphère d'influence a exactement 9 colonnes (la colonne au centre et les 8 colonnes qui ont exactement 1 différence (4 endroits et 2 possibilités)).

Regardons uniquement les sphères d'influence des 9 colonnes qui nous intéressent. Ces neufs colonnes ont la propriété essentielle suivante.

Il y a toujours 3 différences entre deux colonnes choisies parmi les 9.

Cela signifie que les 9 sphères d'influence sont disjointes.

On a donc 9 sphères d'influence disjointes contenant chacune 9 éléments. Ainsi ces 9 sphères contiennent au total $9 \cdot 9 = 81$ grilles. Il s'agit de toutes les grilles possibles.

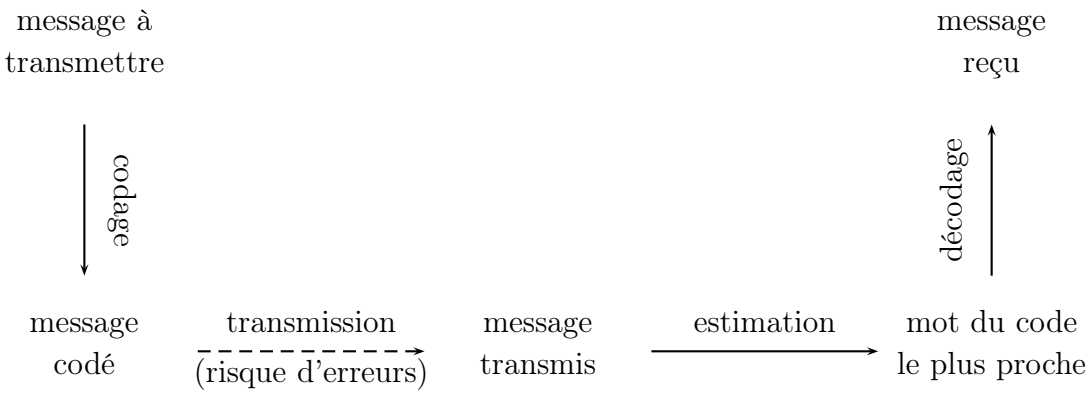
Ainsi, chaque grille (parmi les 81 possibles) se trouve à l'intérieur d'une seule sphère d'influence centrée en une grille parmi les 9 colonnes ci-dessus. La colonne correspondante est celle qui donne 3 ou 4 points. □

Bien évidemment les créateurs sont maintenant au courant de cette astuce et proposent des sport-toto à plus de 13 matchs (il y a aussi une technique similaire permettant d'assurer 12 points sur 13 matchs, mais elle coûte plus cher qu'elle ne rapporte).

9.2 Codes correcteurs d'erreurs

Contrairement aux codes en cryptographie (qui consistent à camoufler un message), les codes correcteurs d'erreurs ont été inventés pour pouvoir détecter et éventuellement corriger des erreurs qui s'y seraient glissées (de manière accidentelle). Voici une méthode bien connue des spécialistes du radar.

Voici le schéma à avoir en tête lorsqu'on pense aux codes correcteurs.



9.2.1 La méthode des spécialistes du radar

Si lors d'une transmission horizontale, une vingtaine de caractères consécutifs sont perdus, il peut être très difficile de les retrouver !

```
Souvent, pour s'amuser, les hommes d'équipage
Prennent des albatros, vastes oiseaux des mers,
Qui suivent, indolents compagnons de voyage,
Le navire glissant sur les gouffres amers.
A peine les ont-ils déposés sur les planches,
Que ces rois de l'azur, maladroits et honteux,
********** leurs grandes ailes blanches
Comme des avirons traîner à coté d'eux.
Ce voyageur ailé, comme il est gauche et veule!
Lui, naguère si beau, qu'il est comique et laid!
L'un agace son bec avec un brûle-gueule,
L'autre mime, en boitant, l'infirme qui volait!
Le Poête est semblable au prince des nuées
Qui hante la tempête et se rit de l'archer;
Exilé sur le sol au milieu des huées,
Ses ailes de géant l'empêchent de marcher.
Charles Baudelaire (Les fleurs du mal)
```

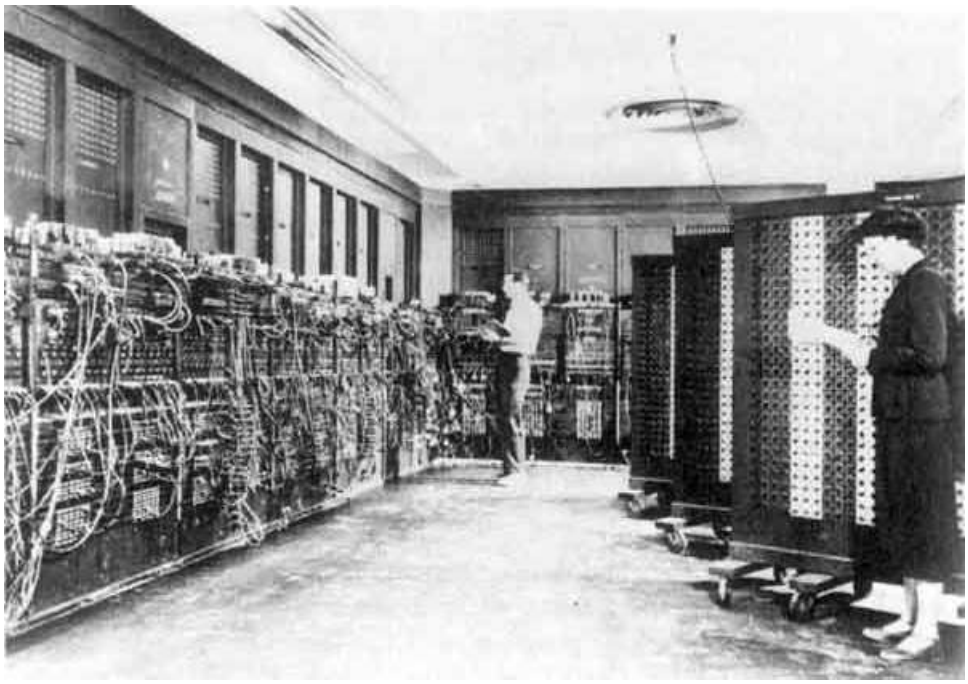
Par contre, si on transmet le texte verticalement, c'est un jeu d'enfant de retrouver vingt caractères consécutifs perdus.

```
Souvent, pour s'amu**r, les hommes d'équipage
Prennent des albatr**, vastes oiseaux des mers,
Qui suivent, indole**s compagnons de voyage,
Le navire glissant *ur les gouffres amers.
A peine les ont-ils*déposés sur les planches,
Que ces rois de l'a*ur, maladroits et honteux,
Laissent piteusemen* leurs grandes ailes blanches
Comme des avirons t*aîner à coté d'eux.
Ce voyageur ailé, c*mme il est gauche et veule!
Lui, naguère si bea*, qu'il est comique et laid!
L'un agace son bec *vec un brûle-gueule,
L'autre mime, en bo*tant, l'infirme qui volait!
Le Poête est sembla*le au prince des nuées
Qui hante la tempêt* et se rit de l'archer;
Exilé sur le sol au*milieu des huées,
Ses ailes de géant *'empêchent de marcher.
Charles Baudelaire *Les fleurs du mal)
```

Malheureusement, cette méthode ne fonctionne pas pour des messages composés de chiffres ou de lettres disposées de manière apparemment aléatoire (penser aux documents numériques : images, sons, musiques, vidéos).

9.3 Le code de Hamming

En 1947, Richard W. Hamming avait accès à un ordinateur de l'armée seulement pendant les week-ends. Les ordinateurs de l'époque étaient très grands (voir photo ci-dessous) et extrêmement lents par rapport à ceux d'aujourd'hui.



Cette photo provient de l'armée américaine et est dans le domaine publique.

L'ordinateur sur lequel Hamming travaillait avait un code détecteur d'erreur, appelé 2-sur-5. On disposait les nombres de 0 à 9 sur des rampes de 5 lampes dont 2 étaient allumées et 3 étaient éteintes.

1	1	1	0	0	0
2	1	0	1	0	0
3	0	1	1	0	0
4	1	0	0	1	0
5	0	1	0	1	0
6	0	0	1	1	0
7	1	0	0	0	1
8	0	1	0	0	1
9	0	0	1	0	1
0	0	0	0	1	1

On voit que toutes les combinaisons possibles de deux lampes allumées sont représentées. Ainsi, si on voit qu'il n'y a pas exactement deux lampes allumées, on sait qu'une erreur s'est produite. Les opérateurs pouvaient retrouver l'erreur (en examinant ce qu'il s'était passé avant), mais ils n'étaient pas présent le week-end et l'ordinateur devait être redémarré (en perdant beaucoup de temps).

Après qu'un calcul a été stoppé de cette manière deux week-ends consécutifs, Hamming était frustré et ennuyé et il s'est demandé pourquoi si l'ordinateur pouvait détecter, il ne pouvait pas trouver sa position et la corriger.

Il inventa ainsi le premier code correcteur de l'Histoire en 1947.

Il s'est placé dans l'anneau $\mathbb{Z}_2 = \{0, 1\}$ avec les règles d'addition suivantes.

$$0 + 1 = 1 = 1 + 0 \qquad 0 + 0 = 0 \qquad 1 + 1 = 0$$

Si on écrit les nombres de 0 à 9 en base 2, on a besoin de 4 lampes.

	0	1	2	3	4	5	6	7	8	9
1	0	1	0	1	0	1	0	1	0	1
2	0	0	1	1	0	0	1	1	0	0
4	0	0	0	0	1	1	1	1	0	0
8	0	0	0	0	0	0	0	0	1	1

Et il eu l'idée d'écrire le tableau suivant.

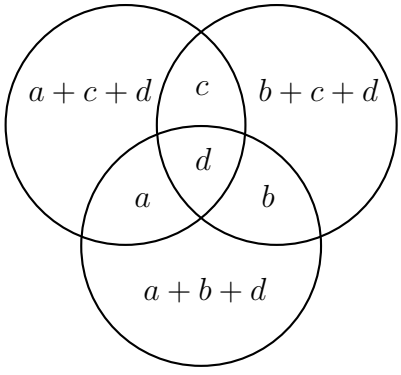
a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d$

Cela donnait une application de codage où chaque chiffre était représenté par un mot (a, b, c, d) avec a, b, c et $d \in \mathbb{Z}_2$. On y associait le mot codé

$$(a,\ b,\ c,\ d,\ \underbrace{a+b,\ c+d,\ a+c,\ b+d,\ a+b+c+d}_{\text{caractères de contrôle}})$$

Si un des neufs éléments de ce mot est changé (un 1 en un 0 ou l'inverse), alors on peut dire qu'il y a une erreur et on peut même situer où elle se trouve. Ainsi faisant, on s'aperçoit (comme Hamming) qu'on peut se passer du caractère de contrôle $a+b+c+d$. Ainsi, on a besoin de 8 lampes pour corriger une erreur (4 pour le chiffre auxquelles on en ajoute 4 pour les caractères de contrôle).

Mais Hamming a réussi à faire encore mieux, grâce à l'idée suivante (Leonhard Euler a eu l'idée d'utiliser les diagrammes de Venn).



On a donc l'application de codage suivante (ordre alphabétique).

$$(a, b, c, d) \mapsto (a,\ b,\ c,\ d,\ \underbrace{a+b+d,\ a+c+d,\ b+c+d}_{\text{caractères de contrôle}})$$

Ce code plus astucieux permet de corriger une erreur et de n'utiliser que 7 lampes. Il a été démontré qu'on ne peux pas corriger une erreur avec moins de 7 lampes.

Proposition

S'il existe un code binaire 1-correcteur d'erreur de longueur n systématique sur les r premières positions (cela signifie que sur les r premières positions on retrouve le message à coder : le code de Hamming ci-dessus est de longueur 7 et systématique sur les 4 premières positions). Alors

$$2^n \geqslant (n+1)2^r$$

Preuve

Il y a 2^r mots du code (un par message à coder) et 2^n mots de longueur n (potentiellement recevable après la transmission et ses multiples erreurs possibles).

Si le code est 1-correcteur, cela signifie que les sphères d'influence des mots du code sont disjointes (rappelons que la sphère d'influence d'un mot du code contient tous les mots qui ont au plus une différence par rapport au mode du code).

Comme on parle de code binaire de longueur n, chaque sphère d'influence d'un mot du code contient $(n+1)$ mots (n modifications possibles d'un zéro ou d'un un et le mot du code lui-même).

On a ainsi

$$\overbrace{2^n}^{\text{nombre de mots de longueur } n} \geqslant \underbrace{\underbrace{2^r}_{\text{nombre de sphères d'influences centrées en les mots du code}} \underbrace{(n+1)}_{\text{nombre de mots dans chaque sphère}}}_{\text{nombre de mots dans toutes les sphères}}$$

On remarque, en bonus, qu'on a l'égalité lorsque tout mot de longueur n se trouve dans une unique sphère d'influence centrée en un mot du code. □

Cas d'égalité

L'égalité de l'équation de la proposition se produit pour

1. $n = 3$ et $r = 1$.

   Il s'agit du code 1-correcteur élémentaire donné par l'application de codage

   $$(a) \mapsto (a, a, a)$$

   En effet, si on triple l'information, on a un code qui corrige une erreur.

2. $n = 7$ et $r = 4$.

   C'est le code de Hamming vu précédemment.

3. $n = 15$ et $r = 11$.

   Il existe un tel code qui a été utilisé à l'époque dans les transmissions US.

4. $n = 31$ et $r = 26$.

   Il est théoriquement possible qu'un tel code existe, mais pour en avoir la certitude, il faudrait l'exhiber. Or même s'il existait, un tel code ne serait pas si utile car il ne permettrait que de corriger une erreur sur les 31 positions possibles.

5. Pour tous les nombres n entre 2 et 31 qui n'apparaissent pas ci-dessus, la valeur de r n'est pas entière. On est donc sûr qu'il n'y a pas de codes binaires de ces longueurs n.

9.4 Les codes ISBN

En 1972, on a commencé à assigner un numéro ISBN (International Standard Book Number) aux livres (certaines informations y sont cachées, comme la zone linguistique, etc). En 2007, le code ISBN-13 est apparu pour principalement deux raisons : augmenter la capacité de numérotation des ouvrages et s'aligner avec les codes barres.

9.4.1 Le code ISBN-10

On note $\mathcal{N} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ l'ensemble des chiffres. Le code ISBN-10 est un code de longueur 10 sur l'alphabet $\mathcal{A} = \mathcal{N} \cup \{\text{X}\}$.

L'application de codage est la suivante.

$$(x_1, \ldots, x_9) \mapsto (x_1, \ldots, x_9, \underbrace{x_{10}}_{\text{caractère de contrôle}})$$

Le caractère de contrôle x_{10} est le reste de division par 11 du nombre $\sum_{i=1}^{9} i \cdot x_i$.
Autrement dit :

$$x_{10} \equiv \sum_{i=1}^{9} i \cdot x_i \pmod{11}$$

On écrit X à la place de x_{10} si x_{10} vaut 10. Ainsi, seul le 10-ième caractère peut être un X.

Le code ISBN permet de coder 10^9 livres.

9.4.2 Le code ISBN-13

On note $\mathcal{N} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ l'ensemble des chiffres. Le code ISBN-13 est un code de longueur 13 sur l'alphabet $\mathcal{N}$.

L'application de codage est la suivante.

$$(x_1, \ldots, x_{12}) \mapsto (x_1, \ldots, x_{12}, \underbrace{x_{13}}_{\text{caractère de contrôle}})$$

Le caractère de contrôle x_{13} est le reste de division par 10 du nombre $-\sum_{\substack{i \text{ impair} \\ i \neq 13}} x_i \; - \; 3\sum_{i \text{ pair}} x_i$.
Autrement dit :

$$\sum_{i \text{ impair}} x_i \; + \; 3 \sum_{i \text{ pair}} x_i \equiv 0 \pmod{10}$$

Le code ISBN permet de coder 10^{12} livres.

Compatibilité en ISBN-10 et ISBN-13

Pour passer d'un code ISBN-10 à un code ISBN-13, on enlève le caractère de contrôle, on ajoute `978` (pour la plupart des ouvrages) et avec le code à 12 chiffres obtenus, on calcule le caractère de contrôle en suivant la méthode du code ISBN-13. Par exemple, on a

ISBN-10	étape 1	étape 2	ISBN-13
`047144779X`	`047144779`	`978047144779`	`9780471447795`
`2980859737`	`298085973`	`978298085973`	`9782980859731`

On ne peut pas faire la démarche à l'envers, car certains nouveaux ouvrages n'ont pas `978` au début de leur code ISBN-13.

Chapitre 10

Les colorations de Pólya

10.1 Groupes de permutations

Permutations : de l'intuitif au formalisme

On permute cinq objets appelés a, b, c, d et e.

avant permutation	$\boxed{a}$	$\boxed{b}$	$\boxed{c}$	$\boxed{d}$	$\boxed{e}$
après permutation	$\boxed{b}$	$\boxed{a}$	$\boxed{e}$	$\boxed{d}$	$\boxed{c}$

On voit que l'objet $\boxed{a}$ quitte la première position pour aller en deuxième position. Les autres objets vont aussi bouger (même s'il se trouve que $\boxed{d}$ reste à la même place).

Symboliquement, on peut représenter le déplacement ainsi.

position d'un objet avant le déplacement	1	2	3	4	5	
	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	σ
position du même objet après le déplacement	2	1	5	4	3	

On voit qu'une permutation de 5 objets est une *fonction bijective* d'un ensemble de 5 objets dans lui-même. Les mathématiciens utilisent une notation encore meilleure en représentant la permutation σ de la manière suivante.

$$\sigma = (12)(35)(4)$$

L'ensemble de toutes les permutations de 5 objets est appelé Sym(5) ou S_5. Lorsqu'il y a n objets, on note Sym(n) ou S_n (on dit « Symétrique n »).

Composition de permutations

Mathématiquement parlant, composer deux permutations revient à faire une permutation, puis une deuxième. Pour voir ce qu'il se passe, superposons deux permutations σ_1 et σ_2.

position d'un objet avant la permutation	1	2	3	4	5	
	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	σ_1
position du même objet après la première permutation	2	1	5	4	3	
	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	σ_2
position du même objet après la deuxième permutation	3	5	1	2	4	

On voit que la permutation composée est

position d'un objet avant la permutation	1	2	3	4	5	
	↓	↓	↓	↓	↓	$\sigma_2 \circ \sigma_1$
position du même objet après les deux permutations	3	5	1	2	4	

C'est en fait une composition de fonctions[1] (d'où la notation $\sigma_2 \circ \sigma_1$).
Quand on travaille avec des permutations, on omet le symbole $\circ$ et on a

$$\sigma_2 \circ \sigma_1 = \sigma_2\,\sigma_1 = (15)(234) \circ (12)(35)(4) = (15)(234)(12)(35)(4) = (13)(254)$$

Lorsque l'on compose des permutations, il faut lire de droite à gauche (à cause de la composition de fonctions).

Vocabulaire

1. Dans $\mathrm{Sym}(n)$, la permutation $(1)(2)(3)\cdots(n)$ est appelée id (comme *identité*).
2. Dans une permutation σ écrite en notation simplifiée, une parenthèse contenant n objets est appelée un *n-cycle*.
3. Si une permutation σ est écrite en notation simplifiée et qu'aucun nombre n'apparaît plusieurs fois, on dit que la permutation σ est écrite *en produit de cycles disjoints*.
4. Le *type* d'une permutation $\sigma \in \mathrm{Sym}(n)$ est défini par

   $$(t_1, t_2, \ldots, t_n)$$

   où t_i est le nombre de i-cycles dans la permutation lorsqu'elle est écrite en produit de cycles disjoints.

Remarque On n'est pas obligé d'écrire les 1-cycles.

Formule intéressante

Si $(t_1, t_2, \ldots, t_n)$ est le type d'une permutation, alors on a la formule évidente suivante.

$$t_1 + 2t_2 + 3t_3 + \cdots + nt_n = n$$

1. Rappelons que $(g \circ f)(x) = g(f(x))$. On applique la fonction f à l'élément x, puis la fonction g à l'élément $f(x)$ résultant de la première opération.

10.2 Groupes

Les permutations forment ce qu'on appelle aujourd'hui un groupe.

Définition

Un *groupe* est un ensemble G muni d'une opération, appelée *loi de composition* et notée ici $\star$, qui satisfait les propriétés suivantes.

1. Pour chaque paire d'éléments de G, notés g_1 et g_2, il existe un unique élément $g_1 \star g_2$.
2. Quelque soit g_1, g_2 et g_3 dans G, on a $(g_1 \star g_2) \star g_3 = g_1 \star (g_2 \star g_3)$.
3. Il existe un élément spécial de G, appelé *neutre* et noté e tel que $g \star e = e \star g = g$.
4. Pour chaque $g \in G$, il existe un inverse, noté g^{-1} tel que $g \star g^{-1} = g^{-1} \star g = e$.

Exemples de groupes

1. Les nombres entiers $\mathbb{Z}$, les nombres rationnels $\mathbb{Q}$, les nombres réels $\mathbb{R}$ et les nombres complexes $\mathbb{C}$ sont tous des groupes dont la loi de composition est l'addition. Le neutre est le zéro et les inverses sont les opposés.
2. Les fonctions réelles bijectives, dont le domaine de définition et le domaine d'arrivée sont $\mathbb{R}$, forment un groupe dont la loi de composition est la composition de fonctions. Le neutre est l'application $\mathrm{id} : \mathbb{R} \to \mathbb{R};\ x \mapsto x$. Les inverses sont les fonctions réciproques (c'est pour cette raison que les fonctions ont besoin d'être bijectives).
3. Les permutations de n éléments, noté $\mathrm{Sym}(n)$, forment un groupe à $n!$ éléments pour lequel la loi de composition est la composition (de fonctions).
4. On découvrira en exercices les groupes de rotations et de symétries des polygones à n côtés (appelés aussi n-gones) et les groupes de rotation des cinq solides platoniciens.

10.3 Les actions de groupes

Une action d'un groupe G sur un ensemble E est une application

$$\begin{array}{ccc} G \times E & \longrightarrow & E \\ (g; x) & \longmapsto & g \cdot x \end{array}$$

qui satisfait les propriétés suivantes.

$$e \cdot x = x \text{ pour tout } x \in E \text{ (e est l'élément neutre de G)}$$

$$g \cdot (h \cdot x) = (gh) \cdot x \text{ pour tout } g, h \in G \text{ et } x \in E$$

Exemples d'actions de groupes

1. Le groupe multiplicatif $\mathbb{R} \setminus \{0\}$ agit sur les vecteurs du plan par multiplication.

$$\begin{array}{ccc} \mathbb{R} \setminus \{0\} \times \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ (\lambda; \vec{v}) & \longmapsto & \lambda \vec{v} \end{array}$$

2. Les groupes de rotations et de symétries agissent sur les ensembles dont ils sont le groupe de rotations et de symétries.

Définitions

Lorsqu'on a une action d'un groupe G sur un ensemble E, on peut définir deux ensembles.

1. Soit $x \in E$. L'*orbite de x* est l'ensemble

$$\text{Orb}(x) = \{g \cdot x : g \in G\}$$

2. Soit $g \in G$. L'*ensemble des points fixes par g* est l'ensemble

$$\text{Fix}(g) = \{x \in E : g \cdot x = x\}$$

Remarques fondamentales

1. Lorsqu'un groupe G agit sur un ensemble E, chaque élément g de G permute les éléments de E.

   En effet, une action associe à chaque $g \in G$, une bijection de E dans E.

2. Les orbites partitionnent l'ensemble E (en d'autres termes les orbites sont des ensembles disjoints dont la réunion est l'ensemble E).

Notation

Si E est un ensemble, on note $|E|$ le nombre de ses éléments.

Théorème de Burnside (sans preuve)

La moyenne du nombre de points fixes est égale au nombre d'orbites n de l'action.

Autrement dit :

$$n = \frac{1}{|G|} \sum_{g \in G} |\text{Fix}(g)|$$

10.4 Les théorèmes de Pólya

L'indicateur des cycles

On considère un ensemble E à m éléments sur lequel un groupe G agit. Par cette action, chaque élément g de G permute les éléments de E avec une permutation de type $(t_1, t_2, \ldots, t_m)$.

On peut ainsi définir l'*indicateur des cycles* de cette action de la manière suivante.

$$\mathcal{Z}(z_1, z_2, \ldots, z_m) = \frac{1}{|G|} \sum_{g \in G} z_1^{t_1} z_2^{t_2} \cdots z_m^{t_m}$$

Le théorème de Pólya 1

Le nombre de colorations inéquivalentes d'un ensemble E de m objets (sous l'action d'un groupe G) à l'aide de k couleurs est $\mathcal{Z}(k, \ldots, k)$.

Le théorème de Pólya 2

On cherche à colorier un ensemble E de m objets (sous l'action d'un groupe G) à l'aide de k couleurs.

Le coefficient $x_1^{j_1} x_2^{j_2} \cdots x_k^{j_k}$ du polynôme

$$\mathcal{Z}(\ x_1 + x_2 + \cdots + x_k\ ,\ x_1^2 + x_2^2 + \cdots + x_k^2\ ,\ \ldots\ ,\ x_1^m + x_2^m + \cdots + x_k^m\ )$$

est égal au nombre de colorations inéquivalentes de E qui utilisent j_1 fois la couleur x_1, j_2 fois la couleur x_2, …, et j_k fois la couleur x_k.

Preuve du théorème de Pólya 1

On considère l'ensemble[2] C de toutes les colorations de E. L'action de groupe de G sur l'ensemble E induit une action de groupe de G sur l'ensemble C dont le nombre de colorations inéquivalentes est exactement le nombre d'orbites de cette action.

Par la formule de Burnside, le nombre d'orbites cherché est égal à

$$\frac{1}{|G|} \sum_{g \in G} |\mathrm{Fix}(g)|$$

Or, $\mathrm{Fix}(g)$ est l'ensemble des colorations de E qui sont fixes par l'élément $g \in G$. Une coloration est fixe si et seulement si dans chaque cycle de g, les éléments de E ont la même couleur. Par conséquent, si on note $(t_1, t_2, \ldots, t_m)$ le type de la permutation de g, l'ensemble $\mathrm{Fix}(g)$ contient exactement $k^{t_1+t_2+\cdots+t_m} = k^{t_1} k^{t_2} \cdots k^{t_m}$ éléments (on a k choix de couleurs pour chaque cycle).

Ainsi le nombre de colorations inéquivalentes est

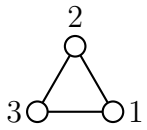
$$\mathcal{Z}(k, \ldots, k) = \frac{1}{|G|} \sum_{g \in G} k^{t_1} k^{t_2} \cdots k^{t_m}$$

□

2. Il s'agit de l'ensemble des applications de E dans $\{1, 2, 3, \ldots, k\}$. Chacune de ces applications assigne à un élément de E, une couleur représentée par un nombre.

Exemple

On a deux sortes de perles : les blanches et les noires. On cherche le nombre de colliers à trois perles que l'on peut construire. Afin de faire apparaître un groupe de symétrie pour représenter les différents mouvements que l'on peut faire subir à un tel collier, on le représente à l'aide d'un triangle régulier dont les sommets sont les perles.



L'action du groupe de rotations et de symétries du triangle sur ses sommets permet de tenir compte des colorations inéquivalentes des perles.

Établissons l'indicateur des cycles de cette action.

Élément du groupe	permutation associée	type de la permutation
identité	id	$(3;0;0)$
rotation de $\frac{2\pi}{3}$	(123)	$(0;0;1)$
rotation de $\frac{4\pi}{3}$	(132)	$(0;0;1)$
symétrie de sommet 1	(23)	$(1;1;0)$
symétrie de sommet 2	(13)	$(1;1;0)$
symétrie de sommet 3	(12)	$(1;1;0)$

Ainsi, l'indicateur est

$$\mathcal{Z}(z_1,z_2,z_3)=\frac{1}{|G|}\sum_{g\in G} z_1^{t_1}z_2^{t_2}z_3^{t_3}=\frac{1}{6}\left(z_1^3+2z_3+3z_1z_2\right)$$

Astuce : il y a deux moyens pour vérifier que l'indicateur n'a pas l'air d'être faux.

1. La somme des coefficients de chaque monôme doit faire $|G|$ (car $Z(1,\ldots,1)=1$). Ici : $1+2+3=6$.
2. On retrouve la formule $t_1+2t_2+3t_3+\cdots+mt_m=m$ concernant le type de chaque permutation sur chaque monôme. Ici, on a bien $1\cdot 3=3$ pour z_1^3, $3\cdot 1=3$ pour z_3^1 et $1\cdot 1+2\cdot 1=3$ pour $z_1^1z_2^1$.

Par Pólya 1, le nombre de colorations à deux couleurs est donné par

$$\mathcal{Z}(2,2,2)=\frac{1}{6}\left(2^3+2\cdot 2+3\cdot 2\cdot 2\right)=\frac{1}{6}\cdot 24=4$$

En appliquant Pólya 2, on établit l'inventaire des figures.

$$\begin{aligned}\mathcal{Z}(x_1+x_2,x_1^2+x_2^2,x_1^3+x_2^3) &= \tfrac{1}{6}\left((x_1+x_2)^3+2(x_1^3+x_2^3)+3(x_1+x_2)(x_1^2+x_2^2)\right)\\ &= \cdots = x_1^3+x_1^2x_2+x_1x_2^2+x_2^3\end{aligned}$$

Si x_1 correspond aux perles blanches et x_2 aux perles noires. On lit sur l'inventaire des figures qu'il y a 1 collier avec 3 perles blanches et 0 perles noires, 1 collier avec 2 perles blanches et 1 perles noires, 1 collier avec 1 perles blanches et 2 perles noires et 1 collier avec 0 perles blanches et 3 perles noires.

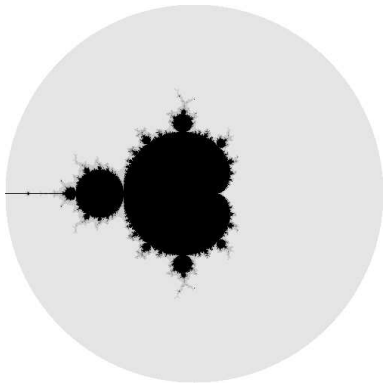
Chapitre 11

Nombres complexes

11.1 Introduction

Les nombres complexes ont été introduits durant la renaissance au XVIe siècle par les mathématiciens italiens Girolamo Cardano (Jérome Cardan pour les français), Raphaël Bombelli, Nicolo Fontana, dit Tartaglia, et Ludovico Ferrari afin d'exprimer les solutions des équations du troisième degré en toute généralité par les formules de Cardan ainsi que les solutions des équations du quatrième degré (méthode de Ferrari).

En mathématiques, les nombres complexes sont utilisés dans le traitement du signal dans les séries de Fourier ; dans le calcul intégral avec les intégrales par résidus ; dans les fractales pour définir le magnifique *ensemble de Mandelbrot* représenté ci-dessous.



En physique, les nombres complexes sont utilisés pour décrire le comportement d'oscillateurs électriques ou les phénomènes ondulatoires en électromagnétisme.

En économie, les nombres complexes mettent en évidence des phénomènes d'oscillations rencontrés dans des problèmes de cycle et de stabilité des équilibres.

Les zéros d'un polynôme du troisième degré (sans preuve)

On considère l'équation $ax^3 + bx^2 + cx + d = 0$ où a, b, c et d sont des nombres réels avec $a \neq 0$. Comme pour les équations du deuxième degré, mais avec plus de difficulté, on peut définir un discriminant donné par

$$\Delta = 18abcd - 4b^3d + b^2c^2 - 4ac^3 - 27a^2d^2$$

Si $\Delta > 0$, l'équation a trois solutions réelles ; si $\Delta = 0$, l'équation n'a qu'une solution réelle (en fait, elle est équivalente à $a(x - x_0)^3 = 0$ où x_0 est le zéro réel) ; si $\Delta < 0$, l'équation a exactement une solution réelle et deux solutions complexes.

11.2 Les nombres complexes

11.2.1 Construction géométrique du nombre imaginaire

Jean-Robert Argand, né le 18 juillet 1768 à Genève et mort le 13 août 1822 à Paris, était un mathématicien suisse.

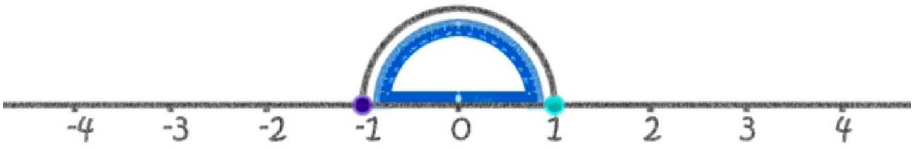
En 1806, alors qu'il tient une librairie à Paris, il publie une interprétation géométrique des nombres complexes, dans un texte intitulé «Essai sur une manière de représenter les quantités imaginaires par des constructions géométriques». Pour cette raison, le plan, vu comme ensemble des nombres complexes, est parfois appelé le plan d'Argand.

Néanmoins, le plan complexe est plus souvent appelé plan de Gauss, ou aussi plan d'Argand-Gauss.

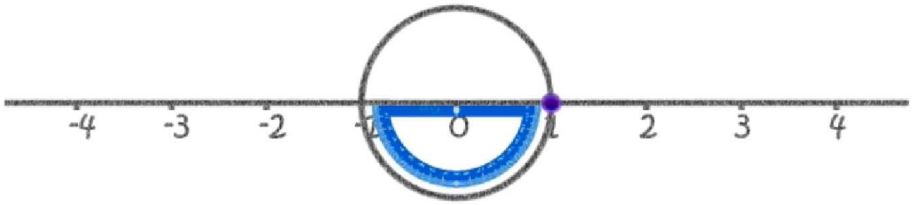
Les images ci-dessous sont issues du film «Dimensions» de Jos Leys, Étienne Ghys et Aurélien Alvarez, que l'on peut gratuitement visualiser ou télécharger en plusieurs langues sur `http://www.dimensions-math.org/`.

Dans son essai, Jean-Robert Argand a observé le phénomène suivant sur la droite réelle.

Lorsque qu'on multiplie 1 par -1 on effectue une rotation d'un demi-tour.



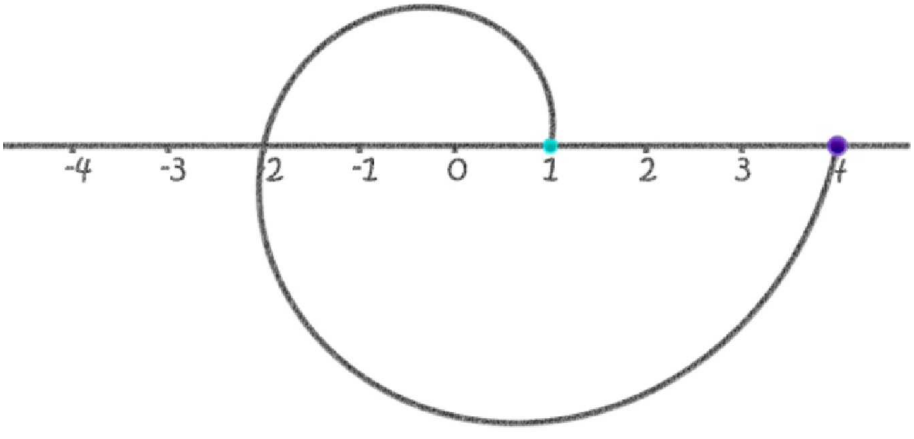
En multipliant deux fois par -1 on effectue une rotation d'un tour complet.



On a donc $1 \cdot (-1) \cdot (-1) = 1 \iff (-1)^2 = 1$.

On dit que 1 est le carré de -1, et que -1 est une racine carrée de 1.

Lorsqu'on multiplie deux fois par -2, on obtient 4.



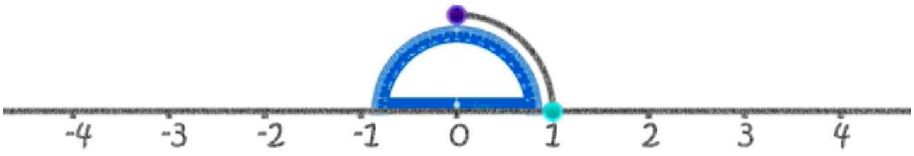
On a donc $1 \cdot (-2) \cdot (-2) = 4 \iff (-2)^2 = 4$.

On dit que 4 est le carré de -2, et que -2 est une racine carrée de 4.

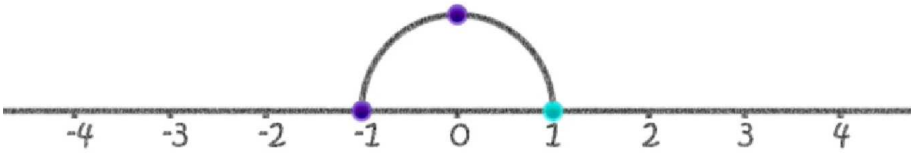
On sait que les carrés sont des nombres positifs ou nuls,
et donc que -1 n'a aucune racine carrée.

L'idée de Jean-Robert Argand est la suivante.

Il imagine un point en dehors de la droite réelle.
Pour aller sur ce point, on effectue une rotation d'un quart de tour.

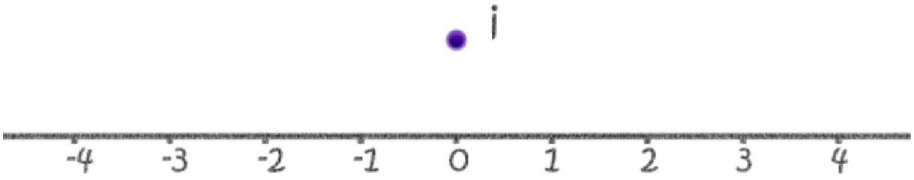


Si on effectue deux fois ce quart de tour, on arrive sur le nombre -1.

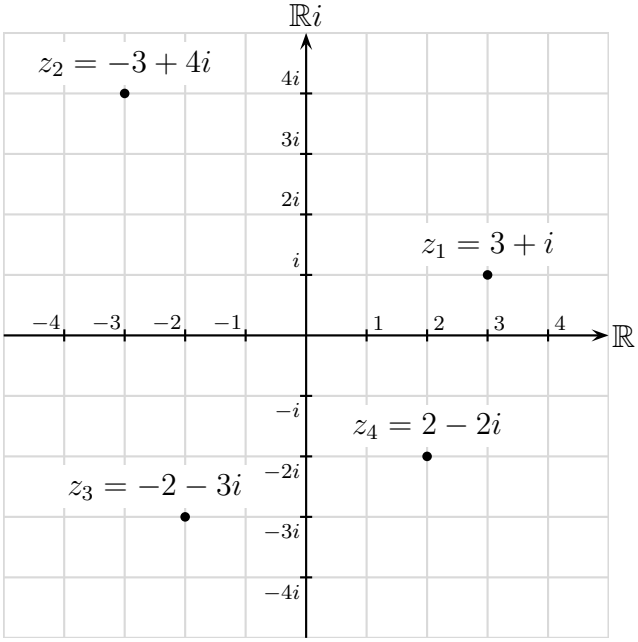


Ainsi, on a réussi à créer géométriquement une racine carrée de -1

Ce nouveau nombre, qui n'est pas un nombre réel,
a été appelé *nombre imaginaire*, et noté i.



Cela a permis de créer un nouvel ensemble : les nombres complexes,
représentés ainsi dans le plan d'Armand-Gauss.



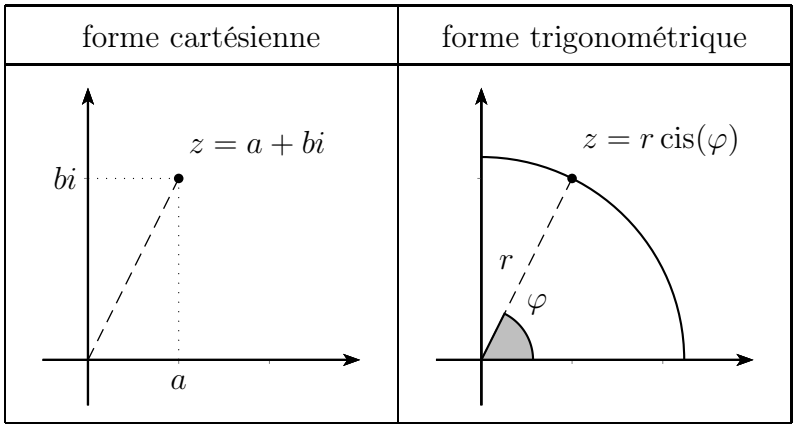
On en tire un fait **important** : dans le plan de Gauss, la multiplication par le nombre i correspond à une rotation de 90° dans le sens trigonométrique.

L'ensemble des nombres complexes est défini à l'aide du nombre i et des nombres réels de la façon suivante.

$$\mathbb{C} = \{a + b\,i : a, b \in \mathbb{R}\} \qquad \text{avec} \qquad i^2 = -1$$

11.2.2 Les deux façons de décrire un nombre complexe

On considère le nombre complexe z. On peut représenter les nombres complexes par des points dans le plan de Gauss et les décrire des deux manières différentes suivantes.



Définitions

1. La *forme cartésienne* d'un nombre complexe est donnée par $z = a + bi$ où
   - le nombre a s'appelle la *partie réelle* du nombre complexe z, notée $\mathrm{Re}(z)$;
   - le nombre b s'appelle la *partie imaginaire* du nombre complexe z, notée $\mathrm{Im}(z)$.
2. La *forme trigonométrique* d'un nombre complexe est donnée par $z = r\,\mathrm{cis}(\varphi)$ où
   - le rayon r du cercle sur lequel z se trouve est appelé le *module de* z, noté $|z|$;
   - l'angle trigonométrique φ qui définit z est appelé l'*argument de* z, noté $\arg(z)$.

Pour passer de la forme cartésienne à la forme trigonométrique

1. Par Pythagore, on a $r = \sqrt{a^2 + b^2}$.
2. Par «tan-opp-adj», on a $\varphi = \tan^{-1}\left(\frac{b}{a}\right)$ si z est dans le premier cadran. S'il n'est pas dans le premier cadran, on calcule $\tan^{-1}\left|\frac{b}{a}\right|$ et on fait un schéma pour savoir comment corriger l'angle trouvé pour avoir l'argument φ.

Pour passer de la forme trigonométrique à la forme cartésienne

1. Par définition de $\cos(\varphi)$ et un facteur d'homothétie r, on a $a = r\cos(\varphi)$.
2. Par définition de $\sin(\varphi)$ et un facteur d'homothétie r, on a $b = r\sin(\varphi)$.

On comprend maintenant que $z = a + bi = r\cos(\varphi) + r\sin(\varphi)i = r\big(\cos(\varphi) + i\sin(\varphi)\big)$ se note simplement $z = r\,\mathrm{cis}(\varphi)$ (c pour cos, i pour i et s pour sin).

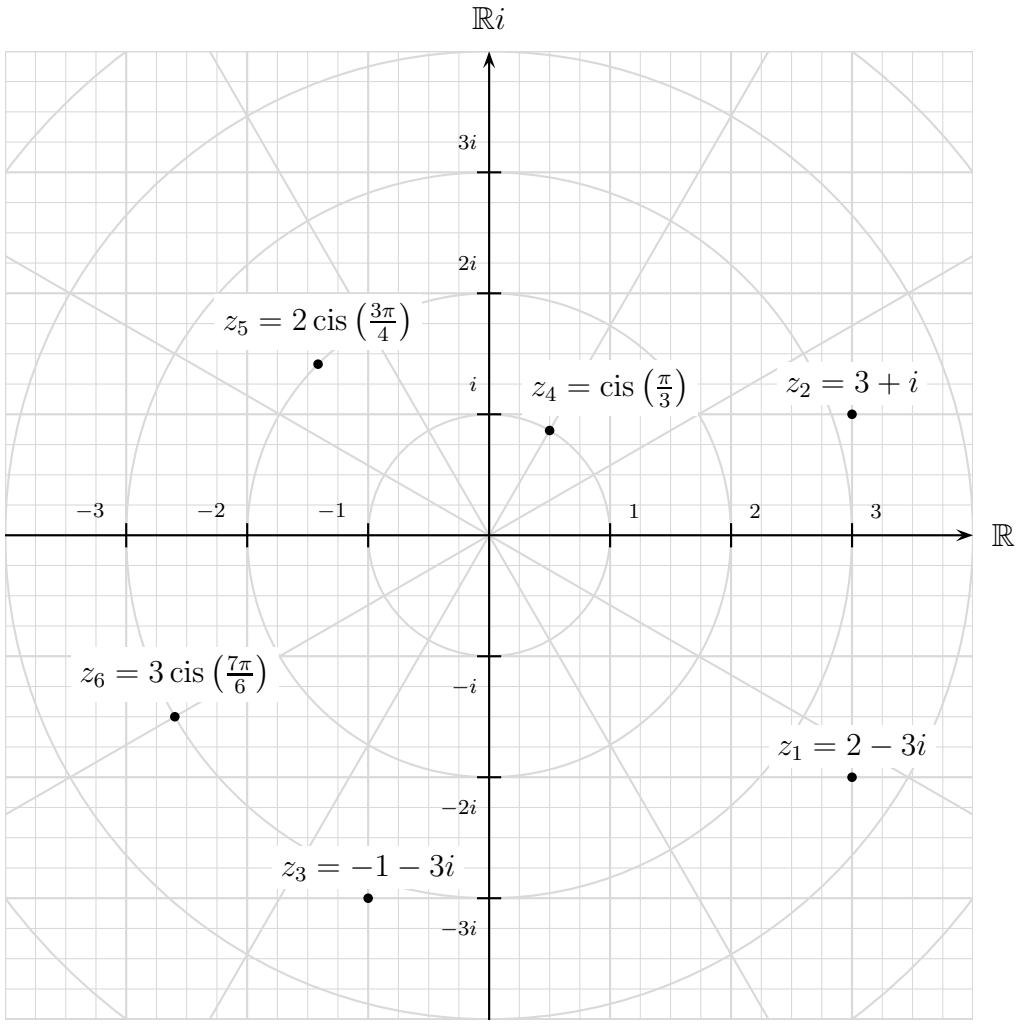
La notation avec l'exponentielle

En travaillant avec les développements de Maclaurin (se référer au chapitre sur les séries et développements de Taylor du cours OS sur le site web `www.vive-les-maths.net`), on peut démontrer que

$$r\,\mathrm{cis}(\varphi) = re^{i\varphi}$$

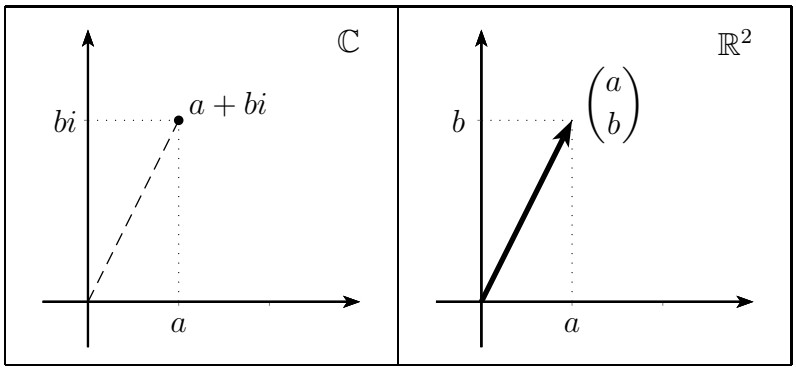
où e est le nombre d'Euler qui vaut environ 2.71828.

Voici le plan de Gauss avec une superposition des repères cartésien et trigonométrique, et quelques nombres complexes donnés sous leur forme cartésienne ou trigonométrique.



11.2.3 L'addition de deux nombres complexes

La bijection $\mathbb{C} \to \mathbb{R}^2; a + bi \mapsto \binom{a}{b}$ permet de voir les nombres complexes comme des vecteurs du plan (attachés à l'origine).

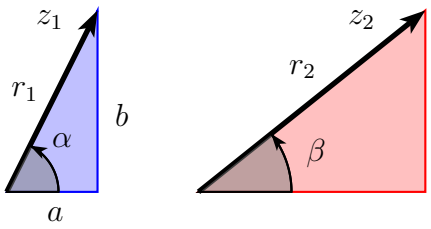


On additionne géométriquement les nombres complexes comme les vecteurs du plan $\mathbb{R}^2$. On a ainsi la correspondance

$(a + bi) + (c + di) = (a + c) + (b + d)i$	$\binom{a}{b} + \binom{c}{d} = \binom{a+c}{b+d}$

11.2.4 La multiplication de deux nombres complexes

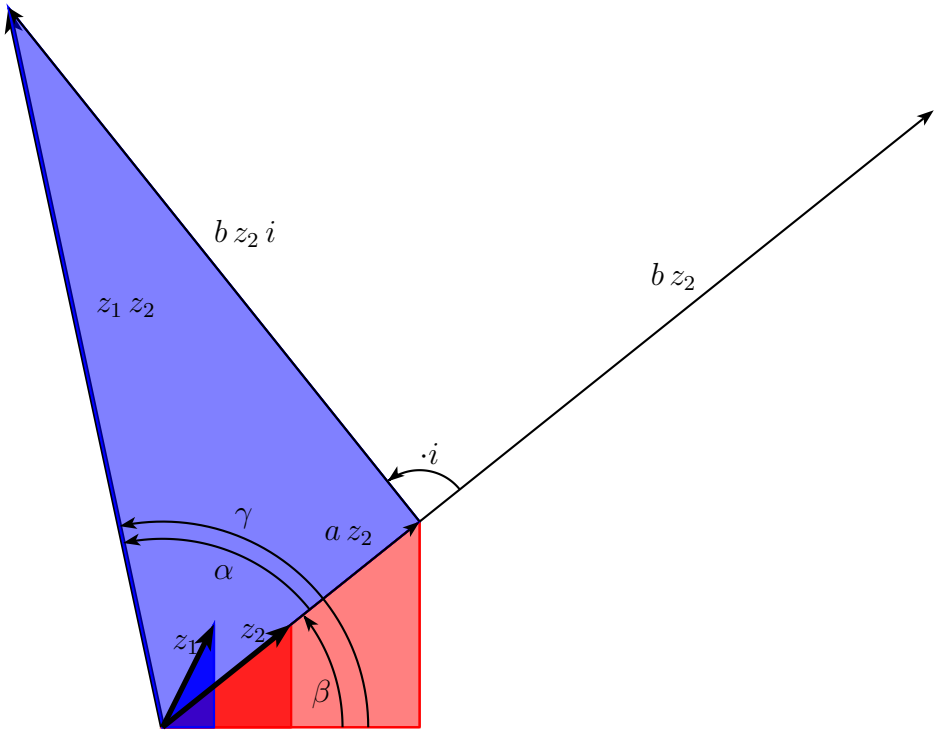
On considère les nombres complexes $z_1 = a + bi = r_1 \operatorname{cis}(\alpha)$ et $z_2 = r_2 \operatorname{cis}(\beta)$. On peut représenter les nombres complexes par des points dans le plan de Gauss ou même des vecteurs en reliant les points depuis l'origine.



Grâce au calcul suivant

$$z_1\, z_2 = (a + bi)\, z_2 = a\, z_2 + b\, z_2\, i$$

on voit que cette multiplication revient à additionner a fois le nombre complexe z_2 et b fois le nombre complexe z_2 tourné d'un quart de tour. On peut illustrer cette situation ainsi



On voit ainsi que $|z_1\, z_2| = |z_1| \cdot |z_2|$. En effet, par Pythagore dans le grand triangle bleu, la longueur de $z_1\, z_2$ vaut $\sqrt{a^2 + b^2}$ fois la longueur de z_2 (et $\sqrt{a^2 + b^2}$ correspond à la longueur de z_1).

On voit aussi que $\arg(z_1\, z_2) = \arg(z_1) + \arg(z_2)$. En effet, le grand triangle bleu est semblable au triangle construit à partir de z_1, on a donc $\gamma = \alpha + \beta$.

Ainsi, on a la formule

$$r_1 \operatorname{cis}(\alpha) \cdot r_2 \operatorname{cis}(\beta) = r_1 r_2 \operatorname{cis}(\alpha + \beta)$$

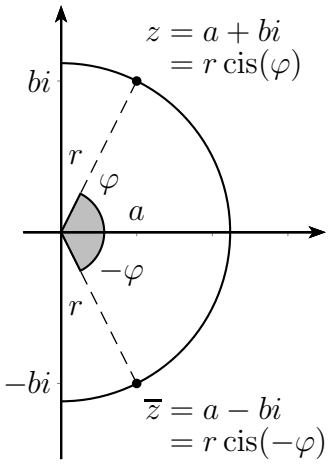
Elle se résume ainsi : «quand on multiplie deux nombres complexes, leurs modules se multiplient et leurs arguments s'additionnent».

Ce qui donne naturellement pour la notation exponentielle

$$r_1 e^{i\alpha} \cdot r_2 e^{i\beta} = r_1 r_2 \cdot e^{i(\alpha+\beta)}$$

11.2.5 Le conjugué d'un nombre complexe

Si z est un nombre complexe, son *conjugué*, noté $\overline{z}$, est le point symétrique par l'axe réel dans le plan de Gauss.



11.2.6 La division de deux nombres complexes

L'idée fondamentale sous-jacente à la division de deux nombres complexes est

«Pour calculer $\frac{z_1}{z_2}$, on amplifie la fraction par le conjugué de z_2»

Avec la forme cartésienne : si $z_1 = a + bi$ et $z_2 = c + di$, on a

$$\frac{z_1}{z_2} = \frac{a+bi}{c+di} \overset{\text{idée fond.}}{=} \frac{(a+bi)(c-di)}{(c+di)(c-di)} = \frac{ac - bdi^2 + bci - adi}{c^2 - (di)^2}$$

$$\overset{i^2=-1}{=} \frac{ac + bd + \big(bc - ad\big)i}{c^2 + d^2}$$

$$= \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}i$$

Avec la forme trigonométrique : si $z_1 = r_1\,\mathrm{cis}(\alpha)$ et $z_2 = r_2\,\mathrm{cis}(\beta)$, on a

$$\frac{z_1}{z_2} = \frac{r_1\,\mathrm{cis}(\alpha)}{r_2\,\mathrm{cis}(\beta)} \overset{\text{idée fond.}}{=} \frac{r_1\,\mathrm{cis}(\alpha)\cdot r_2\,\mathrm{cis}(-\beta)}{r_2\,\mathrm{cis}(\beta)\cdot r_2\,\mathrm{cis}(-\beta)}$$

quand on multiplie deux nombres complexes,
leurs modules se multiplient et leurs arguments s'additionnent

$$= \frac{r_1 r_2\,\mathrm{cis}\left(\alpha + (-\beta)\right)}{r_2 r_2\,\mathrm{cis}\left(\beta + (-\beta)\right)} = \frac{r_1\,\mathrm{cis}(\alpha-\beta)}{r_2\,\mathrm{cis}(0)} = \frac{r_1}{r_2}\,\mathrm{cis}(\alpha-\beta)$$

car on a $\mathrm{cis}(0) = \cos(0) + i\sin(0) = 1 + 0i = 1$

Ainsi : «quand on divise deux nombres complexes, leurs modules se divisent et leurs arguments se soustraient».

Ce qui donne naturellement pour la notation exponentielle

$$\frac{r_1 e^{i\alpha}}{r_2 e^{i\beta}} = \frac{r_1}{r_2}\cdot e^{i(\alpha-\beta)}$$

11.2.7 La formule de Moivre

Le mathématicien français Abraham de Moivre (1667-1754) a trouvé la formule trigonométrique suivante.

$$\begin{aligned} &\big(\cos(\varphi)+i\sin(\varphi)\big)^n=\cos(n\varphi)+i\sin(n\varphi) && \text{pour tout } n\geqslant 1\\ \iff\ &(\text{cis}(\varphi))^n=\text{cis}(n\varphi) && \text{pour tout } n\geqslant 1 \end{aligned}$$

Ce qui donne naturellement pour la notation exponentielle

$$(e^{i\varphi})^n=e^{in\varphi} \quad \text{pour tout } n\geqslant 1$$

11.2.8 Les racines énièmes d'un nombre complexe

Lorsqu'on considère une équation sur les nombres complexes, on préfère noter l'inconnue z (plutôt que de la noter x comme on le ferait pour une équation sur les nombres réels).

Définition

Soit z_0 un nombre complexe.
Les solutions de l'équation $z^n=z_0$ sont appelées *racines n-ième* du nombre complexe z_0.

Attention

Un nombre complexe non nul admet n racines n-ièmes complexes et cela soulève une contradiction si on continue d'utiliser la notation $\sqrt[n]{\ }$ pour les racines énièmes.

$$1=\sqrt{1}=\sqrt{(-1)\cdot(-1)}=\sqrt{-1}\cdot\sqrt{-1}=\big(\sqrt{-1}\big)^2=-1$$

Dans les nombres complexes, i et $-i$ sont deux racines carrées de -1. Dans la ligne ci-dessus, $\sqrt{-1}$ représente une fois i et une fois $-i$, car le calcul ci-dessous est juste.

$$1=\sqrt{1}=\sqrt{(-1)\cdot(-1)}=i\cdot(-i)=-i^2=1$$

Il n'y a donc plus *unicité* pour les racines énièmes (dans le cas des nombres réels, $\sqrt[n]{a}$ représente l'unique solution de $x^n=a$, qui est la solution positive ou nulle lorsqu'il y a deux solutions réelles). Dans les nombres complexes, les relations $<$ et $>$ perdent leur sens, et ainsi on ne doit plus utiliser la notation $\sqrt[n]{\ }$ (sauf si elle a un sens dans les nombres réels, ce qui n'est pas le cas de $\sqrt{-1}$ par exemple).

Résultat (démonstration en exercice grâce à de Moivre)

Soit z_0 un nombre complexe non nul. Alors z_0 possède n racines n-ième distinctes.

Du point de vue de l'informatique.

La plupart des logiciels ou calculatrices font le choix suivant pour l'argument.

1. Si la partie imaginaire de z_0 est positive ou nulle, alors $z_0=re^{i\varphi}$ avec $\varphi\in[0,\pi]$.
2. Si la partie imaginaire de z_0 est négative, alors $z_0=re^{-i\varphi}$ avec $\varphi\in\,]0,\pi[$.

De cette façon lorsque l'on active la fonction ou la touche de la racine n-ième, ils livrent la solution z_0 avec le plus petit argument. Le lecteur profitera de l'occasion pour regarder comment sa calculatrice calcule la racine cubique de -1.

Sur l'extraction de racines carrées avec la forme cartésienne

Lorsqu'on cherche les racines carrées du nombre complexe $a+bi$, cela revient à chercher à résoudre l'équation $z^2 = a+bi$. En posant $z = x+yi$, cette équation s'écrit

$$(x+yi)^2 = a+bi \iff x^2-y^2+2xyi = a+bi \overset{\text{unicité de l'écriture}}{\iff} \begin{cases} x^2-y^2 &= a \\ 2xy &= b \end{cases}$$

On peut se contenter de résoudre ce système de deux équations à deux inconnues dans les nombres réels, mais il existe une manière de simplifier cette résolution.

On utilise la propriété du module qui dit que $|z^2| = |z|^2$ pour écrire

$$|z^2| = |a+bi| \iff |z|^2 = |a+bi| \overset{z=x+yi}{\iff} |x+yi|^2 = |a+bi| \iff x^2+y^2 = \sqrt{a^2+b^2}$$

En ajoutant cette équation au système d'équations précédent, on obtient un système de trois équations à deux inconnues extrêmement facile à résoudre et l'équation $2xy = b$ ne sera alors utile que pour dire si x et y sont de signes opposés ou de même signe.

Sur l'extraction de racines énièmes avec la forme trigonométrique

On démontre, en exercice, que les racines n-ièmes de $z = r\,\text{cis}(\varphi)$ sont données par

$$\sqrt[n]{r} \cdot \text{cis}\left(\frac{\varphi}{n} + \frac{2\pi k}{n}\right) = \sqrt[n]{r} \cdot e^{i\left(\frac{\varphi}{n}+\frac{2\pi k}{n}\right)} \quad \text{pour} \quad k \in \{0,1,2,\ldots,n-1\}$$

11.3 Résolution d'équations

11.3.1 Le théorème fondamental de l'algèbre

L'intérêt principal de l'ensemble des nombres complexes réside dans le théorème suivant.

Théorème

Tout polynôme $p(z)$ de degré $n \geqslant 1$ et à coefficients dans $\mathbb{C}$ admet n racines (non nécessairement distinctes) dans $\mathbb{C}$.

11.3.2 Résolution d'équations du premier degré

Il y a aucune différence par rapport aux résolutions d'équation du premier degré dans les nombres réels.

11.3.3 Résolution d'équations du deuxième degré

Il y n'a qu'une seule subtilité qui différentie la résolution des équations du deuxième degré dans les nombres réels de la résolution dans les nombres complexes : le symbole $\sqrt{\Delta}$ ne s'utilise que lorsque $\Delta \geqslant 0$ (ce qui signifie implicitement que $\Delta \in [0,+\infty] \subset \mathbb{R}$). Si $\Delta \notin [0,+\infty]$, alors ses racines carrées existent, mais ne peuvent pas être notées en utilisant le symbole $\sqrt{\ }$. Il faut donc nommer, par exemple r, une racine carrée de Δ (dans ce cas, la deuxième racine carrée de Δ est $-r$) et la formule de Viète devient

$$az^2+bz+c = 0 \iff z = \frac{-b \pm r}{2a}$$

Le lecteur désirant comprendre cette formule est prié de se référer au chapitre 4, section 3 du cours de discipline fondamentale se trouvant sur `www.vive-les-maths.net`.

11.3.4 Résolution d'équations du troisième degré

Pour résoudre les équations du troisième degré de manière générale, il faut passer par les nombres complexes.

Première étape

On transforme l'équation $az^3 + bz^2 + cz + d = 0$ avec $a \neq 0$ en équation de la forme

$$y^3 + py + q = 0$$

Pour cela, on pose $\boxed{z = y - \dfrac{b}{3a}}$ (remarquer que l'on peut diviser par a puisque $a \neq 0$).

En effet, lorsqu'on effectue la substitution, l'équation devient

$$\begin{aligned} az^3 + bz^2 + cz + d &= ay^3 + \left(c - \frac{b^2}{3a}\right) y + \left(\frac{2b^3}{27a^2} + d - \frac{bc}{3a}\right) = 0 \\ &= a\left(y^3 + \left(\frac{c}{a} - \frac{b^2}{3a^2}\right) y + \left(\frac{2b^3}{27a^3} + \frac{d}{a} - \frac{bc}{3a^2}\right)\right) = 0 \end{aligned}$$

L'équation est donc bien équivalente à une équation de la forme

$$\boxed{y^3 + py + q = 0 \quad \text{avec} \quad p = \frac{c}{a} - \frac{b^2}{3a^2} \quad \text{et} \quad q = \frac{2b^3}{27a^3} + \frac{d}{a} - \frac{bc}{3a^2}}$$

Deuxième étape

On trouve une formule permettant de résoudre toutes les équations du troisième degré de la forme

$$y^3 + py + q = 0$$

Comme on sait déjà résoudre cette équation lorsque p ou q sont nuls, on va supposer par la suite, qu'ils ne sont pas nuls.

L'idée, très astucieuse, consiste à transformer cette équation en un système d'équations à deux inconnues. Pour cela, on utilise la substitution $\boxed{y = u + v}$ et on ajoute la condition $\boxed{3uv = -p}$ qui est là pour simplifier l'expression obtenue lorsqu'on remplace y par $u+v$. Ainsi, on a

$$\left\{ \begin{array}{c} (u+v)^3 + p(u+v) + q = 0 \\ 3uv = -p \end{array} \right. \iff \left\{ \begin{array}{c} u^3 + v^3 = -q \\ 3uv = -p \end{array} \right. \iff \left\{ \begin{array}{c} u^3 + v^3 = -q \\ u^3 \cdot v^3 = -\frac{p^3}{27} \\ 3uv = -p \end{array} \right.$$

En effet, en développant la première équation et en remplaçant p par $-3uv$, on trouve

$$u^3 + 3u^2v + 3uv^2 + v^3 - 3u^2v - 3uv^2 + q = 0$$

On peut ensuite simplifier cette expression qui devient

$$u^3 + v^3 + q = 0$$

La deuxième équivalence est obtenue en élevant la deuxième équation au cube. On est obligé de conserver l'équation que l'on a élevé au cube, car dans les nombres complexes, la fonction $f(z) = z^3$ n'est pas bijective.

Cette idée nous a permis d'introduire deux nombres complexes u et v tels que leur cube satisfait

$$\begin{cases} u^3 + v^3 = -q \\ u^3 \cdot v^3 = -\frac{p^3}{27} \end{cases} \qquad \text{et} \qquad 3uv = -p$$

On constate ici, que u^3 et v^3 sont solutions de l'équation du deuxième degré suivante.

$$(x - u^3)(x - v^3) = x^2 - (u^3 + v^3)x + u^3 \cdot v^3 = x^2 + qx - \frac{p^3}{27} = 0$$

Afin d'avoir un discriminant qui soit plus facile à mémoriser, on modifie très légèrement cette équation en la divisant par 2.

$$\frac{1}{2}x^2 + \frac{q}{2}x - \frac{p^3}{2 \cdot 27} = 0$$

Le *discriminant* de cette équation est

$$\Delta = \frac{q^2}{4} + \frac{p^3}{27} = \left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3$$

En notant r pour une des deux racines carrées de Δ (rappelons que la deuxième racine carrée est $-r$), on arrive à exprimer u^3 et v^3 en fonction de p et de q en résolvant cette équation.

$$\boxed{u^3 = -\frac{q}{2} + r \quad \left(\text{et} \quad v^3 = -\frac{q}{2} - r\right) \quad \text{où } r \text{ est une racine carrée de } \Delta = \left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}$$

Pour trouver u, on extrait les trois racines cubiques de u^3, appelées u_1, u_2 et u_3. A chaque valeur de u va correspondre une unique valeur de v donnée par la relation

$$\boxed{3uv = -p}$$

On appellera respectivement v_1, v_2 et v_3 ces valeurs. Remarquons qu'on est obligé d'utiliser cette façon pour trouver les racines cubiques de v associées à u (car si elles étaient indépendantes, il aurait neuf combinaisons possibles pour $u + v$).

Les solutions y sont données par

$$\boxed{y_1 = u_1 + v_1} \quad \boxed{y_2 = u_2 + v_2} \quad \boxed{y_3 = u_3 + v_3}$$

On revient à z pour avoir les solutions de l'équation de départ.

$$\boxed{z_1 = u_1 + v_1 - \frac{b}{3a}} \quad \boxed{z_2 = u_2 + v_2 - \frac{b}{3a}} \quad \boxed{z_3 = u_3 + v_3 - \frac{b}{3a}}$$

La méthode de Cardan

En 1545, Cardan a publié cette méthode trouvée par Scipione del Ferro et Tartaglia dans un cas plus particulier.

Lorsque p et q sont des nombres réels, et que $\Delta > 0$, alors l'équation n'a qu'une solution réelle qui est donnée par

$$\sqrt[3]{-\frac{q}{2} + \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3}}$$

11.4 D'autres valeurs exactes de cosinus et de sinus

Cette idée provient de Paul Jolissaint (qui était professeur de mathématiques au lycée cantonal de Porrentruy). On va calculer d'abord la valeur exacte de $\cos\left(\frac{2\pi}{5}\right)$. Ce qui nous permettra grâce aux formules trigonométriques suivantes, vraies pour $0 \leqslant \alpha \leqslant \frac{\pi}{2}$, de trouver les valeurs exactes de $\sin(\frac{2\pi}{5})$, $\cos(\frac{\pi}{5})$ et de $\sin(\frac{\pi}{5})$.

$$\sin(\alpha) = \sqrt{1-\cos^2(\alpha)} \quad \text{et} \quad \cos(\alpha) = \sqrt{\frac{1+\cos(2\alpha)}{2}}$$

1. Calcul de la valeur exacte de $\cos\left(\frac{2\pi}{5}\right)$ et de $\sin\left(\frac{2\pi}{5}\right)$.

   Posons $z_0 = \cos(\frac{2\pi}{5}) + i\sin(\frac{2\pi}{5}) = e^{\frac{2i\pi}{5}}$. Ce z_0 satisfait l'équation $z^5 - 1 = 0$.

   Or, puisqu'on a la factorisation $z^5 - 1 = (z^4+z^3+z^2+z+1)(z-1)$ et que $z_0 \neq 1$, alors z_0 satisfait les équations équivalentes suivantes.

   $$z^4+z^3+z^2+z+1 = 0 \overset{:z^2}{\iff} z^2+z+1+\frac{1}{z}+\frac{1}{z^2} = 0$$
   $$\iff \left(z^2+\frac{1}{z^2}\right)+\left(z+\frac{1}{z}\right)+1 = 0 \overset{\text{astuce}}{\iff} \left(z^2+2+\frac{1}{z^2}\right)+\left(z+\frac{1}{z}\right)-1 = 0$$
   $$\iff \left(z+\frac{1}{z}\right)^2+\left(z+\frac{1}{z}\right)-1 = 0$$

   Or,
   $$x_0 = z_0 + \frac{1}{z_0} = e^{\frac{2i\pi}{5}} + e^{-\frac{2i\pi}{5}} = 2\cos\left(\frac{2\pi}{5}\right)$$

   Ainsi, x_0 est un nombre réel qui satisfait les équations équivalentes suivantes.

   $$x^2+x-1 = 0 \iff x = \frac{-1\pm\sqrt{5}}{2}$$

   Or, comme $\frac{2\pi}{5} < \frac{\pi}{2}$, on sait que $x_0 > 0$ et donc que $x_0 = \frac{-1+\sqrt{5}}{2}$. Ainsi

   $$\cos\left(\frac{2\pi}{5}\right) = \frac{\sqrt{5}-1}{4} \quad \text{et} \quad \sin\left(\frac{2\pi}{5}\right) = \frac{\sqrt{10+2\sqrt{5}}}{4}$$

   en utilisant la formule qui permet de trouver le sinus à partir du cosinus.

2. Calcul de la valeur exacte de $\cos\left(\frac{\pi}{5}\right)$ et de $\sin\left(\frac{\pi}{5}\right)$.

   On utilise une des formules citées en rappel pour obtenir l'expression suivante.

   $$\cos\left(\frac{\pi}{5}\right) = \sqrt{\frac{1+\cos(\frac{2\pi}{5})}{2}} = \sqrt{\frac{1+\frac{\sqrt{5}-1}{4}}{2}} = \sqrt{\frac{3+\sqrt{5}}{8}} = \sqrt{\frac{1+2\sqrt{5}+5}{16}}$$

   En repérant une identité remarquable, on obtient

   $$\cos\left(\frac{\pi}{5}\right) = \frac{\sqrt{5}+1}{4} \quad \text{et} \quad \sin\left(\frac{\pi}{5}\right) = \frac{\sqrt{10-2\sqrt{5}}}{4}$$

   en utilisant la formule qui permet de trouver le sinus à partir du cosinus.

11.5 Une projection stéréographique

Voici une projection stéréographique p de la sphère de Riemann, notée $\mathbb{S}^2 \setminus \{\mathcal{N}\}$ où $\mathcal{N}$ est le pôle nord, sur le plan de Gauss $\mathbb{C}$. Cette projection étant bijective, elle admet une fonction réciproque notée p^{-1}.

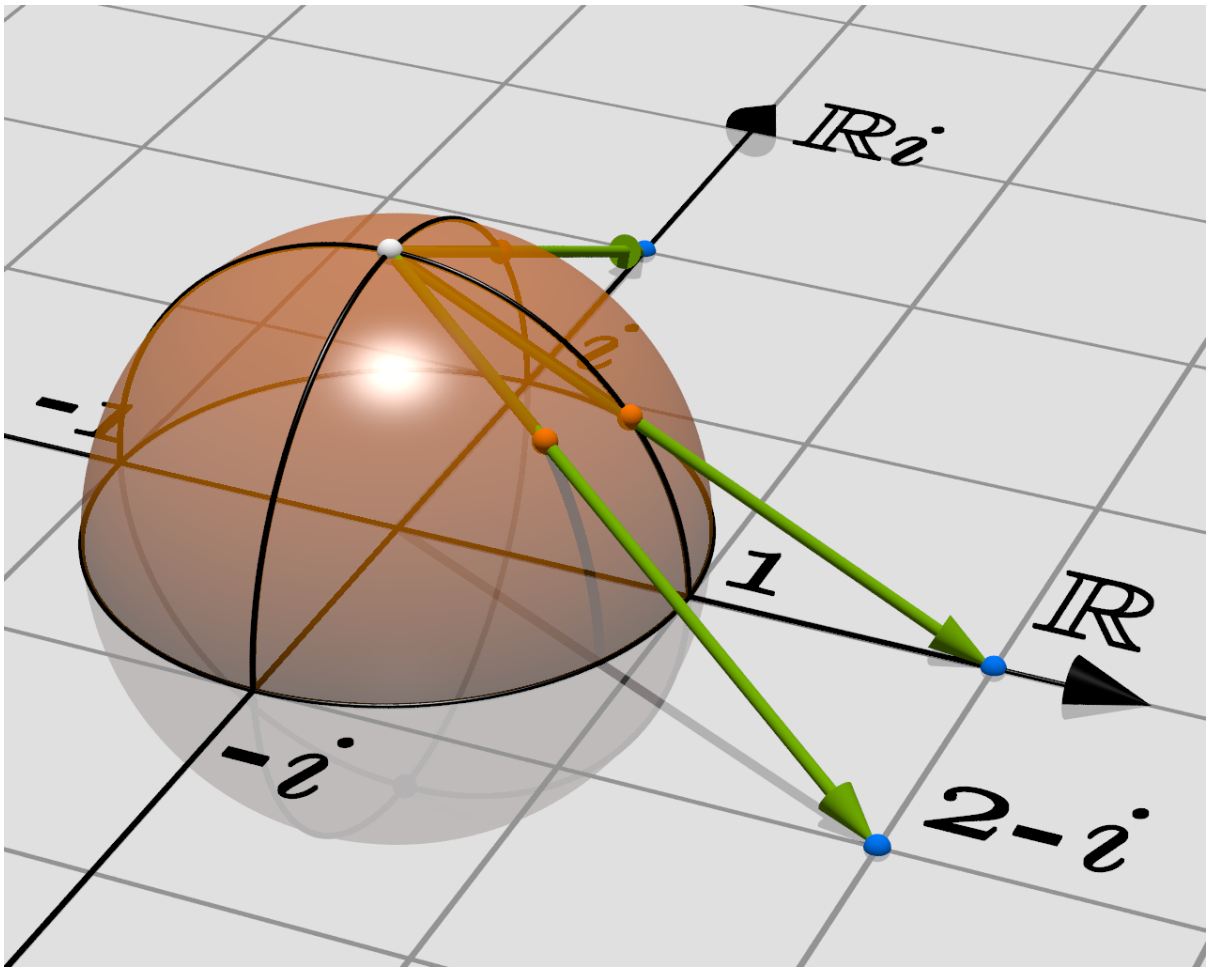
Les descriptions de ces deux applications sont

$$\begin{array}{rccc} p: & \mathbb{S}^2 \setminus \{\mathcal{N}\} & \to & \mathbb{C} \\ & P(x;y;z) & \mapsto & \dfrac{x+yi}{1-z} \end{array} \quad \text{et} \quad \begin{array}{rccc} p^{-1}: & \mathbb{C} & \to & \mathbb{S}^2 \setminus \{\mathcal{N}\} \\ & a+bi & \mapsto & P\left(\frac{2a}{a^2+b^2+1}; \frac{2b}{a^2+b^2+1}; \frac{a^2+b^2-1}{a^2+b^2+1}\right) \end{array}$$

Cette projection stéréographique envoie l'hémisphère nord sur l'extérieur du cercle trigonométrique, l'hémisphère sud sur l'intérieur du cercle trigonométrique et l'équateur sur le cercle trigonométrique.

Cette projection stéréographique consiste a prendre la droite passant par le pôle nord $\mathcal{N}$ et un autre point de la sphère P. La projection de ce point est l'intersection entre cette droite et le plan de Gauss (calcul de géométrie spatiale).

Voici une représentation graphique où les points oranges sont sur la sphère de Riemann $\mathbb{S}^2 \setminus \{\mathcal{N}\}$, le point blanc est le pôle nord $\mathcal{N}$ et les points bleus sont sur le plan de Gauss.



11.6 Les fonctions complexes

A partir de maintenant, la lettre D représente un domaine ou sous-ensemble de l'ensemble des nombres complexes $\mathbb{C}$. Il est possible que le domaine D soit égal à l'ensemble des nombres complexes, on aurait ainsi $D = \mathbb{C}$.

11.6.1 Définition

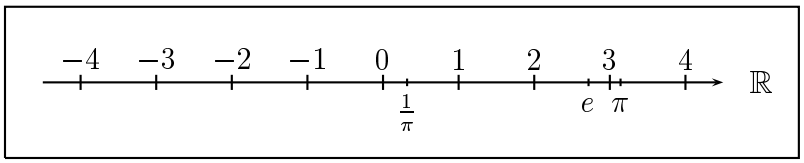
Une *fonction complexe* f est une fonction qui associe à un nombre complexe z (faisant partie d'un domaine D) un nombre complexe $f(z)$ (dépendant généralement de z).

Notation mathématique.

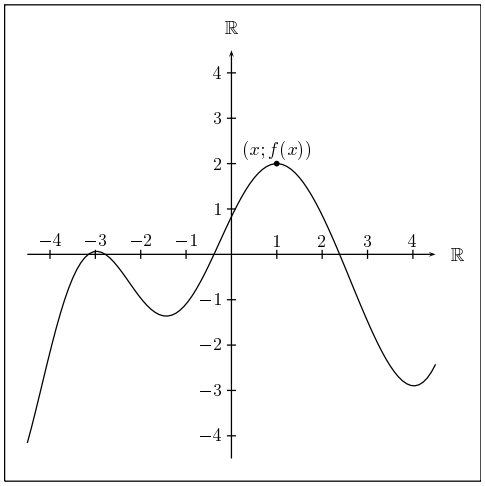
$$\begin{array}{l} f : D \to \mathbb{C} \\ \quad z \mapsto f(z) \end{array} \qquad \text{ou} \qquad f : D \to \mathbb{C}; z \mapsto f(z)$$

11.6.2 Représentation graphique

Contrairement aux fonctions réelles $f : D \to \mathbb{R}$ (où D est un domaine de $\mathbb{R}$), on ne peut pas dessiner de graphes pour les fonctions complexes.
Rappelons ce qui se passe pour les fonctions réelles. L'ensemble des nombres réels est représenté par une droite, appelée la droite réelle, qui est un objet mathématique de dimension 1.

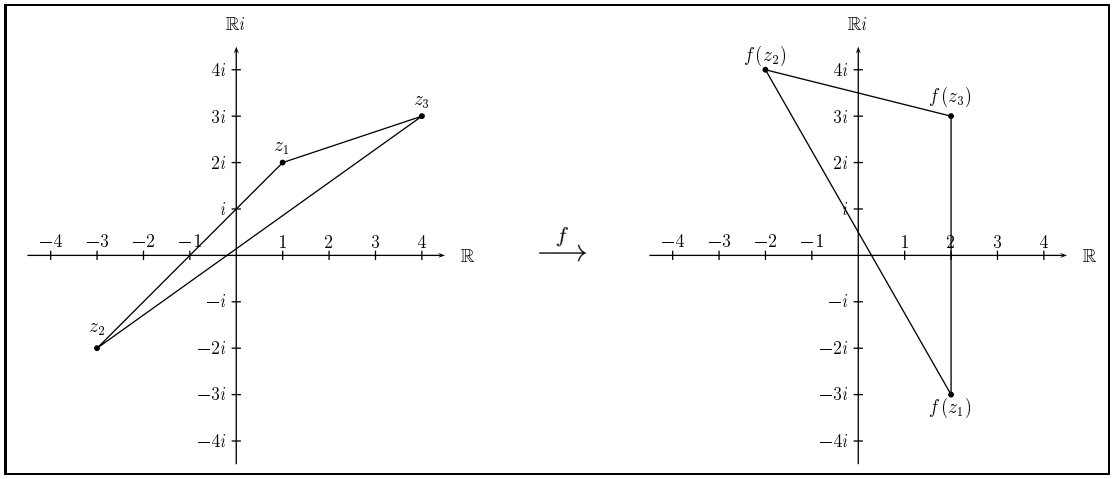


On représente le graphe d'une fonction réelle dans le plan, qui est un objet mathématique de dimension 2, de la manière suivante.



Ainsi, comme le plan de Gauss est un objet mathématique de dimension 2, si l'on tenait à représenter les fonctions complexes de la même manière, il faudrait travailler avec un objet mathématique de dimension 4 ! Ceci n'étant pas possible, on préfère faire autrement.

Pour décrire graphiquement une fonction complexe, on dessine deux plans complexes, l'un représentant le domaine D de départ et l'autre représentant l'ensemble d'arrivée qui est le plan de Gauss.



On peut aussi n'utiliser qu'un seul plan de Gauss, à condition de mettre des couleurs afin de distinguer les ensembles de départ et d'arrivée de la fonction.

11.6.3 Isométries, similitudes et similitudes rétrogrades

Transformations simples

Grâce aux représentations graphiques vues ci-dessus, on peut voir une fonction complexe comme une transformation du plan de Gauss. Les transformations les plus simples sont

1. Les translations.
2. Les homothéties.
3. Les rotations.
4. Les symétries.

Définitions et résultats

Définitions

1. La composition d'un nombre fini de translations et de rotations est une *isométrie*.
2. La composition d'un nombre fini d'isométries ou d'homothéties est appelée *similitude directe* ou *similitude*.

   Deux figures, images l'une de l'autre par similitude, sont dites *semblables*.
3. La composition d'une similitude directe et d'une symétrie axiale est appelée *similitude rétrograde*.

Premiers résultats

1. Toute isométrie s'écrit $f(z) = az + b$ avec $a, b \in \mathbb{C}$ tels que $|a| = 1$.
2. Toute similitude s'écrit $f(z) = az + b$ avec $a, b \in \mathbb{C}$.
3. Toute similitude rétrograde s'écrit $f(z) = a\overline{z} + b$ avec $a, b \in \mathbb{C}$.

Deuxièmes résultats

1. Une similitude directe ou rétrograde envoie une droite sur une droite et un cercle sur un cercle.
2. Une similitude directe conserve les angles (l'image par une similitude directe d'un angle droit est donc un angle droit) et envoie un triangle sur un triangle semblable (de même orientation).
3. Une similitude rétrograde inverse les angles (l'image par une similitude rétrograde d'un angle de $\frac{\pi}{6}$ est un angle de $-\frac{\pi}{6}$) et envoie un triangle sur un triangle presque semblable (dont l'orientation est inversée).

11.6.4 Points fixes

Définition

Soit f une fonction complexe. On dit que z_0 est un *point fixe de* f si $f(z_0) = z_0$.

11.6.5 Deux exercices avec leur corrigé

✠ Exercice 1

On considère la similitude rétrograde $f : z \mapsto f(z) = -2\,i\,\overline{z} + 1 + i$.
Décrire f, puis décrire l'image par f des sous-ensembles de $\mathbb{C}$ caractérisés par les conditions suivantes.

a) $|z| = 1$ b) $\frac{1}{2} < \mathrm{Im}(z) < 2$ c) $\mathrm{Re}(z)^2 = \mathrm{Im}(z)^2$

✠ Exercice 2

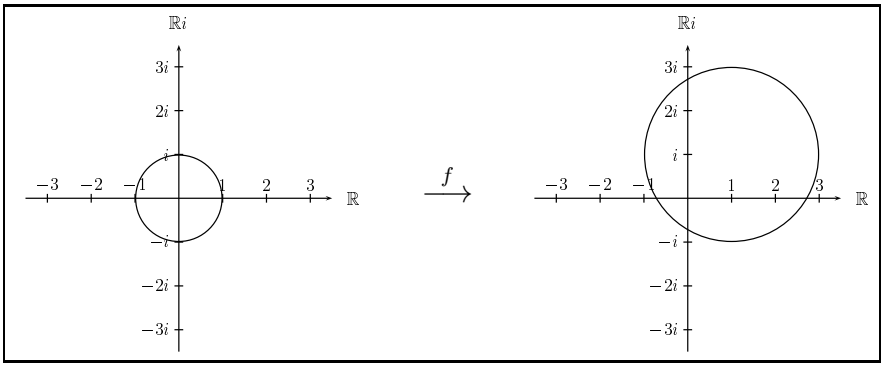
Soit $f : D \to \mathbb{C}; z \mapsto f(z) = \dfrac{z-1}{z+1}$.

1. Déterminer le plus grand domaine de définition D possible.
2. Quelle est l'image de l'axe réel ?
3. Quelle est l'image de l'axe imaginaire ?
4. Quel est l'ensemble des z tels que $f(z)$ soit purement imaginaire ?
5. Quels sont les points fixes de f ?
6. Calculer $f \circ f$.

Correction de l'exercice 1

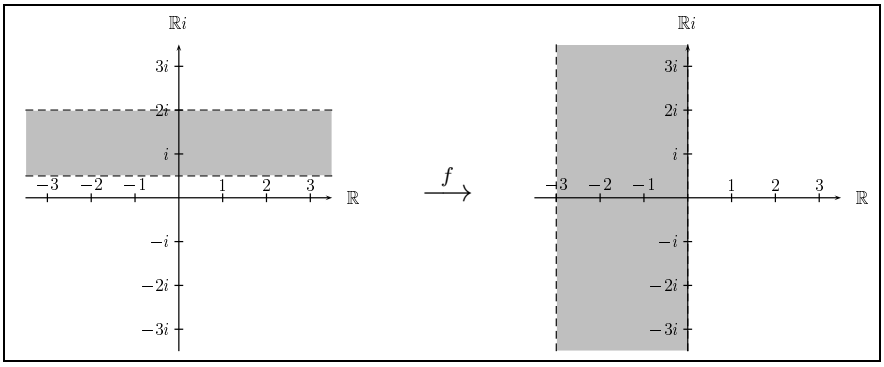
La similitude est une symétrie par rapport à l'axe réel, suivie d'une rotation de 90° autour de 0, puis d'une homothétie de facteur -2 et finalement d'une translation de $1+i$.

a) L'ensemble $\{z \in \mathbb{C} : |z| = 1\}$ représente le cercle de rayon 1 centré en 0.



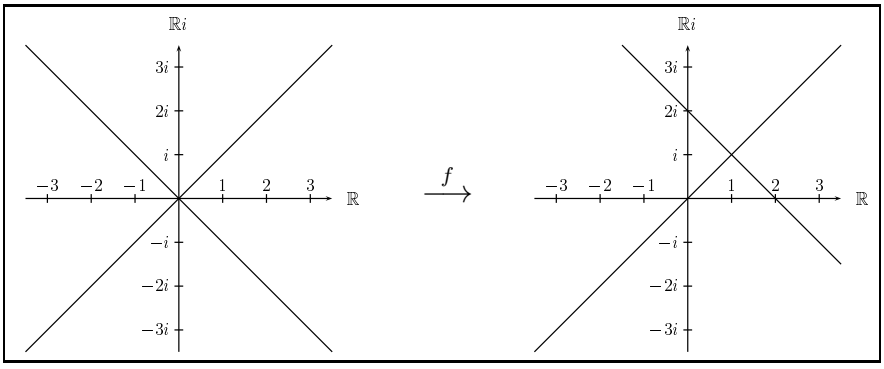
Son image par f est le cercle de rayon 2 centré en $1+i$, décrit par l'ensemble $\{z \in \mathbb{C} : |z-(1+i)| = 2\}$.

b) L'ensemble $\{z \in \mathbb{C} : \frac{1}{2} < \mathrm{Im}(z) < 2\}$ représente une bande horizontale passant entre $\frac{1}{2}i$ et $2i$.



Son image par f est une bande verticale passant entre -3 et 0, décrit par l'ensemble $\{z \in \mathbb{C} : -3 < \mathrm{Re}(z) < 0\}$.

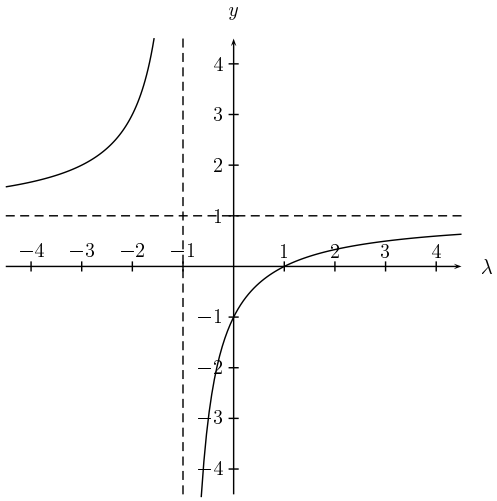
c) L'ensemble $\{z \in \mathbb{C} : \mathrm{Re}(z)^2 = \mathrm{Im}(z)^2\}$ représente les deux diagonales qui traversent le plan de Gauss.



Son image par f est encore deux diagonales, mais décalées de sorte que leur intersection se trouve au point $1+i$. L'ensemble correspondant est décrit comme ceci : $\{z \in \mathbb{C} : (\mathrm{Re}(z)-1)^2 = (\mathrm{Im}(z)-1)^2\}$.

Correction de l'exercice 2

1. Comme on ne peux pas diviser par 0, alors $D = \mathbb{C} \setminus \{-1\}$.
2. L'axe réel est décrit par l'ensemble $\{\lambda : \lambda \in \mathbb{R}\} \subset \mathbb{C}$. Retirons $\lambda = -1$ pour se retrouver dans le domaine de définition de f. Il faut donc examiner les valeurs possible de $f(\lambda)$ avec $\lambda \in \mathbb{R}\setminus\{-1\}$. Ces valeurs correspondent à l'image de l'application réelle $\tilde{f} : \mathbb{R} \setminus \{-1\} \to \mathbb{R}; \lambda \mapsto \frac{\lambda-1}{\lambda+1}$ dont le graphe est



   On voit que tous les nombres réels différents de 1 sont dans l'image. Ainsi l'image de l'axe réel par f est $\{z \in \mathbb{R} : z \neq 1\} \subset \mathbb{C}$.
3. Calculons $f(\lambda i)$ avec $\lambda \in \mathbb{R}$, on a

$$\begin{aligned} f(\lambda i) &= \frac{\lambda i - 1}{\lambda i + 1} = \frac{-(1-\lambda i)}{1+\lambda i} \cdot \frac{1-\lambda i}{1-\lambda i} = \frac{-(1-\lambda i)^2}{1+\lambda^2} = \frac{-(1-2\lambda i + \lambda^2 i^2)}{1+\lambda^2} \\ &= \frac{\lambda^2 + 2\lambda i - 1}{\lambda^2+1} = \frac{\lambda^2-1}{\lambda^2+1} + \frac{2\lambda}{\lambda^2+1} i \end{aligned}$$

   On remarque que $|f(\lambda i)| = 1$ quelque soit $\lambda \in \mathbb{R}$. En effet, on a

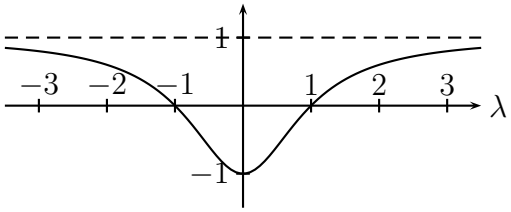
$$|f(\lambda i)| = \frac{(\lambda^2-1)^2 + (2\lambda)^2}{(\lambda^2+1)^2} = \frac{\lambda^4 - 2\lambda^2 + 1 + 4\lambda^2}{\lambda^4 + 2\lambda^2 + 1} = 1$$

   Ainsi, les nombres complexes $f(\lambda i)$ vivent dans le cercle de rayon 1 centré en 0. Mais cela ne permet pas de conclure que tous les points du cercle sont dans l'image.

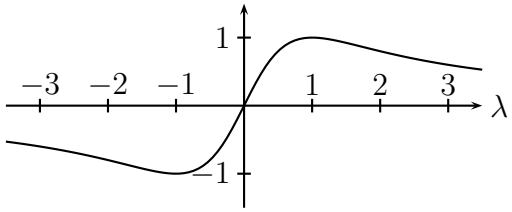
   Afin de déterminer quels points du cercle sont dans l'image de l'axe imaginaire, regardons les graphes des applications réelles.

$$g : \mathbb{R} \to \mathbb{R}; \lambda \mapsto \frac{\lambda^2-1}{\lambda^2+1} \qquad h : \mathbb{R} \to \mathbb{R}; \lambda \mapsto \frac{2\lambda}{\lambda^2+1}$$

$g(\lambda) = \mathrm{Re}(f(\lambda i))$



$h(\lambda) = \mathrm{Im}(f(\lambda i))$



On a les valeurs suivantes.

λ	$\to -\infty$		-1		0		1		$\to +\infty$
$f(\lambda i)$	$\to 1$		$-i$		-1		i		$\to 1$

Puisqu'on sait que $f(\lambda i)$ est sur le cercle trigonométrique et que les fonctions g et h sont continues, l'image par f de l'axe imaginaire est exactement le cercle trigonométrique complexe sans le point 1. Autrement dit :

$$f(\mathbb{R}i) = \{z \in \mathbb{C} : |z| = 1 \text{ et } z \neq 1\}$$

4. Commençons par calculer $f(x+iy)$, on a

$$\begin{aligned} f(x+iy) &= \frac{x+iy-1}{x+iy+1} = \frac{(x+iy-1)(x-iy+1)}{(x+1)^2+y^2} \\ &= \frac{x^2-ixy+x+ixy+y^2+iy-x+iy-1}{(x+1)^2+y^2} = \frac{x^2+y^2-1+2iy}{(x+1)^2+y^2} \\ &= \frac{x^2+y^2-1}{(x+1)^2+y^2} + \frac{2y}{(x+1)^2+y^2}i \end{aligned}$$

On cherche les nombres complexes $z = x+iy$ tels que la partie réelle de $f(z)$ soit nulle, en d'autres termes, on veut que

$$\frac{x^2+y^2-1}{(x+1)^2+y^2} = 0$$

On a donc l'équation $x^2+y^2 = 1$, cela signifie que les nombres complexes recherchés sont tous de module 1. Ainsi, l'ensemble des z tels que $f(z)$ soit purement imaginaire est le cercle de rayon 1 centré en 0 moins le point -1 qui n'est pas dans le domaine de définition. En termes ensemblistes : $\{z \in \mathbb{C} : |z| = 1 \text{ et } z \neq -1\}$.

5. On cherche à résoudre l'équation $f(z) = z$ que l'on peut écrire

$$\frac{z-1}{z+1} = z$$

En multipliant par $z+1$ de chaque côté de l'équation, on obtient

$$z - 1 = z^2 + z$$

En simplifiant par z, on a $z^2 = -1$. Les solutions de cette équation sont i et $-i$. Ce sont les points fixes de f.

6. À calculer : $(f \circ f)(z) = f(f(z))$.
On a

$$f(f(z)) = \frac{f(z)-1}{f(z)+1} = \frac{\frac{z-1}{z+1}-1}{\frac{z-1}{z+1}+1} = \frac{\frac{z-1-(z+1)}{z+1}}{\frac{z-1+(z+1)}{z+1}} = \frac{-2}{2z} = -\frac{1}{z}$$

Chapitre 12

L'ensemble de Mandelbrot et les ensembles de Julia

12.1 Préliminaires

12.1.1 Suites de nombres complexes

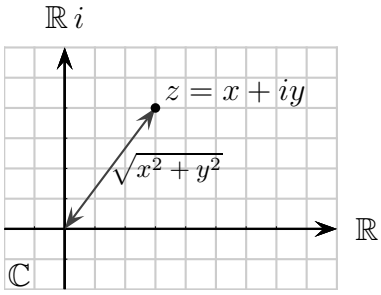
Une *suite de nombres complexes* est une liste ordonnée de nombres complexes, notée $z = (z_n)_{n\in\mathbb{N}} = (z_0, z_1, z_2, \ldots, z_n, \ldots)$.

12.1.2 Module et inégalité triangulaire

Soit $z = x+iy$ un nombre complexe. On définit le *module* de z par

$$|z| = \sqrt{x^2 + y^2}$$

Il s'agit de la distance du nombre z à l'origine dans le plan complexe. Ainsi $|z|$ est un nombre réel positif ou nul, qui n'est nul que pour $z = 0$. Cette définition prolonge celle de la valeur absolue.

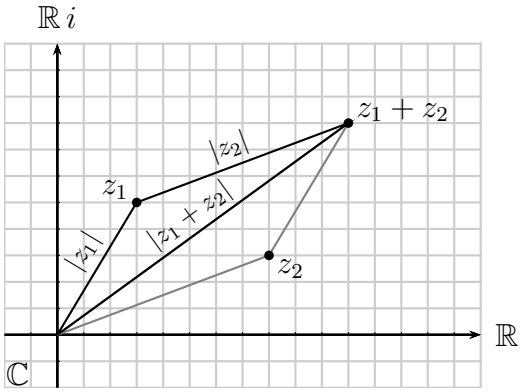


Inégalité triangulaire

La formule ci-dessous s'appelle l'*inégalité triangulaire*.

$$\boxed{|z_1 + z_2| \leqslant |z_1| + |z_2| \quad \text{pour tout} \quad z_1, z_2 \in \mathbb{C}}$$

Cela signifie qu'il est toujours plus court d'aller directement à $z_1 + z_2$ que de passer d'abord par z_1 (ou z_2). Cela revient au même lorsque les points sont alignés !



12.1.3 Inégalité triangulaire renversée (ITR)

On a $|z_1| = |z_1 + z_2 - z_2| \leqslant |z_1 + z_2| + |z_2|$ grâce à l'inégalité triangulaire. Ainsi

$$\boxed{|z_1 + z_2| \geqslant |z_1| - |z_2|}$$

Il s'agit de l'*inégalité triangulaire renversée*.

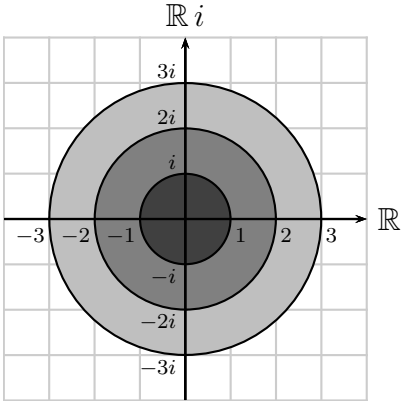
12.1.4 Boules centrées à l'origine

L'ensemble des nombres complexes dont la distance à l'origine est plus petite ou égale à un rayon donné r est noté $B_{\leqslant r}(0)$.

Notation ensembliste

$$B_{\leqslant r}(0) = \{z \in \mathbb{C} : |z| \leqslant r\}$$

Ci-contre, les boules $B_{\leqslant 1}(0)$, $B_{\leqslant 2}(0)$ et $B_{\leqslant 3}(0)$ sont représentées dans le plan complexe.



12.1.5 Suites bornées

Une suite de nombres complexes $(z_n)_{n\in\mathbb{N}}$ est *bornée* s'il existe un rayon r pour lequel tous les éléments de la suite sont dans $B_{\leqslant r}(0)$, c'est-à-dire $z_n \in B_{\leqslant r}(0)$ pour tout $n \in \mathbb{N}$.

Autrement dit, une suite n'est pas bornée si et seulement si

$$|z_n| \longrightarrow +\infty \quad \text{lorsque } n \to +\infty$$

C'est-à-dire, une suite est bornée si et seulement si elle ne s'éloigne pas irrémédiablement de l'origine (ou de tout autre point).

12.1.6 Notation pour les compositions de fonctions

Soit $f : \mathbb{C} \to \mathbb{C}$ une fonction complexe. On note $f^{(\circ n)}$ pour la fonction $f \circ f \circ f \circ \cdots \circ f$ (avec n fois la fonction f).

Autrement dit

$$f^{(\circ n)}(z) = \underbrace{(f \circ f \circ f \circ \cdots \circ f)}_{n \text{ fois la fonction } f}(z) = \underbrace{f(f(\cdots f(z)))}_{n \text{ fois la fonction } f}$$

Par défaut, on pose $f^{(\circ 0)}(z) = z$ («si on n'applique aucune fois la fonction f à z, on obtient z, c'est-à-dire que l'on ne fait rien»).

12.2 L'ensemble de Mandelbrot

Pour chaque $c \in \mathbb{C}$, on se donne la fonction

$$f_c : \mathbb{C} \to \mathbb{C};\ z \mapsto z^2 + c$$

Pour définir l'ensemble de Mandelbrot, on examine les suites de la forme

$$s_c = \left(f_c^{(\circ n)}(0)\right)_{n\in\mathbb{N}} = \big(0,\ f_c(0),\ f_c(f_c(0)),\ \ldots\ ,\ f_c^{(\circ n)}(0),\ \ldots\big)$$

Deux cas exclusifs se produisent.

1. La suite s_c est bornée.
2. La suite s_c n'est pas bornée.

L'ensemble de Mandelbrot, noté M, est l'ensemble des nombres complexes $c \in \mathbb{C}$ pour lesquels les suites s_c sont bornées. En termes mathématiques

$$M = \{c \in \mathbb{C} : s_c \text{ est bornée}\}$$

12.2.1 Une première propriété de l'ensemble de Mandelbrot

On va montrer que l'ensemble de Mandelbrot est contenu dans la boule de rayon 2 centrée à l'origine. Mais pour établir ce résultat, on a besoin d'un lemme. Ce lemme servira par la suite à établir un critère permettant de dessiner l'ensemble de Mandelbrot sur un ordinateur.

Lemme

Soit $c \in \mathbb{C}$.
Supposons qu'il existe un terme, appelé z, de la suite s_c qui satisfait

$$H_1 : |z| \geqslant |c| \quad \text{et} \quad H_2 : |z| > 2$$

Alors, la suite s_c n'est pas bornée.

Preuve

Posons $\alpha = |z| - 1$. Par H_2, on a $\alpha > 1$.

Montrons par récurrence que pour tout $n \in \mathbb{N}$, on a

$$\boxed{\left|f_c^{(\circ n)}(z)\right| \geqslant \alpha^n |z| \geqslant |z| \geqslant |c|}$$

Ancrage : c'est évident pour $n = 0$, car on a

$$|z| \geqslant |z| \geqslant |z| \overset{H_1}{\geqslant} |c|$$

Pas de récurrence : on montre si c'est vrai pour n, alors c'est vrai pour $n+1$.

$$\begin{aligned}
\left|f_c^{(\circ(n+1))}(z)\right| &= \left|f_c\left(f_c^{(\circ n)}(z)\right)\right| = \left|\left(f_c^{(\circ n)}(z)\right)^2 + c\right| \\
&\overset{\text{ITR}}{\geqslant} \left|\left(f_c^{(\circ n)}(z)\right)^2\right| - |c| \\
&\overset{HR}{\geqslant} \left|\left(f_c^{(\circ n)}(z)\right)^2\right| - \left|f_c^{(\circ n)}(z)\right| = \left|f_c^{(\circ n)}(z)\right|^2 - \left|f_c^{(\circ n)}(z)\right| \\
&= \left(\left|f_c^{(\circ n)}(z)\right| - 1\right)\left|f_c^{(\circ n)}(z)\right| \\
&\overset{HR}{\geqslant} \left(|z| - 1\right)\left|f_c^{(\circ n)}(z)\right| = \alpha\left|f_c^{(\circ n)}(z)\right| \\
&\overset{HR}{\geqslant} \alpha \cdot \alpha^n |z| = \alpha^{n+1}|z| \\
&\overset{\alpha>1}{\geqslant} |z| \overset{H_1}{\geqslant} |c|
\end{aligned}$$

Donc

$$|f_c^{(\circ n)}(z)| \geqslant \alpha^n |z| \longrightarrow +\infty \text{ lorsque } n \to +\infty$$

Par conséquent, $|f_c^{\circ n}(z)| \longrightarrow +\infty$ lorsque $n \to +\infty$.
Cela montre que la suite $s_c = (0, f_c(0), \ldots, z, \ldots, f_c^{(\circ n)}(z), \ldots)$ n'est pas bornée.

□

Propriété 1

L'ensemble de Mandelbrot est contenu dans la boule de rayon 2 centrée à l'origine.
En d'autres termes :

$$M \subset B_{\leqslant 2}(0)$$

Preuve

Par contraposée, on suppose que $c \notin B_{\leqslant 2}(0)$ et on montre que $c \notin M$. Autrement dit, on montre que chaque suite s_c, avec $c \in \mathbb{C}$ tel que $|c| > 2$, est non bornée.

Soit donc $c \in \mathbb{C}$ tel $|c| > 2$. On a

$$s_c = (0,\ c,\ f_c(c),\ \ldots)$$

Le deuxième terme, c, satisfait les hypothèses du lemme (H_1 : $|c| \geqslant |c|$ est banal et H_2 : $|c| > 2$ est l'hypothèse de départ de cette propriété).

Ainsi, par le lemme, la suite s_c n'est pas bornée ! □

12.2.2 En route vers les représentations graphiques

Afin de savoir si, pour chaque nombre complexe $c \in \mathbb{C}$, la suite s_c est bornée ou non, on va établir un critère permettant de savoir si une suite va être bornée ou non.

La première propriété nous dit qu'aucune suite s_c avec $|c| > 2$ ne sera bornée. Cela ne suffit pas pour trouver un bon critère, mais cela peut nous en donner une idée.

Critère pour que s_c ne soit pas bornée

Soit $c \in \mathbb{C}$. On calcule chaque terme de la suite s_c un à un. Si à un moment donné, on obtient un élément de la suite s_c qui est à distance plus grande que 2 de l'origine, alors la suite s_c ne sera pas bornée.

Preuve

Deux cas se présentent.

1. $|c| > 2$.

   Dans ce cas, le deuxième terme de la suite s_c est c qui satisfait $|c| > 2$. On applique le lemme pour $z = c$ afin de montrer que s_c n'est pas bornée.

2. $|c| \leqslant 2$.

   Notons z le premier élément de la suite s_c qui satisfait $|z| > 2$. Dans ce cas, on a

   $$|z| > 2 \geqslant |c|$$

   Ainsi, z satisfait les hypothèses du lemme et, par conséquent, la suite s_c n'est pas bornée.

   □

Remarque évidente

La réciproque est vraie puisque si la suite s_c n'est pas bornée, alors il existera un élément de la suite qui est à distance plus grande que 2 de l'origine.

12.2.3 Algorithme en Python

Voici le programme `Python` qui permet de créer les images se trouvant dans la suite de ce document (à quelques détails près pour la palette de couleur). Il commence par l'importation des modules `Image` et `ImageDraw` qui nécessitent la libraire `PIL`.

```
import Image, ImageDraw

def mandelbrot(c,lim) :
    z = 0
    compteur = 0
    while ( abs(z) <= 2.0 and compteur < lim ) :
        z = z**2 + c
        compteur = compteur + 1
    return compteur

def dessineMandelbrot(largeur,hauteur,lim) :
    im = Image.new("RGB", (largeur, hauteur))
    draw = ImageDraw.Draw(im)

    deltaR  = complex((zB.real - zA.real)/largeur, 0)
    deltaRi = complex(0, (zB.imag - zA.imag)/hauteur)
    for y in range (0, hauteur):
        for x in range (0, largeur):
            nb = int(255 - 2.55*mandelbrot(zA + x*deltaR + y*deltaRi, lim))
            draw.point((x, hauteur-1-y), (nb,nb,nb))

    im.save("mandelbrot.png", "PNG")

###  Programme principal  ###
zA = complex(-2.0,-2.0)
zB = complex( 2.0, 2.0)
largeur = 500
hauteur = 500
lim     = 100
dessineMandelbrot(largeur,hauteur,lim)
```

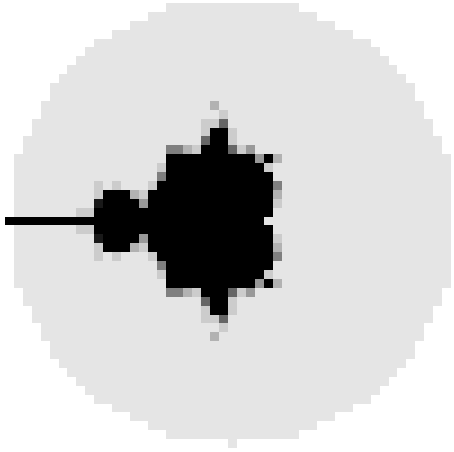
Dans le programme principal, on trouve la définition du *cadre* donné par le coin inférieur gauche `zA` et le coin supérieur droit `zB`. La résolution de l'image est donnée par `hauteur` et `largeur`.

La fonction `mandelbrot` calcule (à l'aide du `compteur`) le nombre de termes de la suite s_c qui sont à distance plus petite ou égale à 2 de l'origine. Bien sûr, si la suite est bornée, le critère indique que la suite reste dans la boule $B_{\leqslant 2}(0)$. Dans ce cas, le `compteur` ira à l'infini. Or, un ordinateur ne pouvant pas effectuer une infinité de calculs, il faut fixer une limite `lim` à partir de laquelle on ne calcule pas le terme suivant de la suite s_c.

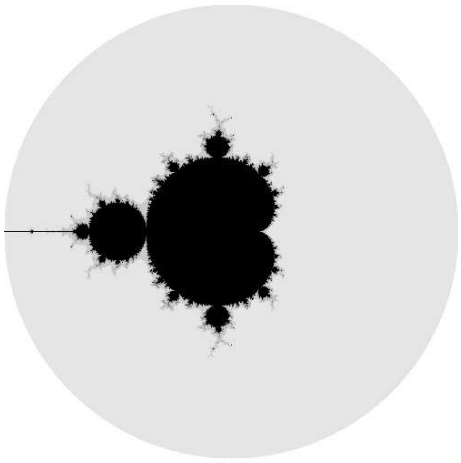
Puis la commande `dessineMandelbrot` va découper le *cadre* en morceaux. Dans chaque zone, un nombre complexe c sera choisi et le `compteur` sera calculé pour ce nombre. La zone correspondante sera ainsi coloriée selon la valeur du `compteur`. Plus vite la suite quitte la boule $B_{\leqslant 2}(0)$, plus la couleur sera claire. Un carré noir ne signifie pas que la suite correspondante au nombre c choisi dans ce carré sera bornée, cela signifie qu'avant la limite fixée (ici : `lim` $= 100$), la suite sera toujours dans la boule $B_{\leqslant 2}(0)$.

12.2.4 Représentations graphiques de l'ensemble de Mandelbrot

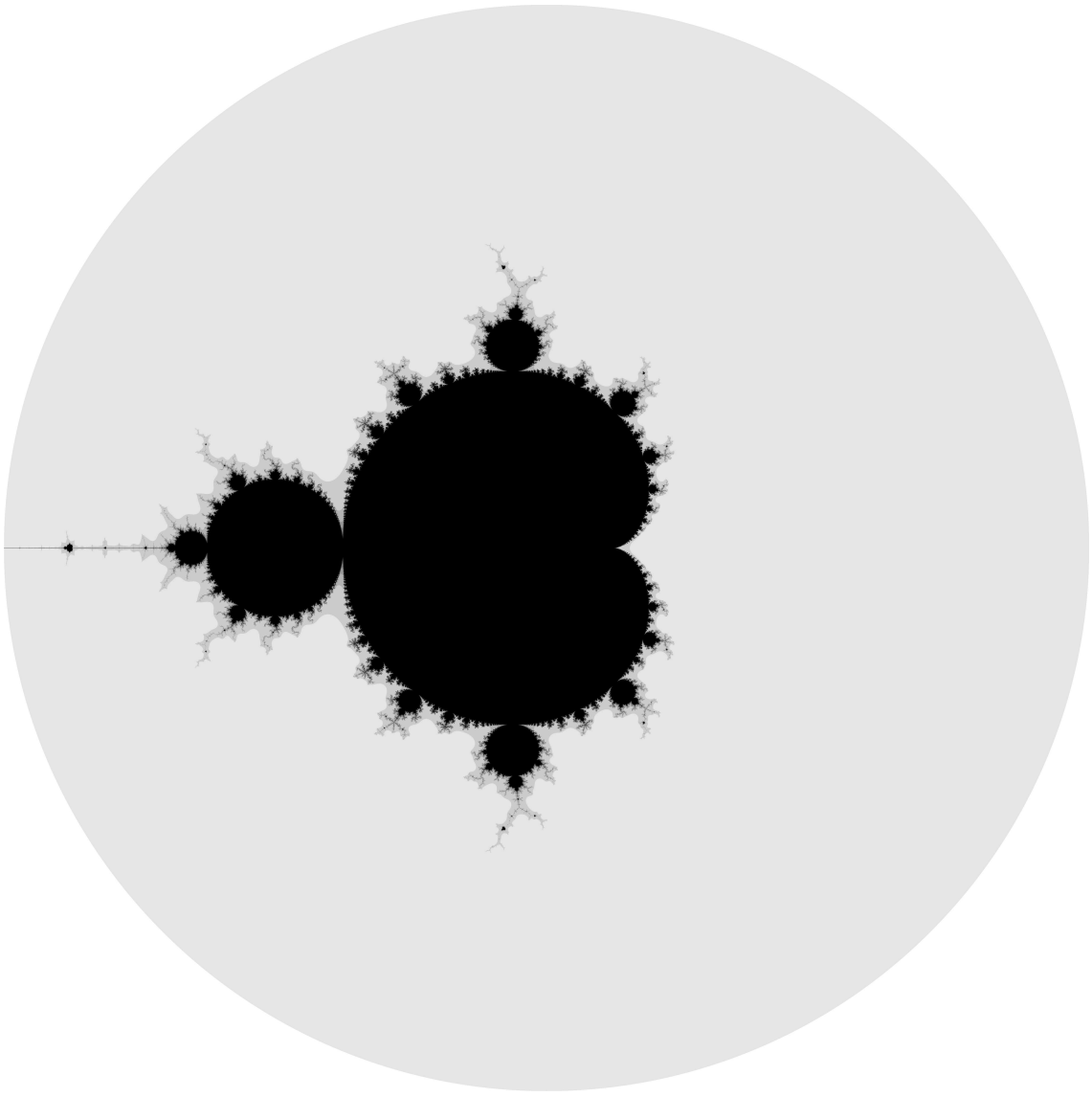
Voici plusieurs représentations effectuées à l'aide du programme précédent. En gris clair, on voit la boule $B_{\leqslant 2}(0)$.



avec une résolution de 50 sur 50



avec une résolution de 500 sur 500



avec une résolution de 5000 sur 5000

12.2.5 Une autre propriété de l'ensemble de Mandelbrot

Propriété 2 (sans preuve)

Les nombres réels contenus dans l'ensemble de Mandelbrot sont exactement les nombres de -2 à $1/4$. Autrement dit :

$$M \cap \mathbb{R} = \left[-2, \tfrac{1}{4}\right]$$

Cette propriété répond aux incertitudes soulevées dans le premier exercice.

Remarque

Les représentations graphiques sont des approximations de l'ensemble de Mandelbrot. Selon le choix du *cadre*, l'ensemble $[-2, \frac{1}{4}]$ pourrait ne pas apparaître pas comme il le devrait. C'est le cas, par exemple, si on prend $z_A = -2 - 1.9i$ et $z_B = 2 + 2i$ avec une résolution de 500.

12.3 Les ensembles de (Gaston) Julia

Ces ensembles sont construits d'une manière similaire à celui de Mandelbrot.

Pour chaque $c \in \mathbb{C}$, on se donne la fonction

$$f_c : \mathbb{C} \to \mathbb{C};\ z \mapsto z^2 + c$$

Il y a un ensemble de Julia pour chaque nombre complexe $c \in \mathbb{C}$. Pour définir l'ensemble de Julia correspondant au nombre c, on examine, pour chaque $z \in \mathbb{C}$, les suites de la forme

$$s_c(z) = \left(f_c^{(\circ n)}(z)\right)_{n\in\mathbb{N}} = \left(z,\ f_c(z),\ f_c(f_c(z)),\ \ldots\ ,\ f_c^{(\circ n)}(z),\ \ldots\right)$$

Deux cas exclusifs se produisent.

1. La suite $s_c(z)$ est bornée.
2. La suite $s_c(z)$ n'est pas bornée.

L'ensemble de Julia associé au nombre complexe c, noté J_c, est l'ensemble des nombres complexes $z \in \mathbb{C}$ pour lesquels les suites $s_c(z)$ sont bornées. En termes mathématiques

$$J_c = \{z \in \mathbb{C} : s_c(z) \text{ est bornée}\}$$

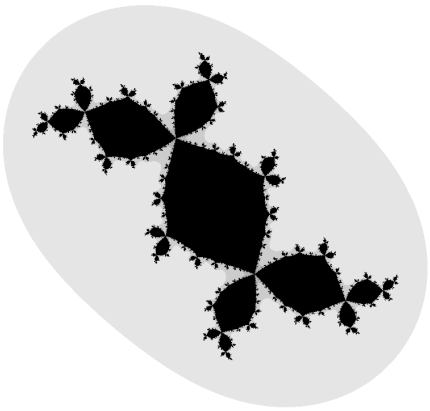
Pour représenter ces ensembles, on utilise le même critère que pour l'ensemble de Mandelbrot (il faut tout de même adapter les preuves, mais ceci est une autre histoire...).

Voici l'algorithme permettant de calculer le `compteur` pour les ensembles de Julia sous `Python`.

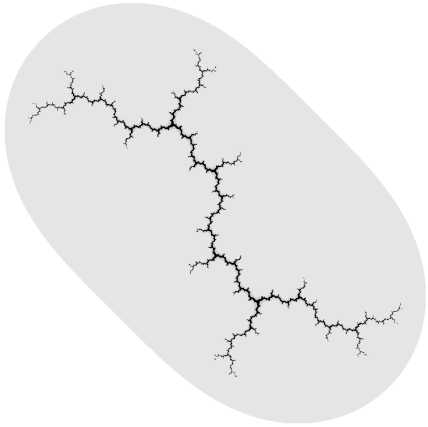
```
def julia(z0,c,lim) :
    z = z0
    compteur = 0
    while ( abs(z) <= 2.0 and compteur < lim ) :
        z = z**2 + c
        compteur = compteur + 1
    return compteur
```

La programmation de la fonction `dessineJulia` est laissée au lecteur. Il s'agit simplement de réadapter la fonction `dessineMandelbrot` donnée précédemment.

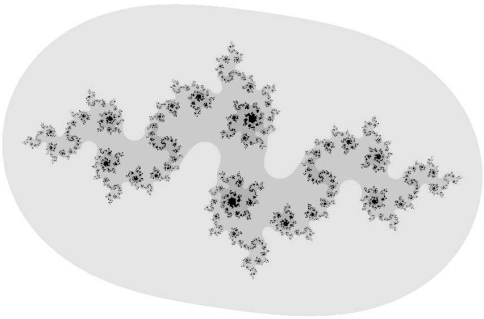
12.3.1 Représentations graphiques d'ensembles de Julia



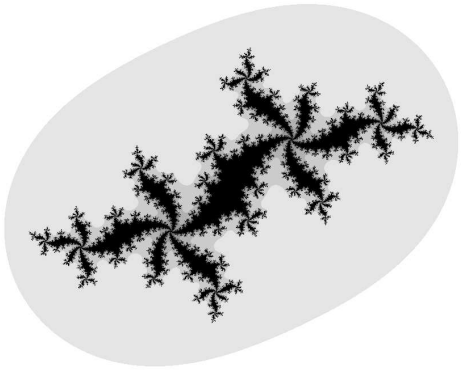
$c = -0.123 + 0.745i$



$c = i$



$c = -0.85 + 0.25i$



$c = -0.5 - 0.6i$

Le lecteur désirant visualiser ces ensembles (ou de manière quasi instantanée, ce qui est pratique pour faire des zooms successifs) peut se rendre sur le site suivant où se trouve une sympathique applet Java.

`http://aleph0.clarku.edu/~djoyce/julia/explorer.html`

L'auteur y définit l'ensemble de Mandelbrot et les ensembles de Julia à l'aide de la fonction $f_c(z) = z^2 - c$, cela a pour effet de produire une symétrie sur notre ensemble de Mandelbrot, mais cela ne le change pas outre mesure.

Il y a aussi le chapitre 6 ($7'15''$) du merveilleux documentaire qui se trouve sur le site

`https://www.dimensions-math.org/`

Chapitre 13

Séries et développements de Taylor

13.1 Les séries arithmétiques et géométriques

Il s'agit d'un résumé du chapitre 3 du cours DF (`http://www.vive-les-maths.net`).

13.1.1 Le symbole somme

Le symbole $\sum$ permet de condenser l'écriture de grandes sommes.

$$\sum_{k=1}^{n} a_k = a_1 + a_2 + a_3 + \cdots + a_n = \sum_{k=0}^{n-1} a_{k+1}$$

13.1.2 Séries arithmétiques

Définition

Une *série arithmétique* est une somme finie de termes $a_1 \overset{+r}{+} a_2 \overset{+r}{+} a_3 \overset{+r}{+} \cdots \overset{+r}{+} a_n$ pour laquelle il existe un nombre r, $r \neq 0$, appelé *raison*, qui satisfait $\boxed{a_{k+1} = a_k + r}$.

Théorème

Si $a_1 + a_2 + \cdots + a_{n-1} + a_n$ est une série arithmétique de raison r. Alors, on a :

$$a_1 \overset{+r}{+} a_2 \overset{+r}{+} a_3 \overset{+r}{+} \cdots \overset{+r}{+} a_n = \boxed{\sum_{k=0}^{n-1} \Big(a_1 + r\,k\Big) = n \cdot \frac{a_1 + a_n}{2}}$$

13.1.3 Séries géométriques

Définition

Une *série géométrique* est une somme finie de termes $a_1 \overset{\cdot r}{+} a_2 \overset{\cdot r}{+} a_3 \overset{\cdot r}{+} \cdots \overset{\cdot r}{+} a_n$ pour laquelle il existe un nombre r, $r \neq 1$, appelé *raison*, qui satisfait $\boxed{a_{k+1} = a_k \cdot r}$.

Théorème

Si $a_1 + a_2 + \cdots + a_{n-1} + a_n$ est une série géométrique de raison r. Alors, on a :

$$a_1 \overset{\cdot r}{+} a_2 \overset{\cdot r}{+} \cdots \overset{\cdot r}{+} a_n = \boxed{\sum_{k=0}^{n-1} \Big(a_1 \cdot r^k\Big) = a_1 \cdot \sum_{k=0}^{n-1} r^k = a_1 \frac{1 - r^n}{1 - r}}$$

13.2 Une propriété fondamentale des nombres réels

Soit $(a_k)_{k\geqslant k_0}$ une suite[1] de nombres réels. S'il existe un rang $K \geqslant k_0$ tel que

1. $(a_k)_{k\geqslant K}$ est *monotone croissante* ($a_{k+1} \geqslant a_k$ pour tout $k \geqslant K$) ;
2. $(a_k)_{k\geqslant K}$ est *majorée* (il existe $M \in \mathbb{R}$ tel que $a_k \leqslant M$ pour tout $k \geqslant K$).

Alors, la suite $(a_k)_{k\geqslant k_0}$ est convergente dans $\mathbb{R}$ (il existe $a \in \mathbb{R}$ tel que $\lim\limits_{k\to+\infty} a_k = a$).

13.3 Séries infinies et critères de convergence

13.3.1 Séries infinies

Soit $(a_k)_{k\geqslant k_0}$ une suite de nombres réels. On définit une série infinie comme étant la limite de ses *sommes partielles* (si cette limite existe). On dit que a_k est le *terme général de cette série.*

$$\underbrace{\sum_{k\geqslant k_0} a_k = \sum_{k=k_0}^{+\infty} a_k}_{\text{ce sont deux façons de noter la même somme}} \overset{\text{définition}}{=} \lim_{n\to+\infty} \underbrace{\sum_{k=k_0}^{n} a_k}_{\text{somme partielle}}$$

13.3.2 La série géométrique infinie et sa convergence

Soit $z \in \mathbb{C}$ ($z = 1$ compris). La série géométrique infinie de raison z est

$$\sum_{k\geqslant 0} z^k = 1 + z + z^2 + z^3 + \cdots$$

Théorème

1. Si $|z| < 1$, alors la série converge dans $\mathbb{C}$ et vaut $\boxed{\sum\limits_{k\geqslant 0} z^k = \dfrac{1}{1-z}}$.
2. Si $|z| \geqslant 1$, alors la série ne converge pas dans $\mathbb{C}$.

Preuve

1. Si $|z| \geqslant 1$, on a $\lim_{k\to+\infty} z^k \neq 0$ et par la contraposée du théorème fondamental de la page suivante, la série ne converge pas (dans $\mathbb{C}$).
2. Si $|z| < 1$, les sommes partielles sont des séries géométriques finies

$$\sum_{k\geqslant 0} z^k = \lim_{n\to+\infty} \sum_{k=0}^{n} z^k \overset{\text{série géométrique}}{=} \lim_{n\to+\infty} \frac{1-z^{n+1}}{1-z}$$

Comme $\lim\limits_{n\to+\infty} z^n = 0$, on a $\sum\limits_{k\geqslant 0} z^k = \dfrac{1}{1-z}$. □

Un cas particulier

Voici une série géométrique infine (cas $z = \frac{1}{2}$).

$$\sum_{k=0}^{+\infty} \frac{1}{2^k} = \lim_{n\to+\infty} \sum_{k=0}^{n} \frac{1}{2^k} = \lim_{n\to+\infty} \frac{1-(\frac{1}{2})^{n+1}}{1-\frac{1}{2}} = \frac{1}{1-\frac{1}{2}} = 2$$

1. Dans ce chapitre, toutes les suites sont indicées par un sous-ensemble des nombres naturels ($k_0 \in \mathbb{N}$).

13.3.3 La série harmonique

Voici une série infinie qui diverge. C'est la *série harmonique*

$$\boxed{\sum_{k=1}^{+\infty} \frac{1}{k} = +\infty}$$

En effet, supposons par l'absurde que la série harmonique converge vers un nombre $s \in \mathbb{R}$. Notons s_n la somme partielle $\sum_{k=1}^{n} \frac{1}{k}$. Ainsi, la suite $(s_n)_{n\geqslant 1}$ converge vers s, tout comme la sous-suite $(s_{2n})_{n\geqslant 1}$. Autrement dit

$$\lim_{n\to+\infty} s_n = \lim_{n\to+\infty} s_{2n} = s$$

Considérons la suite $(s_{2n} - s_n)_{n\geqslant 1}$. Par les propriétés des limites, on a

$$\lim_{n\to+\infty} (s_{2n} - s_n) = \lim_{n\to+\infty} s_{2n} - \lim_{n\to+\infty} s_n = s - s = 0 \ (\bigstar)$$

Mais on a, pour $n \geqslant 1$,

$$\begin{aligned} s_{2n} - s_n &= 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{2n} - \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right) = \sum_{k=1}^{2n} \frac{1}{k} - \sum_{k=1}^{n} \frac{1}{k} \\ &= \underbrace{\frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n-1} + \frac{1}{2n}}_{n \text{ fractions}} = \sum_{k=n+1}^{2n} \frac{1}{k} \end{aligned}$$

Or si k est un nombre tel que $k \leqslant 2n$, alors $\frac{1}{k} \geqslant \frac{1}{2n}$, donc

$$s_{2n} - s_n = \sum_{k=n+1}^{2n} \frac{1}{k} \geqslant \sum_{k=n+1}^{2n} \frac{1}{2n} = \frac{n}{2n} = \frac{1}{2}$$

On vient de voir que chaque terme de la suite $(s_{2n} - s_n)_{n\geqslant 1}$ est plus grand ou égal à $\frac{1}{2}$. Cette suite ne peut donc pas converger vers 0, ce qui contredit $(\bigstar)$. □

13.3.4 Théorème fondamental sur les convergences de séries

Soit $\sum_{k\geqslant k_0} a_k$ une série avec $a_k \in \mathbb{C}$. Si cette série converge dans $\mathbb{C}$, alors son terme général tend vers zéro. Autrement dit

$$\sum_{k\geqslant k_0} a_k = s \implies \lim_{k\to+\infty} a_k = 0$$

Preuve

Notons s_n la somme partielle $\sum_{k=k_0}^{n} a_k$. Par les propriétés des limites, on a

$$\lim_{n\to+\infty} a_n = \lim_{n\to+\infty} \left(s_n - s_{n-1}\right) = \lim_{n\to+\infty} s_n - \lim_{n\to+\infty} s_{n-1} = s - s = 0$$

□

La réciproque du théorème fondamental est fausse

En effet, la série harmonique fournit un contre-exemple.

13.3.5 Critère de comparaison

Soit $(a_k)_{k\geqslant k_0}$ et $(b_k)_{k\geqslant k_0}$ deux suites de nombres réels tels que, à partir d'un rang $K \geqslant k_0$, on a $0 \leqslant a_k \leqslant b_k$ pour chaque $k \geqslant K$, alors :

1. Si $\sum\limits_{k\geqslant k_0} b_k$ converge dans $\mathbb{R}$, alors $\sum\limits_{k\geqslant k_0} a_k$ converge dans $\mathbb{R}$.
2. Si $\sum\limits_{k\geqslant k_0} a_k$ diverge, alors $\sum\limits_{k\geqslant k_0} b_k$ diverge.

Preuve

1. Puisque $\sum\limits_{k\geqslant k_0} b_k$ converge, il existe $t \in \mathbb{R}$ tel que $\sum\limits_{k\geqslant K} b_k = t$.

   Notons s_n la somme partielle $\sum\limits_{k=K}^{n} a_k$ et t_n la somme partielle $\sum\limits_{k=K}^{n} b_k$.

   Comme $a_k \geqslant 0$ et $b_k \geqslant 0$ pour tout $k \geqslant K$, il est évident que les suites des sommes partielles $(s_n)_{n\geqslant K}$ et $(t_n)_{n\geqslant K}$ sont monotones croissantes.

   Comme pour tout $k \geqslant K$ on a $a_k \leqslant b_k$, on sait que $s_n \leqslant t_n \leqslant t$ pour tout $n \geqslant K$ (car $(t_n)_{n\geqslant K}$ est monotone croissante).

   Ainsi, $(s_n)_{n\geqslant K}$ est une suite monotone croissante majorée. Par la propriété fondamentale des nombres réels vue en page 128, on conclut que la suite des sommes partielles $(s_n)_{n\geqslant K}$ est convergente et qu'ainsi $\sum\limits_{k\geqslant k_0} a_k$ converge.

2. Il s'agit de la contraposée du point 1. □

Application du critère de comparaison

On va comparer la série $\sum\limits_{k\geqslant 1} \dfrac{1}{k^2}$ à la série convergente suivante.

$$\sum_{k\geqslant 1} \frac{2}{(k+1)(k+2)} = 1$$

On démontre la convergence de cette dernière série grâce à une somme télescopique.

$$\lim_{n\to+\infty} \sum_{k=1}^{n} \frac{2}{(k+1)(k+2)} = \lim_{n\to+\infty} \sum_{k=1}^{n} \left(\frac{2}{k+1} - \frac{2}{k+2}\right) = \lim_{n\to+\infty} \left(1 - \frac{2}{n+2}\right) = 1$$

De plus, on a, pour $k \geqslant 4$

$$\boxed{0 \leqslant \frac{1}{k^2} < \frac{2}{(k+1)(k+2)}} \bigstar$$

En effet, on a

$$f(k) = \frac{2}{(k+1)(k+2)} - \frac{1}{k^2} = \frac{2k^2 - (k+1)(k+2)}{k^2(k+1)(k+2)} = \frac{k^2 - 3k - 2}{k^2(k+1)(k+2)}$$

Le seul zéro positif est $k = \frac{3+\sqrt{17}}{2} \cong 3.56$, ainsi, après $k = 4$, la fonction ne changera plus de signes. Comme $f(4) > 0$, on a bien $f(k) > 0$ pour $k \geqslant 4$.

Grâce au critère de comparaison, $\bigstar$ montre que $\sum\limits_{k\geqslant 1} \frac{1}{k^2}$ converge puisque $\sum\limits_{k\geqslant 1} \frac{2}{(k+1)(k+2)}$ converge (sa limite est 1).

Remarque. Ce n'est pas facile, mais on peut montrer que $\sum\limits_{k\geqslant 1} \dfrac{1}{k^2} = \dfrac{\pi^2}{6}$.

13.3.6 Critère de la racine (ou de Cauchy)

On considère la série infinie $\sum_{k\geqslant k_0} a_k$ avec $a_k \geqslant 0$ pour tout $k \geqslant k_0$.

Si $\lim_{k\to+\infty} \sqrt[k]{a_k} = c$ (si cette limite existe) et $\begin{cases} c<1, \text{ alors la série converge dans } \mathbb{R} \\ c=1, \text{ alors il y a doute} \\ c>1, \text{ alors la série diverge} \end{cases}$

Preuve

On suppose que la limite existe et vaut c. On distingue trois cas :

1. $c < 1$.

   Soit $r \in]c, 1[$. Puisque $\lim_{k\to+\infty} \sqrt[k]{a_k} = c < 1$, à partir d'un certain rang $K \geqslant k_0$, on a

   $$\sqrt[k]{a_k} \leqslant r < 1 \quad \text{pour tout} \quad k \geqslant K$$

   De plus, comme $a_k \geqslant 0$ pour tout $k \geqslant k_0$, on sait que $c \geqslant 0$. Donc $r \in [0, 1[$.

   Ainsi, on a

   $$\begin{aligned} \sum_{k\geqslant K} a_k &= a_K + a_{K+1} + a_{K+2} + a_{K+3} + \cdots \\ &= \left(\sqrt[K]{a_K}\right)^K + \left(\sqrt[K+1]{a_{K+1}}\right)^{K+1} + \left(\sqrt[K+2]{a_{K+2}}\right)^{K+2} + \left(\sqrt[K+3]{a_{K+3}}\right)^{K+3} + \cdots \\ &\leqslant r^K + r^{K+1} + r^{K+2} + r^{K+3} + \cdots \\ &= r^K \cdot (1 + r + r^2 + r^3 + \cdots) \overset{r\in[0,1[}{=} r^K \cdot \frac{1}{1-r} = \frac{r^K}{1-r} \end{aligned}$$

   On a ainsi montré que la suite des sommes partielles est majorée.

   $$\sum_{k=k_0}^{n} a_k \overset{a_k\geqslant 0}{\leqslant} \sum_{k=k_0}^{+\infty} a_k = \sum_{k=k_0}^{K-1} a_k + \sum_{k=K}^{+\infty} a_k \leqslant \sum_{k=k_0}^{K-1} a_k + \frac{r^K}{1-r}$$

   Donc $\sum_{k\geqslant k_0} a_k$ converge, puisque la suite des sommes partielles est monotone croissante ($a_k \geqslant 0$) et majorée.

2. $c > 1$.

   Puisque $\lim_{k\to+\infty} \sqrt[k]{a_k} = c > 1$, à partir d'un certain rang $K \geqslant k_0$, on a

   $$\sqrt[k]{a_k} > 1 \iff a_k > 1^k = 1 \quad \text{pour tout} \quad k \geqslant K$$

   Ainsi le terme général ne tend pas vers 0. Donc, par la contraposée du théorème fondamental sur les convergences de séries, la série ne converge pas (dans $\mathbb{C}$).

3. $c = 1$.

   La série harmonique $\sum \frac{1}{k}$ ne converge pas et la série $\sum \frac{1}{k^2}$ converge, pourtant pour ces deux séries, on a $c = 1$. En effet, pour $m = 1$ et $m = 2$, on a

   $$c = \lim_{k\to+\infty} \sqrt[k]{\frac{1}{k^m}} = \lim_{k\to+\infty} \sqrt[k]{e^{\ln(\frac{1}{k^m})}} = \lim_{k\to+\infty} e^{-\frac{m\ln(k)}{k}} = e^{-\lim \frac{m\ln(k)}{k}} \overset{\text{Hosp.}}{=} e^{-\lim \frac{m}{k}} = e^0$$

   □

13.3.7 Critère du quotient (ou d'Alembert)

On considère la série infinie $\sum_{k\geqslant K_0} a_k$ avec $a_k \geqslant 0$ pour tout $k \geqslant k_0$.

Si $\lim_{k\to+\infty} \frac{a_{k+1}}{a_k} = c$ (si cette limite existe) et $\begin{cases} c<1, \text{ alors la série converge dans } \mathbb{R} \\ c=1, \text{ alors il y a doute} \\ c>1, \text{ alors la série diverge} \end{cases}$

Preuve

On suppose que la limite existe et vaut c. On distingue trois cas :

1. $c < 1$.

   Soit $r \in]c, 1[$. Puisque $\lim_{k\to+\infty} \frac{a_{k+1}}{a_k} = c < 1$, à partir d'un certain rang $K \geqslant k_0$, on a

   $$\frac{a_{k+1}}{a_k} \leqslant r < 1 \qquad \text{pour tout} \qquad k \geqslant K$$

   De plus, comme $a_k \geqslant 0$ pour tout $k \geqslant k_0$, on sait que $c \geqslant 0$. Donc $r \in [0, 1[$.

   Ainsi, avec un peu d'astuce, on a

   $$\begin{aligned} \sum_{k\geqslant K} a_k &= a_K + a_{K+1} + a_{K+2} + a_{K+3} + \cdots \\ &= a_K \cdot \left(1 + \frac{a_{K+1}}{a_K} + \frac{a_{K+2}}{a_{K+1}} \cdot \frac{a_{K+1}}{a_K} + \frac{a_{K+3}}{a_{K+2}} \cdot \frac{a_{K+2}}{a_{K+1}} \cdot \frac{a_{K+1}}{a_K} + \cdots\right) \\ &\leqslant a_K \cdot (1 + r + r^2 + r^3 + \cdots) \overset{r\in[0,1[}{=} a_K \cdot \frac{1}{1-r} = \frac{a_K}{1-r} \end{aligned}$$

   On a ainsi montré que la suite des sommes partielles est majorée.

   $$\sum_{k=k_0}^{n} a_k \overset{a_k\geqslant 0}{\leqslant} \sum_{k=k_0}^{+\infty} a_k = \sum_{k=k_0}^{K-1} a_k + \sum_{k=K}^{+\infty} a_k \leqslant \sum_{k=k_0}^{K-1} a_k + \frac{a_K}{1-r}$$

   Donc $\sum_{k\geqslant k_0} a_k$ converge, puisque la suite des sommes partielles est monotone croissante ($a_k \geqslant 0$) et majorée.

2. $c > 1$.

   Puisque $\lim_{k\to+\infty} \frac{a_{k+1}}{a_k} = c > 1$, à partir d'un certain rang $K \geqslant k_0$, on a

   $$\frac{a_{k+1}}{a_k} > 1 \iff a_{k+1} > a_k \qquad \text{pour tout} \qquad k \geqslant K$$

   Ainsi, la série finissant par être monotone croissante, son terme général ne tend pas vers 0. Donc, par la contraposée du théorème fondamental sur les convergences de séries, la série ne converge pas (dans $\mathbb{C}$).

3. $c = 1$.

   La série harmonique $\sum \frac{1}{k}$ ne converge pas et la série $\sum \frac{1}{k^2}$ converge, pourtant pour ces deux séries, on a $c = 1$. En effet, pour $m = 1$ et $m = 2$, on a

   $$c = \lim_{k\to+\infty} \frac{\frac{1}{(k+1)^m}}{\frac{1}{k^m}} = \lim_{k\to+\infty} \frac{k^m}{(k+1)^m} = 1$$

□

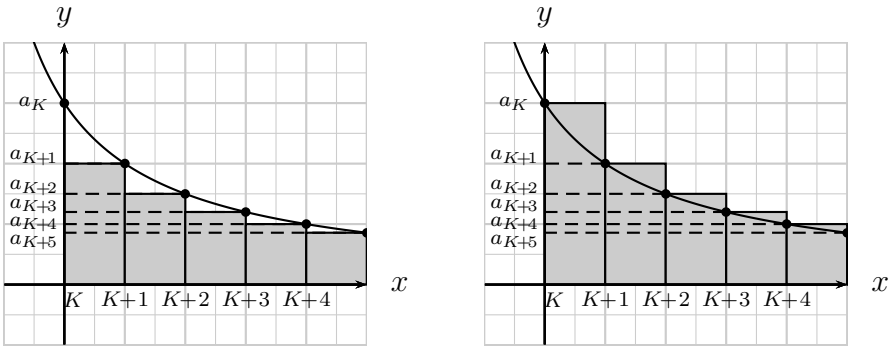
13.3.8 Critère de l'intégrale

On considère la série infinie $\sum_{k\geqslant k_0} a_k$ où $(a_k)_{k\geqslant k_0}$ est une suite décroissante avec $a_k \geqslant 0$ pour tout $k \geqslant k_0$. Supposons qu'il existe un rang $K \geqslant k_0$ et une fonction décroissante définie sur $[K, +\infty[$ telle que $f(k) = a_k$ (pour $k \geqslant K$). Alors

1. si l'intégrale converge dans $\mathbb{R}$, alors la série converge dans $\mathbb{R}$, et réciproquement ;
2. si la série diverge, alors l'intégrale diverge, et réciproquement.

Preuve du critère de l'intégrale

Les hypothèses nous placent dans une situation graphique semblable aux dessins suivants.



On en déduit les chaînes d'inégalités suivantes pour tout $n \geqslant K+1$ ($n = K+5$ sur les schémas).

$$0 \leqslant \sum_{k=K+1}^{n} a_k \leqslant \int_K^n f(x)\,dx \leqslant \sum_{k=K}^{n-1} a_k$$

Donc, comme la suite $\left(\int_K^n f(x)\,dx\right)_{n\geqslant K+1}$ et la suite des sommes partielles sont monotones croissantes, on a les majorations suivantes.

$$\int_K^n f(x)\,dx \leqslant \sum_{k=K}^{+\infty} a_k \quad \text{et} \quad \sum_{k=K+1}^{n} a_k \leqslant \int_K^{+\infty} f(x)\,dx$$

Par conséquent, on peut démontrer les points 1 et 2.

1. si la série est convergente, la suite $\left(\int_K^n f(x)\,dx\right)$ est monotone croissante et majorée, donc converge ! Pour la réciproque, si l'intégrale est convergente, la suite $\left(\sum_{k=k_0}^n a_k\right)$ est monotone croissante et majorée, donc converge.
2. Ce sont les contraposées des deux propositions du point précédent. Elles sont donc aussi vraies. □

Application du critère de l'intégrale

On peut très facilement retrouver les résultats démontrés aux pages 129 et 130 grâce à ce critère. Les fonctions utilisées ci-dessous satisfont bien les hypothèses du critère.

$$\int_1^{+\infty} \frac{1}{x}\,dx \text{ diverge} \implies \sum_{k\geqslant 1} \frac{1}{k} \text{ diverge}$$

$$\int_1^{+\infty} \frac{1}{x^2}\,dx \text{ converge dans } \mathbb{R} \implies \sum_{k\geqslant 1} \frac{1}{k^2} \text{ converge dans } \mathbb{R}$$

13.3.9 Convergence des séries alternées

Définition

Une série est dite *alternée* si elle peut s'écrire de la manière suivante.

$$\sum_{k\geqslant k_0}(-1)^k a_k \quad \text{avec} \quad a_k \geqslant 0$$

Théorème des séries alternées

Une série alternée est convergente dans $\mathbb{R}$ si, pour tout $k \geqslant k_0$, on a

$$\underbrace{a_{k+1} \leqslant a_k}_{\text{la suite } a_k \text{ est monotone décroissante}} \qquad \text{et} \qquad \underbrace{\lim_{k\to+\infty} a_k = 0}_{\text{le terme général de la série tend vers 0}}$$

Preuve

On considère la suite des sommes partielles $(s_n)_{n\geqslant k_0}$ où $s_n = \sum_{k=k_0}^{n}(-1)^k a_k$. On considère encore les sous-suites de sommes partielles $(c_n)_{n\geqslant k_0}$ et $(d_n)_{n\geqslant k_0}$ définies par

$$c_n = s_{2n+1} \qquad \text{et} \qquad d_n = s_{2n}$$

Ces suites satisfont les propriétés suivantes.

1. (c_n) est monotone croissante. En effet, on a

$$\begin{aligned}
&(c_n) \text{ est monotone croissante} \iff c_{n+1} \geqslant c_n \iff s_{2n+3} \geqslant s_{2n+1}\\
&\iff \sum_{k=k_0}^{2n+3}(-1)^k a_k \geqslant \sum_{k=k_0}^{2n+1}(-1)^k a_k \iff \sum_{k=k_0}^{2n+3}(-1)^k a_k - \sum_{k=k_0}^{2n+1}(-1)^k a_k \geqslant 0\\
&\iff \sum_{k=2n+2}^{2n+3}(-1)^k a_k \geqslant 0 \iff a_{2n+2} - a_{2n+3} \geqslant 0 \iff a_{2n+2} \geqslant a_{2n+3}\\
&\iff (a_n) \text{ est monotone décroissante}
\end{aligned}$$

2. (d_n) est monotone décroissante. En effet, on a

$$\begin{aligned}
&(d_n) \text{ est monotone décroissante} \iff d_{n+1} \leqslant d_n \iff s_{2n+2} \leqslant s_{2n}\\
&\iff \sum_{k=k_0}^{2n+2}(-1)^k a_k \leqslant \sum_{k=k_0}^{2n}(-1)^k a_k \iff \sum_{k=k_0}^{2n+2}(-1)^k a_k - \sum_{k=k_0}^{2n}(-1)^k a_k \leqslant 0\\
&\iff \sum_{k=2n+1}^{2n+2}(-1)^k a_k \leqslant 0 \iff -a_{2n+1} + a_{2n+2} \leqslant 0 \iff a_{2n+2} \leqslant a_{2n+1}\\
&\iff (a_n) \text{ est monotone décroissante}
\end{aligned}$$

3. $(d_n - c_n)$ converge vers 0. En effet, on a

$$d_n - c_n = \sum_{k=k_0}^{2n}(-1)^k a_k - \sum_{k=k_0}^{2n+1}(-1)^k a_k = -(-1)^{2n+1} a_{2n+1} = a_{2n+1} \overset{n\to+\infty}{\longrightarrow} 0$$

On peut affirmer (en exercice) que les suites (c_n) et (d_n) convergent vers la même valeur $l \in \mathbb{R}$ et que $c_n \leqslant l \leqslant d_n$ pour tout $n \geqslant k_0$. Ainsi la suite des sommes partielles converge aussi vers cette valeur l. □

Bonus de la preuve. La limite l satisfait $s_{2n+1} \leqslant l \leqslant s_{2n}$ pour tout $n \geqslant k_0$.

La réciproque du théorème des séries alternées est fausse

Si le terme général ne tend pas vers 0, alors la série alternée diverge (c'est la contraposée du théorème fondamental sur les convergences de séries). Cependant, voici deux exemples de séries alternées où la condition de décroissance n'est pas respectée : la première série est convergente dans $\mathbb{R}$; la deuxième ne l'est pas.

Exemple 1

On considère la série alternée

$$\sum_{k\geqslant 1}(-1)^{k+1}a_k \qquad \text{où} \qquad a_k = \begin{cases} \frac{1}{k^2} & \text{si } k \text{ est pair} \\ \frac{2}{k^2} & \text{si } k \text{ est impair} \end{cases}$$

Autrement dit, on a

$$\sum_{k\geqslant 1}(-1)^{k+1}a_k = \frac{2}{1^2} - \frac{1}{2^2} + \frac{2}{3^2} - \frac{1}{4^2} + \frac{2}{5^2} - \frac{1}{6^2} + - \cdots$$

Soit $k \geqslant 4$, k pair, alors on a $a_k < a_{k+1}$. En effet

$$\frac{1}{k^2} < \frac{2}{(k+1)^2} \overset{k>0}{\iff} (k+1)^2 < 2k^2 \iff k^2 - 2k - 1 > 0 \overset{k>0}{\iff} k > 1+\sqrt{2} \cong 2.41$$

Donc la suite $(a_k)_{k\geqslant 1}$ n'est pas monotone décroissante. Pourtant, on a

$$\sum_{k\geqslant 1}(-1)^{k+1}a_k = \underbrace{\frac{1}{1^2} - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \frac{1}{5^2} - \frac{1}{6^2} + - \cdots}_{\text{convergente par le théorème des séries alternées}} + \overbrace{\frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \cdots}^{\text{série croissante et majorée par } \sum_{k\geqslant 1}\frac{1}{k^2}\text{, donc convergente}}$$

Donc cette série alternée converge [2].

Exemple 2

On considère la série alternée

$$\sum_{k\geqslant 1}(-1)^{k+1}a_k \qquad \text{où} \qquad a_k = \begin{cases} \frac{1}{k} & \text{si } k \text{ est pair} \\ \frac{2}{k} & \text{si } k \text{ est impair} \end{cases}$$

Autrement dit, on a

$$\sum_{k\geqslant 1}(-1)^{k+1}a_k = \frac{2}{1} - \frac{1}{2} + \frac{2}{3} - \frac{1}{4} + \frac{2}{5} - \frac{1}{6} + - \cdots$$

Soit $k \geqslant 2$, k pair, alors on a $a_k < a_{k+1}$. En effet

$$\frac{1}{k} < \frac{2}{k+1} \overset{k>0}{\iff} k+1 < 2k \iff k > 1$$

Donc la suite $(a_k)_{k\geqslant 1}$ n'est pas monotone décroissante. Pourtant, on a

$$\sum_{k\geqslant 1}(-1)^{k+1}a_k = \underbrace{\left(\frac{1}{1} - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{5} - \frac{1}{6}\right) + \cdots}_{\geqslant 0} + \frac{1}{1} + \frac{1}{3} + \frac{1}{5} + \cdots \geqslant \overbrace{\frac{1}{1} + \frac{1}{3} + \frac{1}{5} + \cdots}^{\substack{\text{diverge par le} \\ \text{critère de l'intégrale} \\ \int_0^{+\infty}\frac{1}{2x+1}dx = +\infty}}$$

Donc cette série alternée diverge[2].

2. Il faudrait montrer cela formellement avec les suites partielles afin d'être certain d'éviter la subtilité vue en page 142.

13.3.10 Le théorème de la convergence absolue

Définition

Soit une série infinie $\sum_{k\geqslant 0} a_k$ avec $a_k \in \mathbb{C}$.

On dit que cette série *converge absolument* si $\sum_{k\geqslant 0} |a_k|$ converge dans $\mathbb{R}$.

Théorème des séries absolument convergentes

Si une série converge absolument, alors elle converge dans $\mathbb{C}$.

Autrement dit

$$\sum_{k\geqslant 0} |a_k| \text{ converge dans } \mathbb{R} \implies \sum_{k\geqslant 0} a_k \text{ converge dans } \mathbb{C}$$

Preuve (dans le cas où $a_k \in \mathbb{R}$)

On crée les suites $(a_k^+)_{k\geqslant 0}$ et $(a_k^-)_{k\geqslant 0}$ en posant [3]

$$a_k^+ = \max(a_k, 0) \qquad \text{et } a_k^- = \max(-a_k, 0)$$

Ainsi,

si a_k est positif, on a $a_k^+ = a_k$ et $a_k^- = 0$
si a_k est négatif, on a $a_k^+ = 0$ et $a_k^- = -a_k$ $\implies$ Dans tous les cas, on a $a_k = a_k^+ - a_k^-$
si $a_k = 0$, on a $a_k^+ = 0$ et $a_k^- = 0$

Il est donc évident que

$$0 \leqslant a_k^+ \leqslant |a_k| \qquad \text{et } 0 \leqslant a_k^- \leqslant |a_k|$$

Puisque la série converge absolument ($\sum_{k\geqslant 0} |a_k|$ converge), le critère de comparaison affirme que les séries $\sum_{k\geqslant 0} a_n^+$ et $\sum_{k\geqslant 0} a_n^-$ sont convergentes.

Par conséquent, la série $\sum_{k\geqslant 0} a_k$ est convergente. En effet

$$\sum_{k\geqslant 0} a_k = \sum_{k\geqslant 0} \left(a_k^+ - a_k^-\right) = \sum_{k\geqslant 0} a_k^+ \; - \; \sum_{k\geqslant 0} a_k^-$$

□

La réciproque du théorème des séries absolument convergentes est fausse

En effet, la série harmonique alternée

$$\sum_{k\geqslant 0} (-1)^k \frac{1}{k+1} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + - \cdots$$

est une série convergente (grâce au théorème des séries alternées), mais qui n'est pas absolument convergente, car la série harmonique

$$\sum_{k\geqslant 0} \frac{1}{k+1} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \cdots$$

est une série divergente.

3. Cette manière de faire ne fonctionne pas dans $\mathbb{C}$.

13.4 Séries entières et rayon de convergence

Définition

Soit $(a_k)_{k\geqslant 0}$ une suite de nombres (qui pourraient être complexes). On considère une variable x (qui peut aussi s'écrire z si on travaille dans les nombres complexes).

La série $S(x) = \sum_{k\geqslant 0} a_k x^k$ est appelée *série entière* et les coefficients a_k sont appelés *coefficients de la série*.

Conséquences du critère du quotient

En étudiant la convergence absolue de la série entière $S(x) = \sum_{k\geqslant 0} a_k x^k$, c'est-à-dire la convergence de la série $\sum_{k\geqslant 0} |a_k| \cdot |x|^k$, à l'aide du critère du quotient, on obtient

$$c = \lim_{k\to+\infty} \frac{|a_{k+1}| \cdot |x|^{k+1}}{|a_k| \cdot |x|^k} = \lim_{k\to+\infty} \frac{|a_{k+1}|}{|a_k|} \cdot |x|$$

Posons $R = \lim\limits_{k\to+\infty} \dfrac{|a_k|}{|a_{k+1}|}$. On distingue trois cas.

1. $c < 1 \iff |x| < R$.
   La série $S(x)$ converge absolument, donc converge.
2. $c = 1 \iff |x| = R$.
   Il y a doute : il faut étudier la convergence de $S(x)$ pour chaque x tel que $|x| = R$.
3. $c > 1 \iff |x| > R$.
   La série $S(x)$ ne converge pas absolument ; mais a priori elle pourrait converger.

Théorème (sans preuve)

Si $c > 1$, c'est-à-dire si $|x| > R$, alors la série $S(x)$ diverge !

Définition

R est appelé *rayon de convergence de la série entière*. Il est possible que $R = +\infty$.

Exemple

On considère la série entière réelle $S(x) = \sum\limits_{k\geqslant 0} kx^k$.

1. Recherche du rayon de convergence.
   On étudie la convergence absolue de la série grâce au critère du quotient. Après un petit calcul, on trouve que le rayon de convergence est $R = 1$.
2. Étude de la convergence de la série pour $|x| = R$.
   (a) Pour $x = 1$, la série $S(1) = 1 + 2 + 3 + 4 + 5 + 6 + \cdots$ diverge évidemment.
   (b) Pour $x = -1$, la série $S(-1) = \underbrace{(-1)+2}_{=1} + \underbrace{(-3)+4}_{=1} + \underbrace{(-5)+6}_{=1} + \cdots$ diverge.

En conclusion, cette série converge absolument lorsque $x \in]-1, 1[$ et ne converge pas lorsque $x \notin]-1, 1[$.

13.5 Développements de Taylor

13.5.1 Rappel : la tangente à une courbe en un point

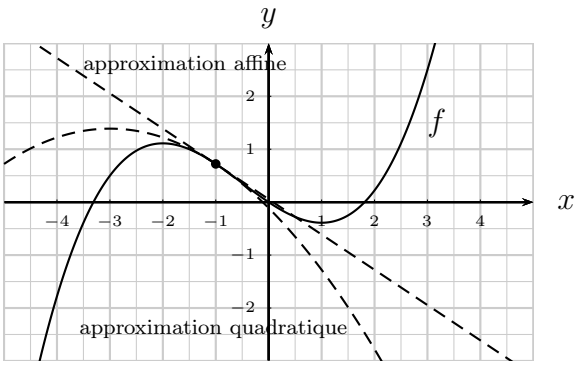
Si $f : D \to A$ est une fonction réelle dérivable, alors on sait que l'équation de la tangente à f en un point $x_0 \in D$ est

$$y = f(x_0) + f'(x_0)(x - x_0)$$

On peut donc approximer la fonction f autour de x_0 par la fonction affine

$$d(x) = f(x_0) + f'(x_0)(x - x_0)$$

Cette fonction satisfait les propriétés $d(x_0) = f(x_0)$ et $d'(x_0) = f'(x_0)$.



Si on cherche une approximation par une fonction quadratique $p(x)$, on va vouloir que cette fonction satisfasse :

$$p(x_0) = f(x_0) \quad \text{et} \quad p'(x_0) = f'(x_0) \quad \text{et} \quad p''(x_0) = f''(x_0)$$

Une telle fonction existe et vaut

$$p(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2$$

13.5.2 Théorème de Taylor

Théorème de Taylor

Soit $n \in \mathbb{N}$, $n > 0$. Soit $f : [a, b] \to \mathbb{R}$ une fonction que l'on peut dériver $n + 1$ fois et dont toutes les dérivées jusqu'à l'ordre $n + 1$ sont continues. Soit x et $x_0 \in [a, b]$. Alors il existe ξ entre x et x_0 tel que

$$f(x) = \underbrace{\sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k}_{\text{approximation de } f \text{ par un polynôme de degré } n} + \underbrace{\frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1}}_{\text{Reste de Lagrange}}$$

Il s'agit du *développement limité de Taylor d'ordre n en x_0 de la fonction f*. Lorsque $x_0 = 0$, on parle de *développement limité de Maclaurin d'ordre n de f*.

Si on note $p_n(x)$ le polynôme de degré n ci-dessus et qu'on déplie la somme, on a :

$$p_n(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

Ce polynôme satisfait les propriétés suivantes.

$$p_n(x_0) = f(x_0)\,,\ p_n'(x_0) = f'(x_0)\,,\ \ldots\,,\ p_n^{(m)}(x_0) = f^{(m)}(x_0)\,,\ \ldots\,,\ p_n^{(n)}(x_0) = f^{(n)}(x_0)$$

Preuve

Idée : on fixe $x \in [a,b]$ et on considère la fonction en une nouvelle variable y suivante.

$$\tilde{p}_x(y) = f(y) + f'(y)(x-y) + \frac{f''(y)}{2}(x-y)^2 + \cdots + \frac{f^{(n)}(y)}{n!}(x-y)^n$$

Puis, on considère la fonction φ en y suivante.

$$\varphi(y) = f(x) - \tilde{p}_x(y)$$

Avec cette façon de noter, $\tilde{p}_x(x_0)$ est l'approximation de f par un polynôme de degré n et on trouve une formule équivalente à la formule de l'encadré du théorème de Taylor.

$$f(x) = \tilde{p}_x(x_0) + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1} \iff f(x) - \tilde{p}_x(x_0) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}$$

Ainsi, il faut montrer qu'il existe ξ entre x et x_0 tel que $\varphi(x_0) = \dfrac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}$.
La fonction φ satisfait :

$$i) \quad \varphi'(y) = -\frac{f^{(n+1)}(y)}{n!}(x-y)^n \qquad ii) \quad \varphi(x) = 0$$

En effet $i)$ s'obtient en dérivant directement par rapport à y (somme télescopique) et $ii)$ s'obtient grâce au fait que $\tilde{p}_x(x) = f(x)$.

Le terme ξ apparaît grâce au théorème de Rolle. Pour pouvoir appliquer le théorème de Rolle, il faut trouver une fonction continue F qui s'annule en x_0 et en x. Rolle nous permettra ainsi d'affirmer qu'il existe ξ entre x et x_0 tel que $F'(\xi) = 0$. L'astuce finale consiste à bien choisir F ! Voici cette fonction :

$$F(y) = \varphi(y)(x-x_0)^{n+1} - \varphi(x_0)(x-y)^{n+1} = \begin{vmatrix} \varphi(y) & (x-y)^{n+1} \\ \varphi(x_0) & (x-x_0)^{n+1} \end{vmatrix}$$

On a $F(x) = \varphi(x)(x-x_0)^{n+1} \overset{ii)}{=} 0$ et évidemment $F(x_0) = 0$. Donc, par Rolle, il existe ξ entre x et x_0 tel que $F'(\xi) = 0$. Or, la dérivée de F par rapport à y est :

$$\begin{aligned} F'(y) &= \varphi'(y)(x-x_0)^{n+1} - \varphi(x_0)(n+1)(x-y)^n(-1) \\ &\overset{i)}{=} -\frac{f^{(n+1)}(y)}{n!}(x-y)^n(x-x_0)^{n+1} + \varphi(x_0)(n+1)(x-y)^n \end{aligned}$$

Donc

$$0 = F'(\xi) = -\frac{f^{(n+1)}(\xi)}{n!}(x-\xi)^n(x-x_0)^{n+1} + \varphi(x_0)(n+1)(x-\xi)^n$$

En simplifiant par $(x-\xi)^n$, on obtient ce qu'on voulait

$$\begin{aligned} 0 = -\frac{f^{(n+1)}(\xi)}{n!}(x-x_0)^{n+1} + \varphi(x_0)(n+1) &\iff (n+1)\varphi(x_0) = \frac{f^{(n+1)}(\xi)}{n!}(x-x_0)^{n+1} \\ &\iff \varphi(x_0) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1} \end{aligned}$$

□

Conséquence

Le reste de Lagrange permet d'estimer l'erreur commise par l'estimation à l'aide du développement limité. En effet, on a

$$|f(x)-p_n(x)| = \left|\frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}\right| \leqslant \frac{|x-x_0|^{n+1}}{(n+1)!}\cdot \max_{\xi \text{ entre } x_0 \text{ et } x} |f^{(n+1)}(\xi)|$$

Définition

Si lorsque $n \to +\infty$, le développement limité de Taylor d'ordre n en x_0 de la fonction f s'approche de f (il faut pour cela que le reste de Lagrange tende vers 0 et que la série converge), on a

$$f(x) = \sum_{k\geqslant 0} \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k$$

C'est *la série de Taylor de f*. Lorsque $x_0 = 0$, c'est *la série de Maclaurin de f*.

13.5.3 Les séries de Maclaurin des fonctions exp, cos et sin

1. Pour $f(x) = e^x$, le reste de Lagrange tend vers 0 lorsque $n \to +\infty$. De plus, la série de Maclaurin suivante est convergente pour tout $x \in \mathbb{R}$ (même $x \in \mathbb{C}$).

   $$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \frac{x^7}{7!} + \cdots \iff e^x = \sum_{k\geqslant 0} \frac{x^k}{k!}$$

   Comme $f(1) = e$, on a une nouvelle formule pour calculer le nombre e :

   $$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + \frac{1}{6!} + \frac{1}{7!} + \cdots \iff e = \sum_{k\geqslant 0} \frac{1}{k!} = \lim_{n\to+\infty} \sum_{k=0}^{n} \frac{1}{k!}$$

   Cette formule est bien plus efficace que $e = \lim\limits_{k\to+\infty} \left(1+\frac{1}{k}\right)^k$.

2. Pour $f(x) = \cos(x)$, le reste de Lagrange tend vers 0 lorsque $n \to +\infty$. De plus, la série de Maclaurin suivante est convergente pour tout $x \in \mathbb{R}$ (même $x \in \mathbb{C}$).

   $$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \frac{x^{10}}{10!} + - \cdots \iff \cos(x) = \sum_{k\geqslant 0} (-1)^k \frac{x^{2k}}{(2k)!}$$

3. Pour $f(x) = \sin(x)$, le reste de Lagrange tend vers 0 lorsque $n \to +\infty$. De plus, la série de Maclaurin suivante est convergente pour tout $x \in \mathbb{R}$ (même $x \in \mathbb{C}$).

   $$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \frac{x^{11}}{11!} + - \cdots \iff \sin(x) = \sum_{k\geqslant 0} (-1)^k \frac{x^{2k+1}}{(2k+1)!}$$

On peut ainsi montrer que

$$\boxed{e^{ix} = \cos(x) + i\sin(x) \quad \text{avec} \quad x \in \mathbb{R}}$$

En effet, on a

$$e^{ix} = 1 + ix - \frac{x^2}{2!} - i\frac{x^3}{3!} + \frac{x^4}{4!} + i\frac{x^5}{5!} - \frac{x^6}{6!} - i\frac{x^7}{7!} + \frac{x^8}{8!} + i\frac{x^9}{9!} - \frac{x^{10}}{10!} - i\frac{x^{11}}{11!} + \cdots$$

En posant $x = \pi$, on obtient une des plus extraordinaires relations des mathématiques.

$$\boxed{\boxed{e^{i\pi} + 1 = 0}} \quad \text{car} \quad e^{i\pi} = \cos(\pi) + i\sin(\pi) = -1 + 0 = -1$$

13.5.4 Une autre façon d'exprimer le reste de Lagrange

Si f est continue sur $[a,b]$ et si ses $(n+1)$ dérivées successives sont continues, alors

$$f(x) - p_n(x) = \int_{x_0}^{x} \frac{f^{(n+1)}(t)}{n!}(x-t)^n dt$$

Cette formule se démontre facilement par récurrence : pour $n = 0$, on retrouve le théorème fondamental du calcul intégral ; pour le pas de récurrence, on écrit l'intégrale ci-dessus de deux manière : la première à l'aide de l'hypothèse de récurrence ; la deuxième en intégrant par parties.

13.5.5 La série de Maclaurin de $\ln(x+1)$

Pour la fonction $f(x) = \ln(x+1)$, la série de Maclaurin est la suivante.

$$\ln(x+1) = S(x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \frac{x^5}{5} - \frac{x^6}{6} + \frac{x^7}{7} - \frac{x^8}{8} + \cdots = \sum_{k\geqslant 1} (-1)^{k+1}\frac{x^k}{k}$$

Étudions la convergence de cette série entière.

1. Recherche du rayon de convergence : on étudie la convergence absolue de la série grâce au critère du quotient (ou d'Alembert).

   $$|x| + \frac{|x|^2}{2} + \frac{|x|^3}{3} + \frac{|x|^4}{4} + \frac{|x|^5}{5} + \frac{|x|^6}{6} + \frac{|x|^7}{7} + \frac{|x|^8}{8} + \cdots$$

   On calcule la valeur du nombre c utilisé dans le critère du quotient

   $$c = \lim_{k\to+\infty} \frac{\frac{|x|^{k+1}}{k+1}}{\frac{|x|^k}{k}} = \lim_{k\to+\infty} |x|\frac{k}{k+1} = |x|$$

   Donc, par le critère du quotient, si $|x| < 1$, la série converge absolument et si $|x| > 1$, la série ne converge pas absolument. Ainsi, le rayon de convergence est $R = 1$.

2. Étude de la convergence de la série pour $|x| = R = 1$.

   (a) Pour $x = -1$, on a

   $$S(-1) = -\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots\right)$$

   Donc $-S(-1)$ est la série harmonique, donc diverge.

   (b) Pour $x = 1$, la série est la série harmonique alternée, donc converge par le théorème des séries alternées.

En conclusion, cette série ne converge que pour $x \in]-1,1]$ (elle ne converge absolument que pour $x \in]-1,1[$).

On peut facilement montrer que si $x \in \left[-\frac{1}{2},1\right]$, alors le reste de Lagrange tend vers 0. C'est bien moins facile pour $x \in \left]-1,-\frac{1}{2}\right[$.

13.5.6 Une dernière subtilité

On a donc trouvé la valeur de la série harmonique alternée.

$$\ln(2) = \ln(1+1) = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \frac{1}{9} - \frac{1}{10} + \cdots$$

Regardons ce qu'il se passe si on change l'ordre dans lequel on additionne les termes de cette série.

$$\begin{aligned}
&\underbrace{1 - \frac{1}{2}}_{=\frac{1}{2}} - \frac{1}{4} + \underbrace{\frac{1}{3} - \frac{1}{6}}_{=\frac{1}{6}} - \frac{1}{8} + \underbrace{\frac{1}{5} - \frac{1}{10}}_{=\frac{1}{10}} - \frac{1}{12} + \underbrace{\frac{1}{7} - \frac{1}{14}}_{=\frac{1}{14}} - \frac{1}{16} + \underbrace{\frac{1}{9} - \frac{1}{18}}_{=\frac{1}{18}} - \frac{1}{20} + \cdots \\
&= \frac{1}{2} - \frac{1}{4} + \frac{1}{6} - \frac{1}{8} + \frac{1}{10} - \frac{1}{12} + \frac{1}{14} - \frac{1}{16} + \frac{1}{18} - \frac{1}{20} + \cdots \\
&= \frac{1}{2}\left(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \frac{1}{9} - \frac{1}{10} + \cdots\right) \\
&= \tfrac{1}{2}\ln(2)
\end{aligned}$$

Ainsi, si on change l'ordre dans lequel l'addition est effectuée, alors on change la valeur de la série.

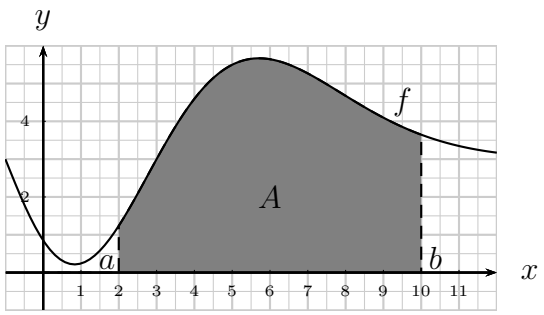
C'est pour cette raison qu'il ne faut jamais oublier qu'une série est une limite de sommes partielles. Le changement d'ordre ci-dessus change complètement les sommes partielles et la série n'a donc rien à voir avec la série de départ !

Chapitre 14

L'intégration numérique

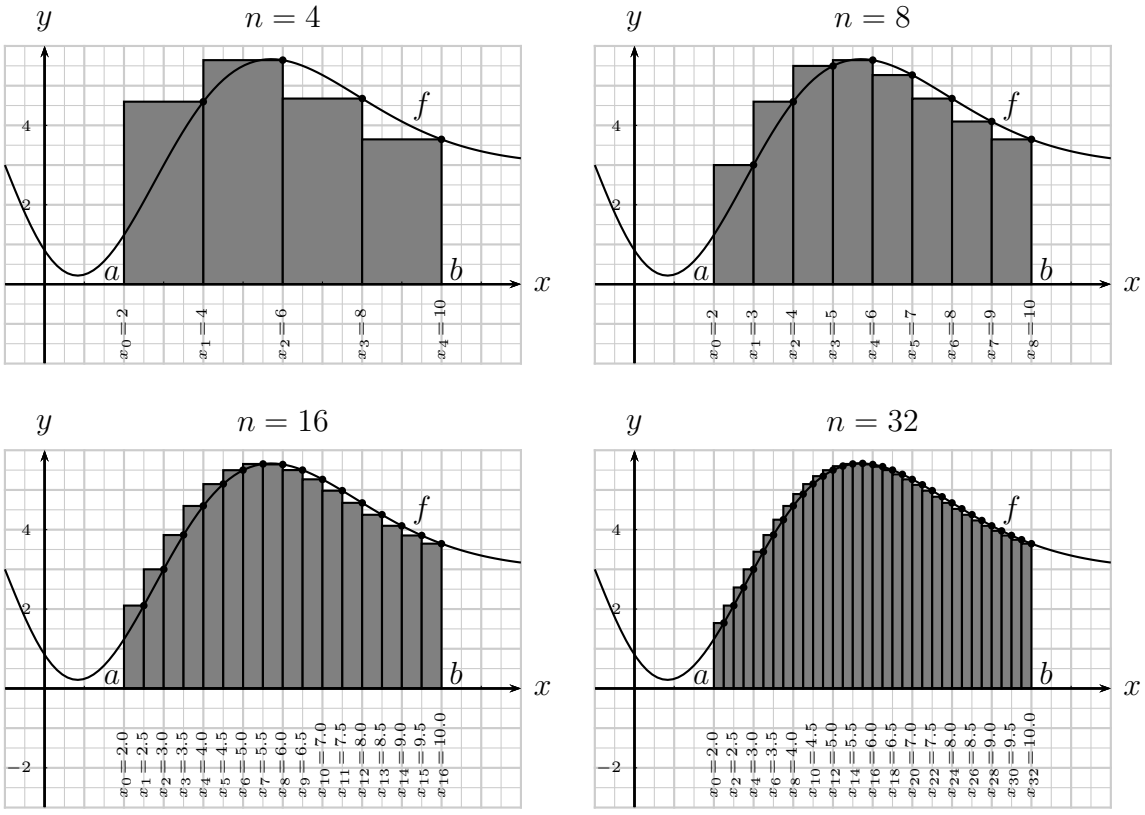
14.1 Définition intuitive

L'*intégrale de la fonction f entre les bornes a et b* est l'aire signée entre la fonction, l'axe des x et les axes verticaux $x = a$ et $x = b$.



14.2 Définition formelle

Commençons par supposer que $a < b$.
Pour calculer cette aire, on découpe l'intervalle $[a, b]$ en n intervalles.



Ainsi, lorsque $n \to +\infty$, alors la somme des aires des rectangles tend vers l'intégrale A.

Formellement, on procède ainsi :

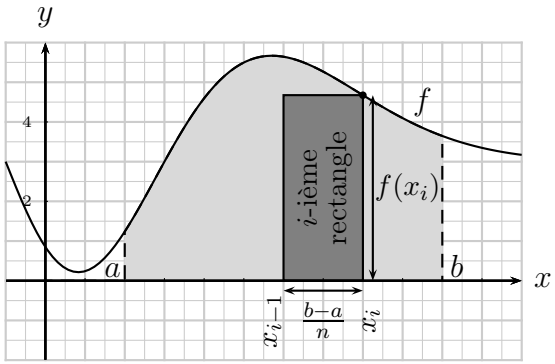
1. On commence par subdiviser l'intervalle $[a, b]$ en n intervalles $[x_{i-1}, x_i]$ avec $i \in \{1, \ldots, n\}$. Cela permet d'approcher l'aire cherchée en calculant l'aire des n rectangles dont le coin droit touche le graphe de la fonction, donc la hauteur du i-ième rectangle est $f(x_i)$.

   L'aire du i-ième rectangle est donnée par la célèbre formule "hauteur fois base", ainsi

   $$\text{Aire du } i\text{-ième rectangle } = f(x_i) \cdot \frac{b-a}{n}$$

   De ce fait, l'aire de tous les rectangles vaut :

$$\sum_{i=1}^{n} \overbrace{\left(\underbrace{f(x_i)}_{\substack{\text{hauteur}\\ \text{du } i\text{-ème}\\ \text{rectangle}}} \cdot \underbrace{\frac{b-a}{n}}_{\substack{\text{base}\\ \text{du } i\text{-ème}\\ \text{rectangle}}} \right)}^{\text{Aire du } i\text{-ième rectangle}}$$



2. On fait ensuite tendre n vers l'infini (et par conséquent la longueur des intervalles de la subdivision vers 0), l'aire totale de tous les rectangles va tendre vers l'aire A cherchée.

On peut donc écrire :

$$A = \lim_{n \to +\infty} \left(\sum_{i=1}^{n} \left(f(x_i) \cdot \frac{b-a}{n} \right) \right)$$

Comme la longueur de chaque intervalle de la subdivision tend vers 0 lorsque n tend vers l'infini, on peut la remplacer par Δx (le lecteur se rappellera le chapitre de la dérivée où Δx symbolisait un nombre étant sensé être très petit). Autrement dit, la formule devient

$$A = \lim_{n \to +\infty} \left(\sum_{i=1}^{n} f(x_i) \cdot \Delta x \right) \quad \text{si on note } \Delta x = \frac{b-a}{n}$$

Notation

De nos jours, on note l'aire sous le graphe de la fonction f entre les points a et b de la façon suivante.

$$\boxed{A = \int_a^b f(x)\, dx}$$

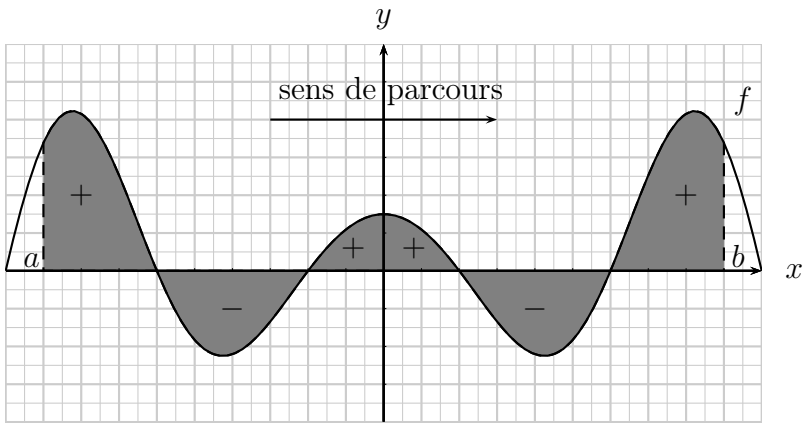
a, b *bornes d'intégration*
x *variable d'intégration*

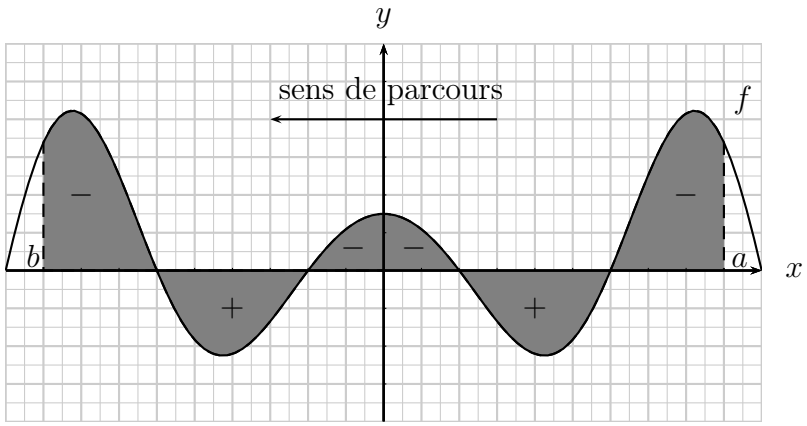
Il s'agit de l'*intégrale (définie) de la fonction f de a à b.*

C'est une transformation visuelle de l'écriture ci-dessus, on remplace Δx par dx et la limite de la somme par un S déformé en $\int$. On bascule aussi les bornes a et b en dessous et en dessus de ce symbole afin de ne pas les oublier.

Conséquence importante de la définition

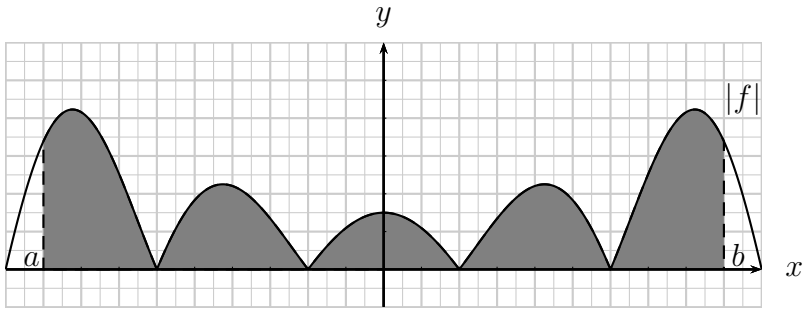
Lorsque la fonction change de signe, l'intégrale ne donne pas l'aire entre le graphe et l'axe horizontal. En effet, l'aire sur chaque morceau sera comptée avec un signe. On a un deuxième changement de signe lorsque $a > b$ (au lieu de $a < b$).





14.2.1 Pour être sûr d'avoir l'aire

Si on désire vraiment calculer la surface entre le graphe et l'axe, on utilise la valeur absolue pour passer tout le graphe au dessus de l'axe horizontal.



On peut réaliser cela grâce à la valeur absolue[1]. Il faut donc calculer

$$\int_a^b |f(x)|\,dx$$

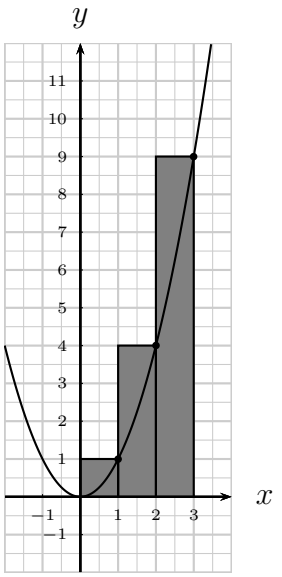
et s'assurer que a est bien plus petit que b.

1. On peut aussi intégrer sur chaque morceau et tenir compte des signes 'à la main'.

14.3 Exemples

Aire sous une parabole

Calculons l'aire sous la parabole $f(x) = x^2$ entre 0 et $b > 0$.



On commence par subdiviser l'intervalle $[0, b]$ en n morceaux (ci-contre, l'intervalle $[0, 3]$ est subdivisé en 3, puis en 6 morceaux).

Les x_i sont ici donnés par $x_i = \frac{b}{n} \cdot i$ pour $i \in \{1, \ldots, n\}$. Ainsi, le i-ième rectangle est de hauteur $f(x_i)$ et de base $\frac{b}{n}$. On calcule l'aire de tous les rectangles, notée A_n, comme suit :

$$A_n = \sum_{i=1}^{n} \left(f(x_i) \cdot \frac{b}{n} \right) = \sum_{i=1}^{n} \left(x_i^2 \cdot \frac{b}{n} \right)$$

En substituant x_i, on obtient :

$$\begin{aligned} A_n &= \sum_{i=1}^{n} \left(\left(\frac{b}{n} \right)^3 i^2 \right) = \left(\frac{b}{n} \right)^3 \sum_{i=1}^{n} i^2 \\ &= \left(\frac{b}{n} \right)^3 \left(1^2 + 2^2 + 3^2 + \cdots + n^2 \right) \end{aligned}$$

On peut faire progresser le calcul en utilisant la formule (qui peut se démontrer par récurrence) suivante :

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n \cdot (n+1) \cdot (2n+1)}{6}$$

Ainsi, on a :

$$\begin{aligned} A_n &= \left(\frac{b}{n} \right)^3 \cdot \frac{n \cdot (n+1) \cdot (2n+1)}{6} \\ &= \frac{b^3}{6} \cdot \frac{n \cdot (n+1) \cdot (2n+1)}{n \cdot n \cdot n} \\ &= \frac{b^3}{6} \left(1 + \frac{1}{n} \right) \left(2 + \frac{1}{n} \right) \end{aligned}$$

Lorsqu'on fait tendre le nombre de tranches n vers l'infini, on obtient l'aire suivante

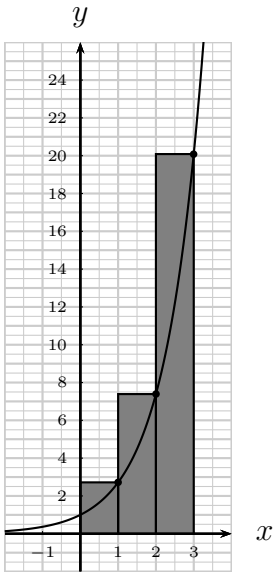
$$A = \lim_{n \to +\infty} A_n = \lim_{n \to +\infty} \frac{b^3}{6} \underbrace{\left(1 + \frac{1}{n} \right)}_{\to 1} \underbrace{\left(2 + \frac{1}{n} \right)}_{\to 2} = \frac{b^3}{3}$$

Ainsi, on a montré que :

$$\int_0^b x^2 \, dx = \frac{b^3}{3}$$

Aire sous une exponentielle

Calculons l'aire sous l'exponentielle $f(x) = e^x$ entre 0 et $b > 0$.



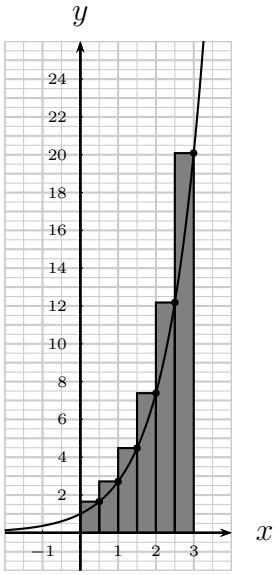
On commence par subdiviser l'intervalle $[0, b]$ en n morceaux (ci-contre, l'intervalle $[0, 3]$ est subdivisé en 3, puis en 6 morceaux).

Les x_i sont aussi donnés par $x_i = \frac{b}{n} \cdot i$ pour $i \in \{1, \ldots, n\}$. Ainsi, le i-ième rectangle est de hauteur $f(x_i)$ et de base $\frac{b}{n}$. On calcule l'aire de tous les rectangles, notée A_n, comme suit :

$$A_n = \sum_{i=1}^{n} \left(f(x_i) \cdot \frac{b}{n} \right) = \sum_{i=1}^{n} \left(e^{x_i} \cdot \frac{b}{n} \right)$$

En substituant x_i, on obtient :

$$\begin{aligned} A_n &= \sum_{i=1}^{n} \left(e^{\frac{b}{n} i} \frac{b}{n} \right) = \frac{b}{n} \sum_{i=1}^{n} e^{\frac{b}{n} i} \\ &= \frac{b}{n} \left(e^{\frac{b}{n}} + \left(e^{\frac{b}{n}}\right)^2 + \left(e^{\frac{b}{n}}\right)^3 + \cdots + \left(e^{\frac{b}{n}}\right)^n \right) \\ &= e^{\frac{b}{n}} \frac{b}{n} \left(1 + \left(e^{\frac{b}{n}}\right) + \left(e^{\frac{b}{n}}\right)^2 + \cdots + \left(e^{\frac{b}{n}}\right)^{n-1} \right) \end{aligned}$$



Il s'agit d'une progression géométrique dont la formule est

$$1 + r + r^2 + r^3 + \cdots + r^{n-1} = \frac{r^n - 1}{r - 1} \quad \left(= \frac{1 - r^n}{1 - r} \right)$$

Ainsi, on a :

$$\begin{aligned} A_n &= e^{\frac{b}{n}} \frac{b}{n} \cdot \frac{\left(e^{\frac{b}{n}}\right)^n - 1}{e^{\frac{b}{n}} - 1} \\ &= e^{\frac{b}{n}} \frac{b}{n} \cdot \frac{e^b - 1}{e^{\frac{b}{n}} - 1} \end{aligned}$$

Pour obtenir l'aire sous la courbe, on utilise le théorème de l'Hospital pour calculer la limite de l'aire A_n lorsque le nombre de tranches n tend vers l'infini (ici, n est la variable).

$$\begin{aligned} A &= \lim_{n \to +\infty} A_n = \lim_{n \to +\infty} e^{\frac{b}{n}} \frac{b}{n} \cdot \frac{e^b - 1}{e^{\frac{b}{n}} - 1} = \lim_{n \to +\infty} (e^b - 1)\, e^{\frac{b}{n}} \cdot \frac{\frac{b}{n}}{e^{\frac{b}{n}} - 1} \\ &= (e^b - 1) \cdot \underbrace{\lim_{n \to +\infty} e^{\frac{b}{n}}}_{\to 1} \cdot \lim_{n \to +\infty} \frac{\frac{b}{n}}{e^{\frac{b}{n}} - 1} \overset{\text{Hospital}}{=} (e^b - 1) \cdot \lim_{n \to +\infty} \frac{-\frac{b}{n^2}}{e^{\frac{b}{n}} \cdot \left(-\frac{b}{n^2}\right)} \\ &= (e^b - 1) \cdot \underbrace{\lim_{n \to +\infty} \frac{1}{e^{\frac{b}{n}}}}_{\to 1} = e^b - 1 \end{aligned}$$

Ainsi, on a montré que :

$$\int_0^b e^x \, dx = e^b - 1$$

14.4 Le théorème fondamental du calcul intégral

Ce théorème permet de calculer une intégrale (sous certaines hypothèses) en passant par le calcul différentiel. La preuve se trouve dans le cours de discipline fondamentale.

14.4.1 Théorème

Soit f une fonction réelle continue définie sur l'intervalle $[a, b]$.
Alors on a

$$\boxed{\int_a^b f(x)\ dx = F(x)\Big|_a^b} \qquad \text{où} \qquad F(x)\Big|_a^b \overset{\text{Notation}}{=} F(b) - F(a)$$

où F est une primitive quelconque de f (c'est-à-dire une fonction telle que $F' = f$).

14.4.2 Exemples de calcul d'intégrales avec ce théorème

Reprenons les intégrales effectuées précédemment à l'aide de la définition.

1. Si $f(x) = x^2$, alors $F(x) = \frac{x^3}{3}$ est une primitive de f et on a :

$$\int_0^b x^2\ dx = F(x)\Big|_0^b = F(b) - F(0) = \frac{b^3}{3} - \frac{0^3}{3} = \frac{b^3}{3}$$

2. Si $f(x) = e^x$, alors $F(x) = e^x + 2$ est une primitive de f et on a :

$$\int_0^b e^x\ dx = F(x)\Big|_0^b = F(b) - F(0) = e^b + 2 - (e^0 + 2) = e^b - 1$$

14.5 Le point de vue numérique

Le théorème fondamental du calcul intégral n'est pas toujours applicable, car il existe des fonctions continues sur $\mathbb{R}$ qui n'admettent pas de primitives explicites. La fonction ci-dessous est une telle fonction qui joue un rôle très important en statistique car elle permet de décrire la densité de la loi normale.

$$f(x) = e^{-x^2}$$

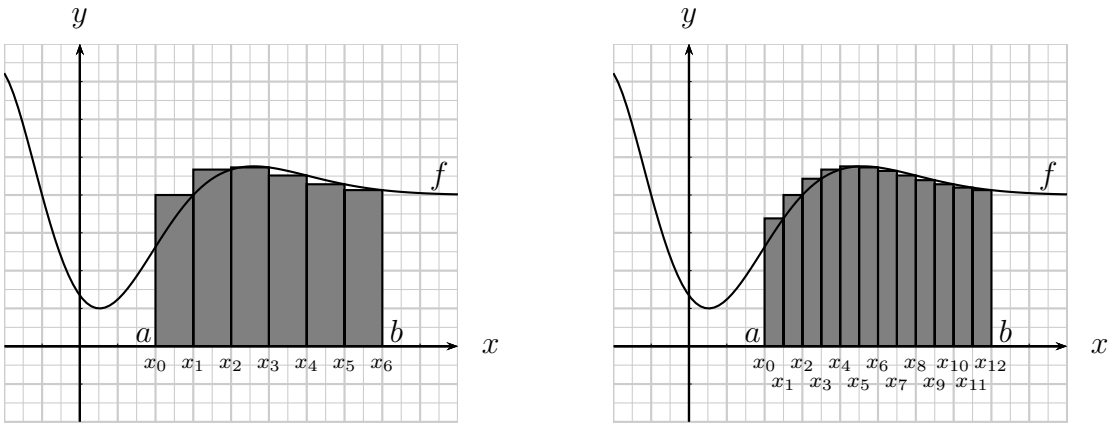
Il est alors nécessaire de revenir à la définition même de l'intégrale. Mais cette dernière étant plus technique, il devient nécessaire d'utiliser un ordinateur pour calculer une approximation de cette intégrale (le passage à la limite discuté dans la définition devient délicat à cause des erreurs d'arrondi).

Les numériciens ont ainsi cherché de nouvelles façons pour faire de meilleures approximations. La première méthode, la méthode des approximations à gauche, est celle de la définition donnée dans ce cours. De manière équivalente, il y a aussi la méthode des approximations à droite. On présentera ensuite la méthode du point médian (aussi appelée méthode des rectangles) et la méthode des trapèzes. Puis, on parlera brièvement de la méthode de Simpson.

14.5.1 La méthode des approximations à droite

On subdivise l'intervalle $[a, b]$ en n intervalles $[x_{i-1}, x_i]$ (avec $i \in \{1, \ldots, n\}$) de même largeur. On a ainsi l'approximation suivante :

$$\int_a^b f(x)\ dx \quad \approx \quad D_n \overset{\text{définition}}{=} \frac{b-a}{n} \cdot \sum_{i=1}^{n} f(x_i) \qquad \begin{array}{c} x_i \text{ est la borne de droite} \\ \text{du } i\text{-ème intervalle} \end{array}$$



Exemple

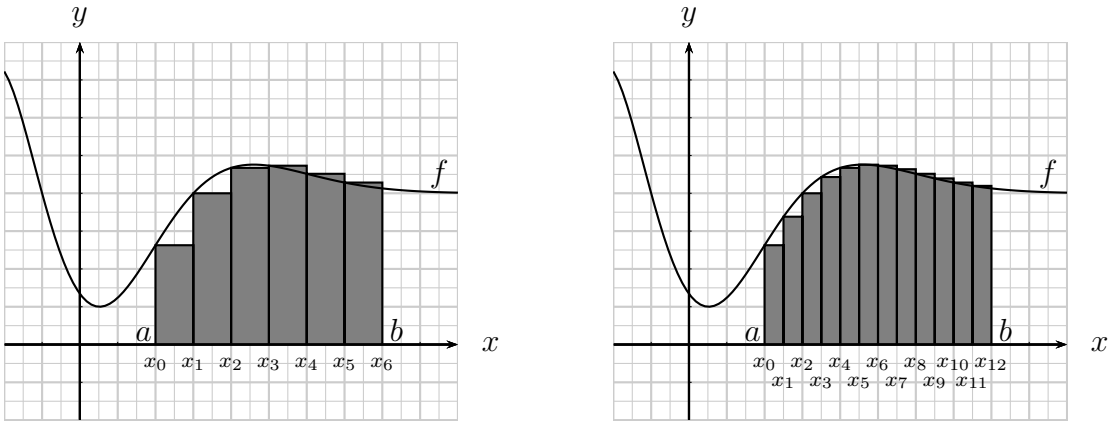
Si on cherche à estimer l'intégrale $\int_0^6 x^2\ dx = \frac{6^3}{3} = 72$, on obtient $D_6 = 91$ et $D_{12} = 81.25$. $(D_6 = 1 \cdot (1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = 91)$.

14.5.2 La méthode des approximations à gauche

On subdivise l'intervalle $[a, b]$ en n intervalles $[x_{i-1}, x_i]$ (avec $i \in \{1, \ldots, n\}$) de même largeur. On a ainsi l'approximation suivante :

$$\int_a^b f(x)\ dx \quad \approx \quad G_n \overset{\text{définition}}{=} \frac{b-a}{n} \cdot \sum_{i=1}^{n} f(x_{i-1}) \qquad \begin{array}{c} x_{i-1} \text{ est la borne de gauche} \\ \text{du } i\text{-ème intervalle} \end{array}$$

C'est la méthode présentée dans la définition.



Exemple

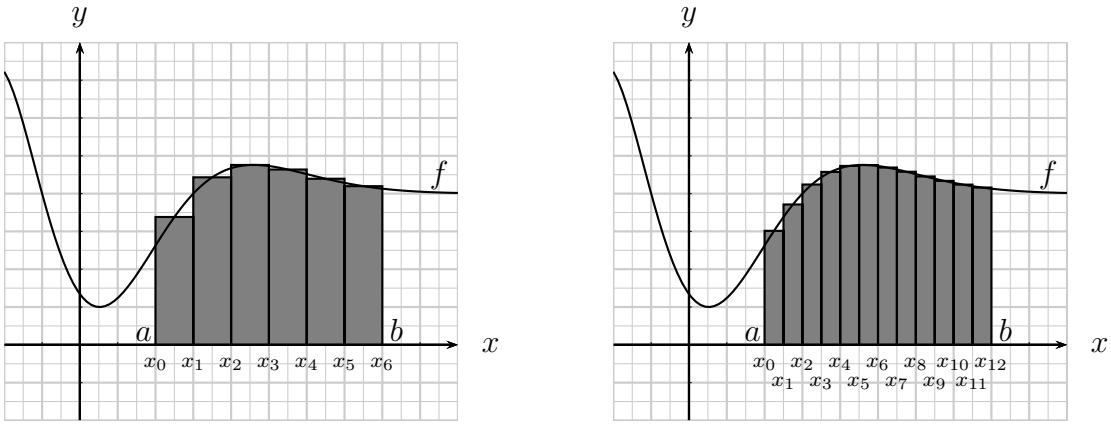
Si on cherche à estimer l'intégrale $\int_0^6 x^2\ dx = \frac{6^3}{3} = 72$, on obtient $G_6 = 55$ et $G_{12} = 63.25$. $(G_6 = 1 \cdot (0^2 + 1^2 + 2^2 + 3^2 + 4^2 + 5^2) = 55)$.

14.5.3 La méthode du point médian (ou méthode des rectangles)

On subdivise l'intervalle $[a, b]$ en n intervalles $[x_{i-1}, x_i]$ (avec $i \in \{1, \ldots, n\}$) de même largeur. On a ainsi l'approximation suivante :

$$\int_a^b f(x)\,dx \quad \approx \quad M_n \overset{\text{définition}}{=} \frac{b-a}{n} \cdot \sum_{i=1}^{n} f\Big(\frac{x_{i-1}+x_i}{2}\Big)$$

$\frac{x_{i-1}+x_i}{2}$ est le milieu du i-ème intervalle



Exemple

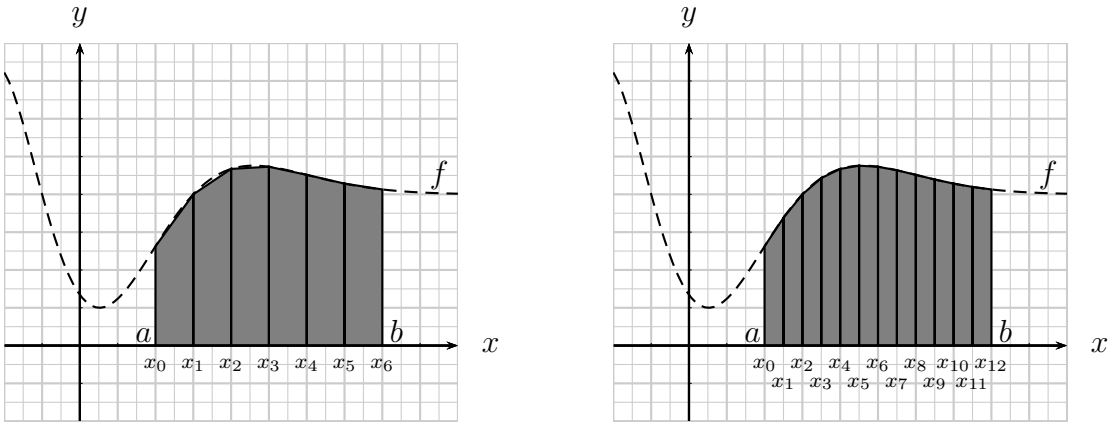
Si on cherche à estimer l'intégrale $\int_0^6 x^2\,dx = \frac{6^3}{3} = 72$, on a $M_6 = 71.5$ et $M_{12} = 71.875$. $(M_6 = 1 \cdot (0.5^2 + 1.5^2 + 2.5^2 + 3.5^2 + 4.5^2 + 5.5^2) = 71.5)$.

14.5.4 La méthode des trapèzes

On subdivise l'intervalle $[a, b]$ en n intervalles $[x_{i-1}, x_i]$ (avec $i \in \{1, \ldots, n\}$) de même largeur. On a ainsi l'approximation suivante :

$$\int_a^b f(x)\,dx \quad \approx \quad T_n \overset{\text{définition}}{=} \frac{b-a}{n} \cdot \sum_{i=1}^{n} \frac{f(x_{i-1}) + f(x_i)}{2}$$

$\frac{f(x_{i-1})+f(x_i)}{2}$ est la hauteur moyenne du i-ème trapèze



Exemple

Si on cherche à estimer l'intégrale $\int_0^6 x^2\,dx = \frac{6^3}{3} = 72$, on a $T_6 = 73$ et $T_{12} = 72.25$. $(T_6 = 1 \cdot (\frac{0^2}{2} + 1^2 + 2^2 + 3^2 + 4^2 + 5^2 + \frac{6^2}{2}) = 73)$.

Remarque. Cette méthode consiste à utiliser la méthode des rectangles sur une interpolation linéaire de la fonction au-dessus des intervalles de la subdivision.

14.5.5 Critères d'arrêts de ces méthodes

Chaque méthode nous donne une suite d'approximations qui converge vers la valeur exacte de l'intégrale. On va donc utiliser le critère d'arrêt basé sur l'erreur absolue au n-ième pas de la suite par rapport à la valeur exacte (ce critère est le même que celui utilisé dans le chapitre sur les résolutions d'équations numériques).

Définition

Si $(x_n)_{n\geqslant 1}$ est une suite qui converge vers x_0. L'*erreur (absolue) au pas n* est définie par :

$$e_n = |x_n - x_0|$$

Théorème

Pour les méthodes des approximations à gauche et à droite, on a les majorations de l'erreur au pas n suivantes :

$$\boxed{|e_n^{(G)}| \leqslant \frac{(b-a)^2}{2n} \cdot \max_{x\in[a,b]} |f'(x)|} \quad \text{et} \quad \boxed{|e_n^{(D)}| \leqslant \frac{(b-a)^2}{2n} \cdot \max_{x\in[a,b]} |f'(x)|}$$

Pour les méthodes du point médian (méthode des rectangles) et la méthode des trapèzes, on a les majorations de l'erreur au pas n suivantes :

$$\boxed{|e_n^{(M)}| \leqslant \frac{(b-a)^3}{24n^2} \cdot \max_{x\in[a,b]} |f''(x)|} \quad \text{et} \quad \boxed{|e_n^{(T)}| \leqslant \frac{(b-a)^3}{6n^2} \cdot \max_{x\in[a,b]} |f''(x)|}$$

Cela signifie que les méthodes des approximations à gauche et à droite sont exactes pour des fonctions constantes (dont la dérivée est nulle). Tandis que les méthodes du point médian et des trapèzes sont exactes pour des fonctions affines (dont la dérivée seconde est nulle).

Ingrédients pour la démonstration du théorème

Le développement de Taylor

On va utiliser le développement de Taylor d'ordre n en x_0 d'une fonction f dont les $n+1$ premières dérivées sont supposées continues. Le nombre ξ ci-dessous est entre x et x_0.

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}$$

Il est important de bien souligner le fait que ξ dépend de la valeur de x_0 et de x.

L'inégalité triangulaire

Cette inégalité permet d'écrire :

$$\Big|x_1 + x_2 + \cdots + x_n\Big| \leqslant |x_1| + |x_2| + \cdots + |x_n| \quad \text{ou} \quad \left|\sum_{i=1}^{n} x_i\right| \leqslant \sum_{i=1}^{n} |x_i|$$

Démonstration de la formule concernant la borne d'erreur pour la méthode des approximations à gauche

On peut commencer à estimer l'erreur en utilisant les définitions, les notations précédentes et l'inégalité triangulaire,

$$\begin{aligned} |e_n^{(G)}| &= \left| \int_a^b f(x)\,dx - G_n \right| = \left| \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,dx - \sum_{i=1}^{n} f(x_{i-1})\Delta x \right| \\ &= \left| \sum_{i=1}^{n} \left(\int_{x_{i-1}}^{x_i} f(x)\,dx - f(x_{i-1})\Delta x \right) \right| \leqslant \sum_{i=1}^{n} \left| \int_{x_{i-1}}^{x_i} f(x)\,dx - f(x_{i-1})\Delta x \right| \end{aligned}$$

Prenons F une primitive de f (F existe car f est continue). Par le théorème fondamental du calcul intégral, on a :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = F(x_i) - F(x_{i-1})$$

On utilise le développement de Taylor d'ordre 1 en x_{i-1} de la fonction F pour simplifier cette intégrale :

$$F(x) = F(x_{i-1}) + f(x_{i-1})(x - x_{i-1}) + \frac{f'(\xi_i)}{2}(x - x_{i-1})^2 \quad (\xi_i \text{ est entre } x_{i-1} \text{ et } x)$$

En évaluant ce développement en x_i, on trouve :

$$F(x_i) = F(x_{i-1}) + f(x_{i-1})\Delta x + \frac{f'(\xi_i^*)}{2}(\Delta x)^2 \quad (\xi_i^* \text{ est entre } x_{i-1} \text{ et } x_i)$$

Ainsi :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = F(x_i) - F(x_{i-1}) = f(x_{i-1})\Delta x + \frac{f'(\xi_i^*)}{2}(\Delta x)^2$$

Par conséquent, on a :

$$|e_n^{(G)}| \leqslant \sum_{i=1}^{n} \left| \int_{x_{i-1}}^{x_i} f(x)\,dx - f(x_{i-1})\Delta x \right| = \sum_{i=1}^{n} \left| f'(\xi_i^*) \cdot \frac{(\Delta x)^2}{2} \right|$$

De plus, on peut simplifier cette expression grâce à la dernière majoration suivante :

$$|f'(\xi_i^*)| \leqslant \max_{x\in[a,b]} |f'(x)| \quad \text{car } \xi_i^* \in [a,b]$$

D'où :

$$|e_n^{(G)}| \leqslant \sum_{i=1}^{n} \left| f'(\xi_i^*) \cdot \frac{(\Delta x)^2}{2} \right| = \sum_{i=1}^{n} |f'(\xi_i^*)| \cdot \frac{(\Delta x)^2}{2} \leqslant \sum_{i=1}^{n} \max_{x\in[a,b]} |f'(x)| \cdot \frac{(\Delta x)^2}{2}$$

En utilisant le fait que $\Delta x = \frac{b-a}{n}$, on obtient la majoration annoncée :

$$|e_n^{(G)}| \leqslant n \cdot \max_{x\in[a,b]} |f'(x)| \cdot \frac{(\Delta x)^2}{2} = \frac{(b-a)^2}{2n} \cdot \max_{x\in[a,b]} |f'(x)|$$

□

A propos de la démonstration de la formule concernant la borne d'erreur pour la méthode des approximations à droite

La démonstration est quasiment la même. On utilise bien sûr D_n au lieu de G_n et on devra utiliser le développement de Taylor d'ordre 1 en x_i de la fonction F (au lieu du développement en x_{i-1}).

Démonstration de la formule concernant la borne d'erreur pour la méthode du point médian

Notons $x_{i-\frac{1}{2}}$ le milieu de l'intervalle $[x_{i-1}, x_i]$. On peut commencer à estimer l'erreur en utilisant les définitions, les notations précédentes et l'inégalité triangulaire,

$$\begin{aligned} |e_n^{(M)}| &= \left| \int_a^b f(x)\,dx - M_n \right| = \left| \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,dx - \sum_{i=1}^{n} f(x_{i-\frac{1}{2}})\Delta x \right| \\ &= \left| \sum_{i=1}^{n} \left(\int_{x_{i-1}}^{x_i} f(x)\,dx - f(x_{i-\frac{1}{2}})\Delta x \right) \right| \leqslant \sum_{i=1}^{n} \left| \int_{x_{i-1}}^{x_i} f(x)\,dx - f(x_{i-\frac{1}{2}})\Delta x \right| \end{aligned}$$

Prenons F une primitive de f (F existe car f est continue). Par le théorème fondamental du calcul intégral, on a :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = F(x_i) - F(x_{i-1})$$

On utilise le développement de Taylor d'ordre 2 en $x_{i-\frac{1}{2}}$ de la fonction F pour simplifier cette intégrale. Il existe ξ_i entre $x_{i-\frac{1}{2}}$ et x tel que :

$$F(x) = F(x_{i-\frac{1}{2}}) + f(x_{i-\frac{1}{2}})(x - x_{i-\frac{1}{2}}) + \frac{f'(x_{i-\frac{1}{2}})}{2}(x - x_{i-\frac{1}{2}})^2 + \frac{f''(\xi_i)}{3!}(x - x_{i-\frac{1}{2}})^3$$

En évaluant ce développement en x_i, il existe ξ_i^* entre $x_{i-\frac{1}{2}}$ et x_i tel que :

$$F(x_i) = F(x_{i-\frac{1}{2}}) + f(x_{i-\frac{1}{2}}) \left(\frac{\Delta x}{2}\right) + \frac{f'(x_{i-\frac{1}{2}})}{2} \left(\frac{\Delta x}{2}\right)^2 + \frac{f''(\xi_i^*)}{3!} \left(\frac{\Delta x}{2}\right)^3$$

En évaluant ce développement en x_{i-1}, il existe ξ_i^{**} entre $x_{i-\frac{1}{2}}$ et x_{i-1} tel que :

$$F(x_{i-1}) = F(x_{i-\frac{1}{2}}) - f(x_{i-\frac{1}{2}}) \left(\frac{\Delta x}{2}\right) + \frac{f'(x_{i-\frac{1}{2}})}{2} \left(\frac{\Delta x}{2}\right)^2 - \frac{f''(\xi_i^{**})}{3!} \left(\frac{\Delta x}{2}\right)^3$$

Ainsi :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = F(x_i) - F(x_{i-1}) = f(x_{i-\frac{1}{2}})\Delta x + \frac{f''(\xi_i^*)}{48}(\Delta x)^3 + \frac{f''(\xi_i^{**})}{48}(\Delta x)^3$$

On peut maintenant simplifier $|e_n^{(M)}|$, puis on utilise l'inégalité triangulaire et les majorations suivantes

$$|f''(\xi_i^*)| \leqslant \max_{x\in[a,b]} |f''(x)| \quad \text{et} \quad |f''(\xi_i^{**})| \leqslant \max_{x\in[a,b]} |f''(x)|$$

afin de trouver :

$$\begin{aligned} |e_n^{(M)}| &\leqslant \sum_{i=1}^{n} \left| \int_{x_{i-1}}^{x_i} f(x)\,dx - f(x_{i-\frac{1}{2}})\Delta x \right| = \sum_{i=1}^{n} \left| f''(\xi_i^*) \cdot \frac{(\Delta x)^3}{48} + f''(\xi_i^{**}) \cdot \frac{(\Delta x)^3}{48} \right| \\ &\leqslant \sum_{i=1}^{n} \left(|f''(\xi_i^*)| \cdot \frac{(\Delta x)^3}{48} + |f''(\xi_i^{**})| \cdot \frac{(\Delta x)^3}{48} \right) \\ &\leqslant \sum_{i=1}^{n} \left(\max_{x\in[a,b]} |f''(x)| \cdot \frac{(\Delta x)^3}{48} + \max_{x\in[a,b]} |f''(x)| \cdot \frac{(\Delta x)^3}{48} \right) \\ &= n \cdot \max_{x\in[a,b]} |f''(x)| \cdot \frac{(\Delta x)^3}{24} = \frac{(b-a)^3}{24n^2} \cdot \max_{x\in[a,b]} |f''(x)| \qquad \text{car } \Delta x = \frac{b-a}{n} \end{aligned}$$

□

Démonstration de la formule concernant la borne d'erreur pour la méthode des trapèzes

Notons $x_{i-\frac{1}{2}}$ le milieu de l'intervalle $[x_{i-1}, x_i]$. On peut commencer à estimer l'erreur en utilisant les définitions, les notations précédentes et l'inégalité triangulaire,

$$\begin{aligned}
|e_n^{(T)}| &= \left| \int_a^b f(x)\,dx - T_n \right| = \left| \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)\,dx \; - \; \sum_{i=1}^{n} \frac{f(x_{i-1}) + f(x_i)}{2} \Delta x \right| \\
&= \left| \sum_{i=1}^{n} \left(\int_{x_{i-1}}^{x_i} f(x)\,dx \; - \; \frac{f(x_{i-1}) + f(x_i)}{2} \Delta x \right) \right| \\
&\leqslant \sum_{i=1}^{n} \left| \int_{x_{i-1}}^{x_i} f(x)\,dx \; - \; \Big(f(x_{i-1}) + f(x_i) \Big) \cdot \frac{\Delta x}{2} \right|
\end{aligned}$$

On va transformer chacun des deux termes qui se trouvent dans la valeur absolue.

1. Prenons F une primitive de f (F existe car f est continue). Par le théorème fondamental du calcul intégral, on a :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = F(x_i) - F(x_{i-1})$$

   En effectuant le même calcul que pour la méthode du point médian, on a :

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = F(x_i) - F(x_{i-1}) = f(x_{i-\frac{1}{2}})\Delta x + \frac{f''(\xi_i^*)}{48}(\Delta x)^3 + \frac{f''(\xi_i^{**})}{48}(\Delta x)^3$$

2. On utilise le développement de Taylor d'ordre 1 en $x_{i-\frac{1}{2}}$ de la fonction f. Ainsi, il existe ζ_i entre $x_{i-\frac{1}{2}}$ et x tel que :

$$f(x) = f(x_{i-\frac{1}{2}}) + f'(x_{i-\frac{1}{2}})(x - x_{i-\frac{1}{2}}) + \frac{f''(\zeta_i)}{2}(x - x_{i-\frac{1}{2}})^2$$

   En évaluant ce développement en x_i, il existe ζ_i^* entre $x_{i-\frac{1}{2}}$ et x_i tel que :

$$f(x_i) = f(x_{i-\frac{1}{2}}) + f'(x_{i-\frac{1}{2}}) \left(\frac{\Delta x}{2} \right) + \frac{f''(\zeta_i^*)}{2!} \left(\frac{\Delta x}{2} \right)^2$$

   En évaluant ce développement en x_{i-1}, il existe ζ_i^{**} entre $x_{i-\frac{1}{2}}$ et x_{i-1} tel que :

$$f(x_{i-1}) = f(x_{i-\frac{1}{2}}) - f'(x_{i-\frac{1}{2}}) \left(\frac{\Delta x}{2} \right) + \frac{f''(\zeta_i^{**})}{2!} \left(\frac{\Delta x}{2} \right)^2$$

   Ainsi :

$$\Big(f(x_{i-1}) + f(x_i) \Big) \cdot \frac{\Delta x}{2} = f(x_{i-\frac{1}{2}})\Delta x + \frac{f''(\zeta_i^*)}{16}(\Delta x)^3 + \frac{f''(\zeta_i^{**})}{16}(\Delta x)^3$$

Par conséquent, on a :

$$\begin{aligned} |e_n^{(T)}| &\leqslant \sum_{i=1}^{n} \left| \int_{x_{i-1}}^{x_i} f(x)\,dx \;-\; \Big(f(x_{i-1}) + f(x_i)\Big) \cdot \frac{\Delta x}{2} \right| \\ &= \sum_{i=1}^{n} \left| \frac{f''(\xi_i^*)}{48}(\Delta x)^3 + \frac{f''(\xi_i^{**})}{48}(\Delta x)^3 - \frac{f''(\zeta_i^*)}{16}(\Delta x)^3 - \frac{f''(\zeta_i^{**})}{16}(\Delta x)^3 \right| \end{aligned}$$

En utilisant l'inégalité triangulaire, on a :

$$|e_n^{(T)}| \leqslant \sum_{i=1}^{n} \left(\frac{|f''(\xi_i^*)|}{48}(\Delta x)^3 + \frac{|f''(\xi_i^{**})|}{48}(\Delta x)^3 + \frac{|f''(\zeta_i^*)|}{16}(\Delta x)^3 + \frac{|f''(\zeta_i^{**})|}{16}(\Delta x)^3 \right)$$

Grâce aux majorations suivantes

$$|f''(\xi_i^*)| \leqslant \max_{x\in[a,b]} |f''(x)| \quad \text{et} \quad |f''(\xi_i^{**})| \leqslant \max_{x\in[a,b]} |f''(x)|$$

$$|f''(\zeta_i^*)| \leqslant \max_{x\in[a,b]} |f''(x)| \quad \text{et} \quad |f''(\zeta_i^{**})| \leqslant \max_{x\in[a,b]} |f''(x)|$$

on obtient :

$$\begin{aligned} |e_n^{(T)}| &\leqslant \sum_{i=1}^{n} \max_{x\in[a,b]} |f''(x)| \cdot (\Delta x)^3 \cdot \left(\frac{1}{48} + \frac{1}{48} + \frac{1}{16} + \frac{1}{16} \right) \\ &= n \cdot \max_{x\in[a,b]} |f''(x)| \cdot (\Delta x)^3 \cdot \left(\frac{1}{24} + \frac{1}{8} \right) \\ &= n \cdot \max_{x\in[a,b]} |f''(x)| \cdot (\Delta x)^3 \cdot \left(\frac{1}{24} + \frac{3}{24} \right) \\ &= n \cdot \max_{x\in[a,b]} |f''(x)| \cdot (\Delta x)^3 \cdot \frac{4}{24} \end{aligned}$$

En utilisant le fait que $\Delta x = \frac{b-a}{n}$, on obtient la majoration de l'erreur annoncée :

$$|e_n^{(T)}| \leqslant n \cdot \max_{x\in[a,b]} |f''(x)| \cdot (\Delta x)^3 \cdot \frac{1}{6} = \frac{(b-a)^3}{6n^2} \cdot \max_{x\in[a,b]} |f''(x)|$$

□

14.5.6 La méthode de Simpson

On peut facilement montrer que la méthode des trapèzes satisfait la relation $T_n = \frac{G_n + D_n}{2}$. Il s'agit de la moyenne entre la méthode des approximations à gauche avec celle des approximations à droite.

On définit la méthode de Simpson par la formule $S_n = \frac{T_n + 2M_n}{3}$. C'est une moyenne pondérée entre la méthode des trapèzes et celle du point médian qui compte double.

Théorème Il existe une constante $C > 1$ telle que : $\boxed{|e_n^{(S)}| \leqslant \frac{(b-a)^5}{C \cdot n^4} \cdot \max_{x\in[a,b]} |f^{(4)}(x)|}$

La méthode de Simpson donne des résultats exacts pour tous les polynômes de degré $\leqslant 3$.

14.6 Formules de quadratures

Dans cette section, on va découvrir un nouveau point de vue qui permet de retrouver les méthodes d'intégrations numériques précédentes et de pouvoir généraliser ces méthodes dans le cas du calcul des intégrales de fonctions à deux variables.

14.6.1 Généralités

Soit Ω un domaine borné de $\mathbb{R}^n$ (dans ce cours, on se restreint à $n = 1$ ou $n = 2$). Soit $f : \Omega \to \mathbb{R};\ x = (x_1, x_2, \ldots, x_n) \mapsto f(x) = f(x_1, x_2, \ldots, x_n)$ une fonction continue de n variables à valeur réelle.

Le but des formules de quadratures est de pouvoir estimer les intégrales suivantes :

$$\int_\Omega f(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \cdots dx_n \ = \ \int_\Omega f(x)\, dx$$

Dans ce but, on choisit N points a_1, a_2, …, a_N de Ω auxquels on associe des poids $w_k \geqslant 0$ tels que :

$$\sum_{k=1}^{N} w_k = \mathrm{Vol}(\Omega) \quad \text{où} \quad \mathrm{Vol}(\Omega) \text{ est le volume de } \Omega$$

On obtient ainsi une formule de quadrature, notée $J(f)$ et définie par :

$$\boxed{J(f) = \sum_{k=1}^{N} w_k f(a_k)}$$

Théorème 1

Sous ces notations, on a la majoration suivante pour l'erreur commise E :

$$E = \left| \int_\Omega f(x)\, dx - J(f) \right| \ \leqslant \ V_\Omega(f) \cdot \mathrm{Vol}(\Omega) \quad \text{où} \quad \begin{array}{c} V_\Omega(f) = \sup_{x\in\Omega} f(x) - \inf_{y\in\Omega} f(y) \\ \text{est l'écart maximal de } f \text{ sur } \Omega \end{array}$$

Preuve du théorème 1

$$\begin{aligned} \text{On a : } \ E \ &= \ \left| \int_\Omega f(x)\, dx - J(f) \right| \ = \ \left| \frac{\sum_{k=1}^{N} w_k}{\mathrm{Vol}(\Omega)} \int_\Omega f(x)\, dx - \sum_{k=1}^{N} w_k f(a_k) \right| \\ &= \ \left| \sum_{k=1}^{N} \frac{w_k}{\mathrm{Vol}(\Omega)} \int_\Omega f(x)\, dx - \sum_{k=1}^{N} \frac{w_k}{\mathrm{Vol}(\Omega)} \int_\Omega f(a_k)\, dx \right| \\ &= \ \left| \sum_{k=1}^{N} \frac{w_k}{\mathrm{Vol}(\Omega)} \int_\Omega \Big(f(x) - f(a_k) \Big)\, dx \right| \ \leqslant \ \sum_{i=1}^{N} \frac{w_k}{\mathrm{Vol}(\Omega)} \int_\Omega |f(x) - f(a_k)|\, dx \end{aligned}$$

La dernière majoration provient de l'inégalité triangulaire et du fait que les poids w_k sont positifs ou nuls.

Or, par définition de $V_\Omega(f)$, on a $|f(x) - f(a_k)| \leqslant V_\Omega(f)$. Donc :

$$E \ \leqslant \ \sum_{k=1}^{N} \frac{w_k}{\mathrm{Vol}(\Omega)} \int_\Omega V_\Omega(f)\, dx \ = \ \int_\Omega V_\Omega(f)\, dx \ = \ V_\Omega(f) \cdot \mathrm{Vol}(\Omega)$$

□

La technique de la partition

On peut aussi faire une partition du domaine Ω en M domaines appelés Ω_i. Ces domaines Ω_i doivent alors satisfaire les conditions suivantes :

$$\text{a)} \quad \bigcup_{i=1}^{M} \Omega_i = \Omega \qquad \text{b)} \quad \text{Les } \Omega_i \text{ sont d'intérieurs disjoints}$$

Pour chaque domaine Ω_i, on choisit N_i points a_k^i et des poids associés $w_k^i \geqslant 0$ qui vérifient :

$$\sum_{k=1}^{N_i} w_k^i = \text{Vol}(\Omega_i)$$

En notant $J_i(f)$ la formule de quadrature associée sur Ω_i, on peut définir une formule de quadrature $J^*(f)$ sur Ω de la manière suivante :

$$\boxed{J^*(f) = \sum_{i=1}^{M} J_i(f) = \sum_{i=1}^{M} \left(\sum_{k=1}^{N_i} w_k^i f(a_k^i) \right)}$$

Bien que les notations sont plus complexes, cette façon de procéder est équivalente aux formules de quadrature décrites sur la page précédente. Il y a tout de même un subtile avantage.

Théorème 2

Sous ces notations, on a la majoration suivante pour l'erreur commise E^* :

$$E^* = \left| \int_\Omega f(x)\,dx - J^*(f) \right| \leqslant \max_i V_{\Omega_i}(f) \cdot \text{Vol}(\Omega)$$

Preuve du théorème 2

$$\begin{aligned}
\text{On a : } E^* &= \left| \int_\Omega f(x)\,dx - J^*(f) \right| = \left| \sum_{i=1}^{M} \int_{\Omega_i} f(x)\,dx - \sum_{i=1}^{M} J_i(f) \right| \\
&= \left| \sum_{i=1}^{M} \left(\int_{\Omega_i} f(x)\,dx - J_i(f) \right) \right| \leqslant \sum_{i=1}^{M} \left| \int_{\Omega_i} f(x)\,dx - J_i(f) \right| \\
&\overset{\text{Thm 1}}{\leqslant} \sum_{i=1}^{M} V_{\Omega_i}(f) \cdot \text{Vol}(\Omega_i) \leqslant \sum_{i=1}^{M} \max_i V_{\Omega_i}(f) \cdot \text{Vol}(\Omega_i) \\
&= \max_i V_{\Omega_i}(f) \cdot \underbrace{\sum_{i=1}^{M} \text{Vol}(\Omega_i)}_{=\text{Vol}(\Omega)} = \max_i V_{\Omega_i}(f) \cdot \text{Vol}(\Omega)
\end{aligned}$$

□

Moralité

Plus la partition du domaine Ω est fine, plus le terme $\max_i V_{\Omega_i}(f)$ deviendra petit, donc plus l'approximation sera bonne !

En effet, la fonction f étant continue, plus les domaines Ω_i seront petits, plus l'écart maximal de f sur Ω_i sera petit.

14.6.2 Applications aux intégrales à une dimension

Dans la section précédente, on a subdivisé l'intervalle $[a,b]$ en n intervalles $[x_{i-1},x_i]$ (avec $i\in\{1,\ldots,n\}$) de même largeur $\Delta x=\frac{b-a}{n}$ et avec $x_i=a+i\cdot\Delta x$ ($x_0=a$ et $x_n=b$).

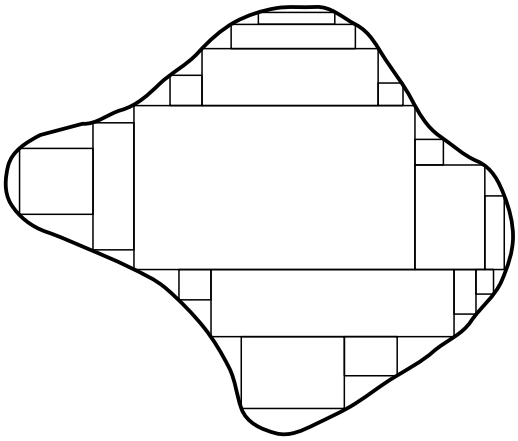
Sous les notations de cette section, on a $\Omega=[a,b]$ et $\Omega_i=[x_{i-1},x_i]$.

Cette technique permet de retrouver les méthodes déjà vues comme le montre le tableau de la page 161.

14.6.3 Applications aux intégrales à deux dimensions

Soit Ω un domaine borné du plan $\mathbb{R}^2$.

On peut toujours estimer Ω par une partition de rectangles. Pour chaque rectangle, on peut effectuer un changement de variables pour le ramener au carré universitaire $\widetilde{\Omega}=[0,1]^2$.



Ainsi, on se ramène à calculer des intégrales de la forme :

$$\int_{\widetilde{\Omega}} f(x,y)\,dxdy \;=\; \int_0^1\Big(\int_0^1 f(x,y)\,dx\Big)\,dy$$

On peut généraliser les méthodes précédentes en travaillant d'abord sur l'intégrale associée à la variable x, puis sur celle associée à y.

Regardons ce que donne cette généralisation pour la méthode du point médian si on découpe l'intervalle $[0,1]$ (sur l'axe des x et des y) en n morceaux.

Le cas $n=1$ donne la formule suivante :

$$\int_{\widetilde{\Omega}} f(x,y)\,dxdy \;=\; \int_0^1\Big(\int_0^1 f(x,y)\,dx\Big)\,dy \;\cong\; \int_0^1 f(\tfrac12,y)dy \;\cong\; f(\tfrac12,\tfrac12)$$

Le cas $n=2$ donne la formule suivante :

$$\begin{aligned}\int_{\widetilde{\Omega}} f(x,y)\,dxdy &= \int_0^1\Big(\int_0^1 f(x,y)\,dx\Big)\,dy \;\cong\; \int_0^1 \tfrac12\Big(f(\tfrac14,y)+f(\tfrac34,y)\Big)dy\\ &\cong \tfrac12\Big(\tfrac12\big(f(\tfrac14,\tfrac14)+f(\tfrac34,\tfrac14)\big)\Big)+\tfrac12\Big(\tfrac12\big(f(\tfrac14,\tfrac34)+f(\tfrac34,\tfrac34)\big)\Big)\\ &= \tfrac14\Big(f(\tfrac14,\tfrac14)+f(\tfrac34,\tfrac14)+f(\tfrac14,\tfrac34)+f(\tfrac34,\tfrac34)\Big)\end{aligned}$$

Le cas général donne la formule suivante :

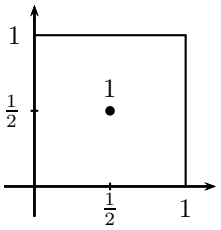
$$\int_{\widetilde{\Omega}} f(x,y)\,dxdy \;\cong\; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} f\Big(\tfrac{i-\frac12}{n},\tfrac{j-\frac12}{n}\Big)$$

On reconnaît des formules de quadrature !

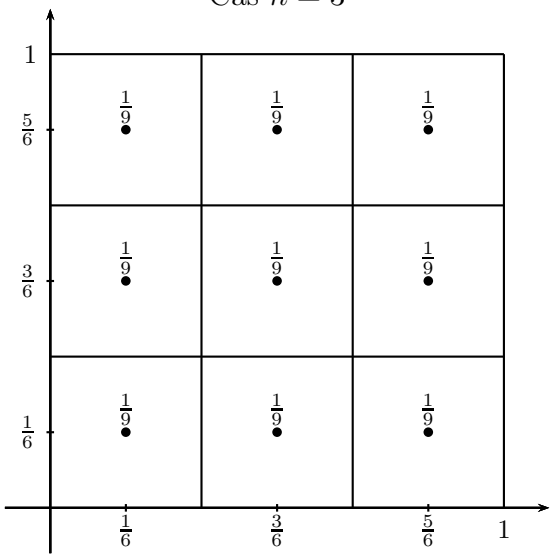
Vision schématique des poids

Pour la méthode généralisée à partir du point médian

Cas $n = 1$



Cas $n = 3$

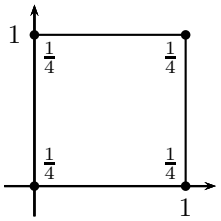


On a la formule de quadrature suivante si on subdivise l'intervalle $[0, 1]$ en n parties.

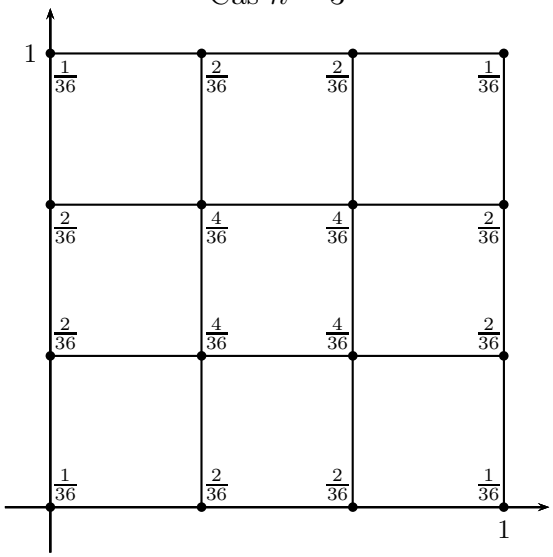
$$M_n^{(2)} \;=\; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} f\Big(\tfrac{i-\frac{1}{2}}{n}, \tfrac{j-\frac{1}{2}}{n}\Big) \;=\; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} f\Big(\tfrac{2i-1}{2n}, \tfrac{2j-1}{2n}\Big)$$

Pour la méthode généralisée à partir des trapèzes

Cas $n = 1$



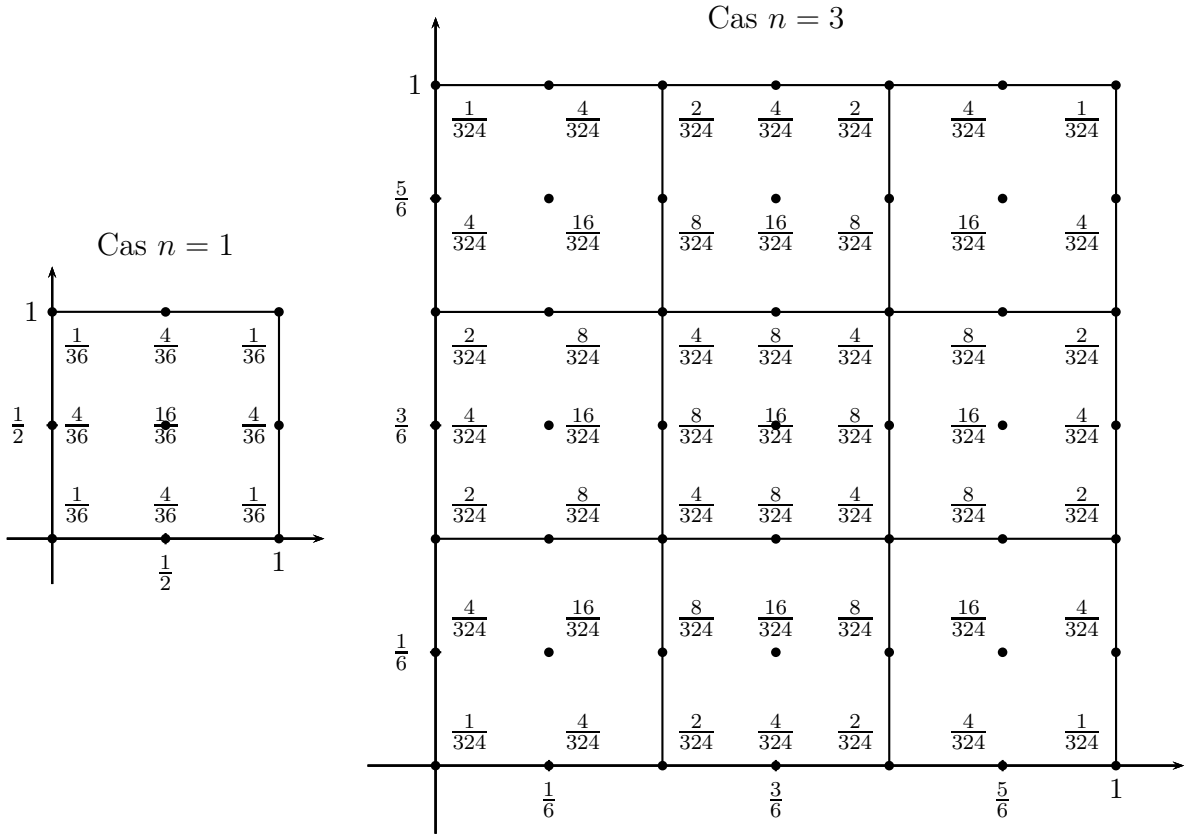
Cas $n = 3$



On a la formule de quadrature suivante si on subdivise l'intervalle $[0, 1]$ en n parties.

$$\begin{aligned} T_n^{(2)} \;=\; \frac{1}{4n^2}\Big(\;& f(0,0)+f(1,0)+f(0,1)+f(1,1) \;+\; 4\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} f\Big(\tfrac{i}{n}, \tfrac{j}{n}\Big) \\ & + 2\sum_{k=1}^{n-1}\Big(f\Big(\tfrac{k}{n},0\Big)+f\Big(\tfrac{k}{n},1\Big)+f\Big(0,\tfrac{k}{n}\Big)+f\Big(1,\tfrac{k}{n}\Big)\Big) \quad \Big) \end{aligned}$$

Pour la méthode généralisée à partir de Simpson



On a la formule de quadrature suivante si on subdivise l'intervalle $[0, 1]$ en n parties.

$$\begin{aligned} S_n^{(2)} \;=\; \frac{1}{36n^2}\Big(\;& f(0,0)+f(1,0)+f(0,1)+f(1,1) \\ & + 2\sum_{k=1}^{n-1}\Big(f\big(\tfrac{k}{n},0\big)+f\big(\tfrac{k}{n},1\big)+f\big(0,\tfrac{k}{n}\big)+f\big(1,\tfrac{k}{n}\big)\Big) \\ & + 4\sum_{k=1}^{n}\Big(f\big(\tfrac{2k-1}{2n},0\big)+f\big(\tfrac{2k-1}{2n},1\big)+f\big(0,\tfrac{2k-1}{2n}\big)+f\big(1,\tfrac{2k-1}{2n}\big)\Big) \\ & + 8\sum_{i=1}^{n-1}\sum_{j=1}^{n}\Big(f\big(\tfrac{i}{n},\tfrac{2j-1}{2n}\big)+f\big(\tfrac{2j-1}{2n},\tfrac{i}{n}\big)\Big) \\ & + 4\sum_{i=1}^{n-1}\sum_{j=1}^{n-1} f\big(\tfrac{i}{n},\tfrac{j}{n}\big) \;+\; 16\sum_{i=1}^{n}\sum_{j=1}^{n} f\big(\tfrac{2i-1}{2n},\tfrac{2j-1}{2n}\big) \quad \Big) \end{aligned}$$

Remarque. Cette généralisation est issue de la formule de quadrature de la méthode de Simpson en dimension 1 qui se trouve sur la page suivante ; malheureusement, on perd la relation que l'on avait avec les méthodes unidimensionnelles :

$$S_n^{(2)} \neq \tfrac{T_n^{(2)}+2M_n^{(2)}}{3}$$

Pire, $S_n^{(2)}$ ne plus s'exprimer comme une moyenne pondérée de $T_n^{(2)}$ et de $M_n^{(2)}$. En effet, $S_n^{(2)}$ a des poids qui n'apparaissent nullement dans $T_n^{(2)}$ et dans $M_n^{(2)}$ (contrairement au cas unidimensionnel).

méthode	sur $[0,1]$	sur Ω_i	sur Ω
approximations à gauche	$f(0)$	$J_i(f) = \Delta x \cdot f(x_{i-1})$	$J^*(f) = \Delta x \cdot \sum_{i=1}^{n} f(x_{i-1})$
approximations à droite	$f(1)$	$J_i(f) = \Delta x \cdot f(x_i)$	$J^*(f) = \Delta x \cdot \sum_{i=1}^{n} f(x_i)$
point médian	$f(\frac{1}{2})$	$J_i(f) = \Delta x \cdot f(x_{i-\frac{1}{2}})$	$J^*(f) = \Delta x \cdot \sum_{i=1}^{n} f(x_{i-\frac{1}{2}})$
des trapèzes	$f(\frac{1}{2})$	$J_i(f) = \frac{\Delta x}{2}\Big(f(x_{i-1}) + f(x_i)\Big)$	$J^*(f) = \frac{\Delta x}{2}\left(f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right)$
de Simpson	$\frac{f(0)}{6} + \frac{f(\frac{1}{2})}{6} + \frac{f(1)}{6}$	$J_i(f) = \frac{\Delta x}{6}\Big(f(x_{i-1}) + 4f(x_{i-\frac{1}{2}}) + f(x_i)\Big)$	$J^*(f) = \frac{\Delta x}{6}\left(f(x_0) + 4\sum_{i=1}^{n} f(x_{i-\frac{1}{2}}) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right)$