

Introduction à la théorie des systèmes de votes

Raphaël Goutmann & Louise Marsollier

12/01/2024

[TOC]

Qu'est ce que la cryptographie ?

La cryptographie est l'art de chiffrer des informations.

Elle a pour but principal de permettre le transfert d'une information d'une personne A à une personne B sans qu'une tierce personne ne puisse la lire et la comprendre. Elle est utilisée de nos jours pour sécuriser les communications sur internet, protégeant la vie privée des utilisateurs et évitant toute sorte de piratage.

Modes de chiffrements “intuitifs”

La cryptographie est apparue bien avant l'informatique (ex: tablette d'argile qui a été retrouvée en mésopotamie (-1500), ou la scyphale spartiate durant l'antiquité).

L'un des premiers systèmes de chiffrement de l'histoire a été conçu par César lui-même pour permettre des communications sécurisées avec ses armées. Lorsque l'empereur leur transmettait des messages, il était en effet essentiel qu'en cas d'interception par un ennemi, celui-ci ne puisse en exploiter des informations stratégiques qui pourraient compromettre l'empire romain.

Ce procédé, dit “chiffrement de César” est particulièrement simple à mettre en œuvre. Admettons que Bruno souhaite transmettre un message à Alice, disons le mot “bonjour”. Il commence par s'accorder avec Alice sur une clé secrète, correspondant à un entier k compris entre 1 et 26 puis décale chaque lettre composant son message de k rang dans l'alphabet. Si Alice et Bruno se s'accordent sur une clé $k = 2$, le message “bonjour” se chiffre alors en “dqqlqwt”. Pour déchiffrer ce message, il suffit de répéter le processus dans le sens inverse en décalant les lettres du message chiffré d'un rang $-k$.

Le chiffrement César présente toutefois une sécurité très faible. Il existe en effet 26 choix possibles de clés. Un tiers malveillant ayant intercepté le message et ayant connaissance du système utilisé pourrait alors tester toutes les possibilités

de clés et déchiffrer le message en un rien de temps. On dit alors que l'espace de clés (c'est-à-dire l'ensemble des clés possibles) est petit.

Un autre système de ce type est le chiffrement par substitution. On associe cette fois-ci à chaque lettre une autre lettre en laquelle elle sera transformée. La clé à communiquer est alors une table des correspondances entre les différentes lettres et le décryptage se fait en partant des lettres d'arrivées pour revenir sur les lettres de départ.

L'avantage de ce système, outre sa simplicité, est son espace de clés gigantesque. Lorsque nous construisons la clé, nous commençons par la lettre A. Nous avons alors 26 choix de correspondances possibles. Une fois choisi, nous avons 25 possibilités pour la lettre B, 24 pour la lettre C etc. Le nombre de choix possibles est alors de $26 \times 25 \times 24 \times \dots \times 2 \times 1 = 26!$ (26 "factorielle") soit près de 4×10^{26} clés. Avec un espace de clés si vaste, même les ordinateurs les plus puissants dont nous disposons actuellement seraient incapable de toutes les tester. La méthode d'attaque vue précédemment dans le cas du chiffrement César ne peut donc être utilisée.

Le chiffrement par substitution présente toutefois une faille majeure. Chaque lettre est toujours chiffrée de la même manière. Si je décide d'associer la lettre A à la lettre Y, chaque A du message initial apparaîtra sous forme de Y dans le message chiffré. Mais dans une langue donnée, toutes les lettres n'apparaissent pas selon la même fréquence. Le E, par exemple, est la lettre la plus souvent rencontrée dans la langue française, loin devant le S, le A etc. En analysant les fréquences des différentes lettres et en usant d'un peu de bon sens, nous pouvons donc déchiffrer le message.

Enigma et Turing

La machine Enigma a été inventée par l'ingénieur électricien allemand Arthur Scherbius, d'après un brevet de l'inventeur néerlandais Hugo Koch déposé en 1919. C'est une machine qui permet, à travers plusieurs réglages, de crypter des messages. Pour pouvoir décrypter le message, il faut disposer d'une autre machine enigma . Elle est largement utilisée par les allemands durant la seconde guerre mondiale et attire l'attention des services secrets français et polonais, qui parviennent à tordre certains de ses secrets. Pour autant, l'armée allemande réussit à complexifier encore davantage sa machine, en changeant de protocole : elle augmente notamment le nombre de rotors et donc de transpositions, rebouchant la faille identifiée auparavant par les chercheurs. Les américains vont à leur tour essayer de résoudre tous ces mystères en lançant l'opération "Ultra". Sept mille personnes s'attellent à cette tâche dans le manoir de Bletchley Park, parmi lesquelles le mathématicien Alan Turing. Ils font notamment construire d'énormes engins électromécaniques appelés "bombe" qui permettent ainsi de tester toutes les clés de chiffrement possible, en essayant en parallèle jusqu'à vingt mille configurations par seconde. Alan Turing parviendra finalement à "casser" Enigma, permettant aux alliés d'intercepter les communications

ennemis. Cette prouesse technologique leur donnera un avantage considérable, réduisant - d'après certains historiens - la guerre de près de 2 ans et sauvant par la même occasion des millions de vies. Alan Turing, condamné quelques années plus tard pour son homosexualité, sera retrouvé mort le 7 juin 1954 avec à ses côtés une pomme, probablement un suicide. Il laissera derrière lui une œuvre mathématique considérable à l'origine, entre autres, de l'ordinateur moderne et de l'intelligence artificielle.

Cryptographie asymétrique

Les systèmes de cryptographie que nous venons de voir nécessitent le partage d'une clé. Toute la sécurité du système repose donc sur ce partage qui ne peut être parfaitement invulnérable. Le problème est d'autant plus important sur internet où le partage d'une clé ne peut se faire physiquement et impose donc de partager la clé par réseau, les interlocuteurs s'exposent - de fait - au risque qu'un pirate intercepte la clé et puisse déchiffrer les futurs messages.

Une solution à ce problème est l'utilisation d'un chiffrement dit "asymétrique". Le principe de ce système repose sur l'utilisation d'un chiffrement "à sens unique".

Pour communiquer avec Bruno, Alice commence par générer deux clés : une clé publique qu'elle partagera librement à Bruno et une clé privée qu'elle gardera pour elle. Bruno chiffre ensuite son message avec la clé publique partagée par Alice à l'aide d'un chiffrement à sens unique. La particularité d'un tel chiffrement est que la clé pour chiffrer un message et celle pour le déchiffrer est différente. Ainsi, une fois le message chiffré avec la clé publique dont dispose Bruno, seule Alice disposant de sa clé privée pourra le déchiffrer. Bruno peut ainsi librement envoyer son message sur internet sans craindre d'être intercepté puisque seule la clé privée d'Alice (qu'elle a gardée secrète) permet de le déchiffrer. Si Alice souhaite à son tour envoyer un message à Bruno, il suffit de réitérer le processus dans l'autre sens : Bruno génère une clé publique et une clé privée, partage sa clé publique, Alice l'utilise pour chiffrer son message, l'envoie sur internet, Bruno l'intercepte et le déchiffre avec sa clé privée.

Bien que les processus mis en œuvre derrière de tels chiffrement soient assez complexes, il est facile d'en saisir intuitivement le fonctionnement.

Prenons le nombre 15. Il est relativement aisé d'en déterminer la factorisation en nombres premiers : 3×5 . De même, une fois les facteurs premiers trouvés, il est très facile de retrouver le nombre de départ (une simple multiplication suffit) : $3 \times 5 = 15$. Mais admettons que nous prenions des nombres plus grands, bien plus grands, gigantesques même avec plusieurs centaines de chiffres. La seule méthode dont nous disposons pour décomposer ce nombre en produit de facteurs premiers est de tester toutes les possibilités. Cette décomposition, facile pour de petits nombres, devient alors terriblement complexe. Même les ordinateurs les plus puissants dont nous disposions auraient besoin de plusieurs milliards d'années pour déterminer cette décomposition. Toutefois, disposant

des facteurs, il est toujours très facile de retrouver le nombre initial. Il suffit de calculer une simple multiplication, réalisable en quelques secondes par un ordinateur. Le processus est donc rapide dans un sens et lent dans l'autre. Il est “facile” de multiplier et “complexe” de décomposer. Et c'est là dessus que repose le chiffrement asymétrique. Le chiffrement à l'aide de la clé publique est un processus “facile” mais le déchiffrement de ce même message à l'aide de la clé publique est “complexe” à moins de disposer de la clé privée qui agit comme une “trappe” secrète pour retrouver le message initial. (voir ce lien pour plus de détails sur le chiffrement RSA, principale système de chiffrement asymétrique)

Annexe

Voici quelques ressources pour approfondir le sujet.

Chacune d'entre elle est précédée d'un certain nombre d'étoiles suivant sa complexité, allant d'une étoile (*) pour les plus faciles à trois étoiles (***) pour les plus complexes.

- (**) Pour en savoir d'avantage sur le chiffrement César et sa formalisation mathématique (avec en bonus une implémentation du système en Python).
Exo7Math — Cryptographie - partie 1 : chiffrement de César
- (**) Pour en savoir plus sur le chiffrement par substitution et un chiffrement semblable dit “de Vigenère” ainsi que leur formalisation mathématique. Exo7Math — Cryptographie - partie 2 : chiffrement de Vigenère
- (**) Pour un savoir plus sur le principe de cryptographie à clé publique / asymétrique et sa formalisation mathématique. Exo7Math — Cryptographie - partie 4 : cryptographie à clé publique
- (***) Pour comprendre en profondeur le fonctionnement du chiffrement RSA, l'un des principaux système de chiffrement asymétrique. RSA est particulièrement complexe, c'est pourquoi deux vidéos vous sont proposées : une présentation des outils arithmétiques utilisés dans la conception de RSA et le système en lui même.
 - Partie 1 (outils arithmétiques pour RSA) : Exo7Math — Cryptographie - partie 5 : arithmétique pour RSA
 - Partie 2 (chiffrement RSA en tant que tel) : Exo7Math — Cryptographie - partie 6 : chiffrement RSA
- (**) Pour comprendre le principe général de la Blockchain (à l'origine notamment des cryptomonnaies), largement basé sur la cryptographie. Maths Adultes — Blockchain : Comment ça marche ?
- (*) Pour en connaître davantage sur Alan Turing, le film *Imitation Game*. Sans doute l'un des meilleurs films de mathématiques jamais réalisé.
- (*) Pour découvrir le mystère qui entoure Satoshi Nakamoto, à l'origine d'une révolution mathématique et économique en ayant créé la Blockchain

et le Bitcoin. Sans doute l'un des meilleurs reportages Arte jamais réalisé.
Arte — Le mystère Satoshi : enquête sur l'inventeur du bitcoin

[TOC]

L'objectif ici est d'établir de jolies formules faisant intervenir ϕ , π et e , trois constantes fondamentales des mathématiques. Nous n'avons ici l'ambition ni d'être exhaustifs, ni même de suivre une quelconque structure. Les résultats sont présentés pour leur intérêt esthétique et pédagogique.

π , ϕ , et e se distinguent principalement des constantes généralement étudiées par leur irrationalité. Aucune d'entre elles ne peut être écrite sous forme d'une fraction de deux entiers. Il en résulte un développement décimal infini sans logique apparente.

π et e présentent par ailleurs la particularité d'être des nombres transcendants, c'est-à-dire qu'il ne sont la racine d'aucun polynôme à coefficients entiers.

Malgré cet apparent "chaos", elles présentent de jolies propriétés qui justifient leur étude.

Phi

Le nombre ϕ , aussi appelé nombre d'or, est une constante dont l'expression décimale commence par 1,618. ϕ est défini comme l'unique rapport a/b entre deux longueurs a et b telles que le rapport de la somme $a+b$ des deux longueurs sur la plus grande (a) soit égal à celui de la plus grande (a) sur la plus petite (b), ce qui s'écrit :

$$\frac{a+b}{a} = \frac{a}{b} = \phi$$

Ainsi, le nombre d'or est l'unique nombre réel positif tel que

$$\phi^2 = \phi + 1$$

La notation ϕ fait référence au sculpteur Phidias, concepteur du parténon.

Partant de l'équation caractéristique du nombre d'or

$$\phi^2 = \phi + 1$$

En réarrangeant les termes nous obtenons

$$\phi^2 - \phi - 1 = 0$$

Il s'agit là d'une équation polynomiale du second degré des plus classiques.

Nous calculons $\Delta = (-1)^2 - 4 \times (-1) = 5$

D'où

$$x_1 = \frac{1 + \sqrt{5}}{2}$$

$$x_2 = \frac{1 - \sqrt{5}}{2}$$

Étant donné que seul x_1 est positif, nous obtenons

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

Une manière très courante de représenter le nombre d'or.

Mais ϕ apparaît également et de manière plus surprenante dans l'évaluation d'expression au caractère "infini".

ϕ est par exemple égal à

$$\sqrt{1 + \sqrt{1 + \sqrt{1 + \dots}}}$$

Démonstration

Ce résultat, d'apparence complexe, est en réalité tout à fait trivial.

$$\text{Notons } x = \sqrt{1 + \sqrt{1 + \sqrt{1 + \dots}}}$$

Étant donné que le motif se répète indéfiniment, nous avons en particulier

$$x = \sqrt{1 + x}$$

d'où

$$x^2 = 1 + x$$

qui correspond à l'équation caractéristique du nombre d'or.

Ainsi

$$x = \sqrt{1 + \sqrt{1 + \sqrt{1 + \dots}}} = \phi$$

Une autre propriété remarquable du nombre d'or concerne son lien étroit avec la suite de Fibonacci.

Considérons la suite (F_n) telle que $F_0 = 1$, $F_1 = 1$ et $F_{n+2} = F_{n+1} + F_n$, autrement dit la suite dont les deux premiers termes sont égaux à 1 et dont chaque terme suivant est la somme des deux termes qui le précèdent. En voici les premiers termes

1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, ...

C'est la suite de *Fibonacci*.

Considérons alors $R_n = \frac{F_{n+1}}{F_n}$, le quotient des termes consécutifs de la suite au rang n . Cette suite R_n possède une limite, laquelle est le nombre d'or, ce qui peut s'exprimer mathématiquement comme

$$\lim_{n \rightarrow +\infty} R_n = \lim_{n \rightarrow +\infty} \frac{F_{n+1}}{F_n} = \phi$$

Démonstration

Rappelons dans un premier temps quelques éléments au sujet de la valeur absolue.

La valeur absolue d'un nombre correspond à sa valeur numérique sans tenir compte du signe.

Elle est notée $|x|$.

Par exemple

$$\begin{aligned} |-3| &= 3 \\ |3| &= 3 \end{aligned}$$

D'un point de vue géométrique, la valeur absolue correspond à la distance séparant un nombre de l'origine du repère.

Une propriété essentielle de la valeur absolue est la suivante. Puisque $(a - b) = -(b - a)$, nous avons $|a - b| = |b - a|$.

Notons dans un premier temps que.

$$\begin{aligned} \phi^2 &= \phi + 1 \\ \Leftrightarrow \phi &= 1 + \frac{1}{\phi} \end{aligned}$$

Par ailleurs, et d'après la définition de la suite de Fibonacci

$$R_n = \frac{F_{n+1}}{F_n} = \frac{F_n + F_{n-1}}{F_n} = \frac{F_n}{F_n} + \frac{F_{n-1}}{F_n} = 1 + \frac{F_{n-1}}{F_n}$$

Or

$$R_{n-1} = \frac{F_n}{F_{n-1}}$$

D'où

$$\frac{1}{R_{n-1}} = \frac{F_{n-1}}{F_n}$$

Et ainsi

$$R_n = 1 + \frac{F_{n-1}}{F_n} = 1 + \frac{1}{R_{n-1}}$$

Nous pouvons alors en déduire que

$$\begin{aligned} & |R_n - \phi| \\ &= \left| \left(1 + \frac{1}{R_{n-1}} \right) - \left(1 + \frac{1}{\phi} \right) \right| \\ &= \left| \frac{1}{R_{n-1}} - \frac{1}{\phi} \right| \\ &= \left| \frac{\phi - R_{n-1}}{\phi R_{n-1}} \right| \end{aligned}$$

Mais puisque ϕ est strictement positif, il en est de même de $\frac{1}{\phi}$ et il est alors possible de le sortir de la valeur absolue

$$= \frac{1}{\phi} \left| \frac{\phi - R_{n-1}}{R_{n-1}} \right|$$

Par ailleurs, R_{n-1} étant plus grand que 1 (quotient de deux termes consécutifs d'une suite croissante), nous avons l'inégalité suivante

$$\begin{aligned} &= \frac{1}{\phi} \left| \frac{\phi - R_{n-1}}{R_{n-1}} \right| \\ &\leq \frac{1}{\phi} |\phi - R_{n-1}| \end{aligned}$$

Que nous pouvons réécrire

$$= \frac{1}{\phi} \left| \frac{\phi - R_{n-1}}{R_{n-1}} \right|$$

$$\leq \frac{1}{\phi} |R_{n-1} - \phi|$$

Nous obtenons en fin de compte

$$|R_n - \phi| \leq \frac{1}{\phi} |R_{n-1} - \phi|$$

Mais nous pouvons ici remarquer que le terme $|R_{n-1} - \phi|$ correspond précisément au terme $|R_n - \phi|$ en $n - 1$

Il est alors possible de répéter le processus de la façon suivante

$$|R_n - \phi|$$

$$\leq \frac{1}{\phi} |R_{n-1} - \phi|$$

$$\leq \left(\frac{1}{\phi} \right)^2 |R_{n-2} - \phi| \leq$$

...

$$\leq \left(\frac{1}{\phi} \right)^{n-1} |R_1 - \phi|$$

Or puisque $0 < 1/\phi < 1$

$$\lim_{n \rightarrow +\infty} \left(\frac{1}{\phi} \right)^{n-1} = 0$$

Ce qui implique

$$\lim_{n \rightarrow +\infty} |R_n - \phi| = 0$$

Autrement dit, R_n est contraint en $+\infty$ à se confondre à ϕ d'où en fin de compte

$$\lim_{n \rightarrow +\infty} \frac{F_{n+1}}{F_n} = \phi$$

QED. ¹

E

Le nombre e , aussi appelé constante de Neper ou nombre exponentiel, est une constante dont l'expression décimale commence par 2,7182. e est défini comme la base du logarithme naturel, c'est à dire le nombre tel que

$$\ln(e) = 1$$

On doit la notation e au mathématicien suisse Euler. De nombreuses conjectures existent quand à l'origine de sa notation. e pourrait être un hommage à Euler ou encore e comme première lettre de exponentielle.

e apparaît régulièrement dans des expressions infinis particulièrement harmonieuses.

En 1737, Euler a obtenu le développement en fraction continue de e suivant

$$e = 2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{4 + \cfrac{1}{1 + \dots}}}}}}$$

Le schéma étant à partir de la deuxième fraction 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1 10, ... 2n, 1, 1, 2(n+1), ... à l'infini

De façon similaire, Euler a démontré que

$$e = 1 + \frac{1}{1} + \frac{1}{1 \times 2} + \frac{1}{1 \times 2 \times 3} + \dots = \sum_{n=0}^{+\infty} \frac{1}{n!}$$

C'est à partir de cette formule que Euler a démontré ce qui est considéré par bon nombre de mathématiciens comme *la plus belle formule des mathématiques*

$$e^{i\pi} + 1 = 0$$

Elle met en relation les quatres piliers des mathématiques que sont l'analyse (e), l'algèbre (i), la géométrie (π) et l'arithmétique (0 et 1).

¹Il est également tout à fait possible, et c'est en réalité le cas le plus fréquent dans la littérature mathématique, de définir ζ sur \mathbb{C} .

e fait également une apparition remarquable dans le calcul de la limite en $+\infty$ de $(1 + \frac{1}{n})^n$

Si l'on calcule successivement les termes de la suite nous obtenons

```

n = 1 -> 2.0
n = 2 -> 2.25
n = 3 -> 2.37037037037037
n = 4 -> 2.44140625
n = 5 -> 2.4883199999999994
n = 6 -> 2.5216263717421135
n = 7 -> 2.546499697040712
n = 8 -> 2.565784513950348
n = 9 -> 2.5811747917131984
n = 10 -> 2.5937424601000023
n = 11 -> 2.6041990118975287
n = 12 -> 2.613035290224676
n = 13 -> 2.6206008878857308
n = 14 -> 2.6271515563008685
n = 15 -> 2.6328787177279187
n = 16 -> 2.6379284973666
n = 17 -> 2.64241437518311
n = 18 -> 2.6464258210976865
n = 19 -> 2.650034326640442
n = 20 -> 2.653297705144422

```

La suite semble converger vers e . Il s'agit en réalité d'un théorème que nous pouvons exprimer de la façon suivante

$$\lim_{n \rightarrow +\infty} \left(1 + \frac{1}{n}\right)^n = e$$

Démonstration

Pour démontrer cette égalité il est tout d'abord important de remarquer que nous sommes en présence d'une forme indéterminée (1^∞) . Nous cherchons donc à lever l'indétermination.

Pour cela, réécrivons dans un premier temps

$$\left(1 + \frac{1}{n}\right)^n$$

en

$$e^{\ln\left(1 + \frac{1}{n}\right)^n}$$

Cette transformation est permise par la propriété suivante

$$e^{\ln x} = \ln e^x = x$$

De plus puisque $\ln a^n = n \ln a$ nous pouvons réécrire l'expression de la façon suivante

$$e^{n \ln(1 + \frac{1}{n})}$$

On pose alors $N = \frac{1}{n}$ de sorte que

$$\begin{aligned} & \lim_{n \rightarrow +\infty} \left(1 + \frac{1}{n}\right)^n \\ &= \lim_{n \rightarrow +\infty} e^{n \ln(1 + \frac{1}{n})} \\ &= \lim_{N \rightarrow 0} e^{\frac{\ln(1+N)}{N}} \end{aligned}$$

$\frac{\ln(1+N)}{N}$ est toujours une forme indéterminée. Or puisque $\ln 1 = 0$ nous pouvons écrire

$$\begin{aligned} & \lim_{N \rightarrow 0} \frac{\ln(1+N)}{N} \\ &= \lim_{N \rightarrow 0} \frac{\ln(1+N) - \ln(1)}{N} \end{aligned}$$

Qui correspond alors précisément à l'expression de la dérivée de la fonction $\ln(1+x)$ en $x = 0$.

Calculons donc cette dérivée

$$[\ln(1+x)]' = \frac{1}{1+x}$$

qui vaut 1 en $x = 0$.

Dès lors

$$\lim_{N \rightarrow 0} \frac{\ln(1+N)}{N} = 1$$

Ainsi

$$\lim_{n \rightarrow +\infty} \left(1 + \frac{1}{n}\right)^n$$

$$\begin{aligned}
&= \lim_{N \rightarrow 0} e^{\frac{\ln(1+N)}{N}} \\
&= e^1 \\
&= e
\end{aligned}$$

QED.

Pi

π est un nombre dont la valeur approchée est 3,1415. Il est défini comme rapport constant de la circonférence d'un cercle à son diamètre.

Depuis le XVIIe siècle, on le représente par la lettre grecque π , première du mot périmètre. On parle également de "constante d'Archimède".

C'est une constante fondamentale en mathématiques et plus généralement en sciences car, au-delà de la géométrie, elle apparaît dans un très grand nombre de formules.

Une première méthode "naïve" pour calculer π est dûe au savant Archimède.

Considérons un cercle de rayon 1.

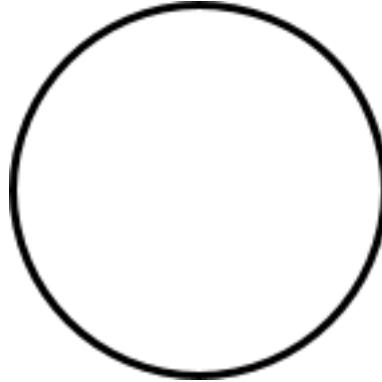


Figure 1: Illustration d'un cercle de rayon $r=1$

Nous noterons P le périmètre de ce cercle.

Par définition de π

$$P = 2 \times \pi \times R$$

Or dans notre cas $R = 1$ d'où

$$P = 2 \times \pi$$

$$\pi = \frac{P}{2}$$

Ainsi, il suffit de connaître le périmètre de notre cercle pour en déduire π .

Pour cela, traçons deux carrés A et B sur notre figure, respectivement inscrit et circonscrit.

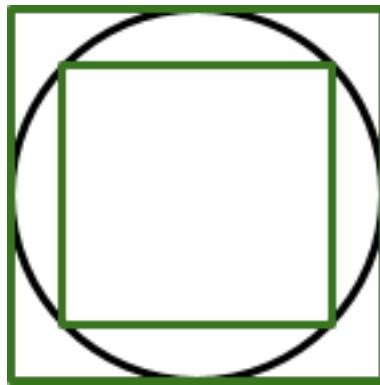


Figure 2: Illustration d'un cercle de rayon $r=1$ encadré par deux carrés respectivement inscrit et circonscrit

De ces carrés, il est très simple de déterminer les périmètres P_A et P_B .

Ces deux périmètres encadrent le périmètre du cercle, de sorte que

$$P_A \leq P \leq P_B$$

Et puisque $\pi = \frac{P}{2}$, nous obtenons un encadrement de π

$$\frac{P_A}{2} \leq \pi \leq \frac{P_B}{2}$$

Pour améliorer la précision de l'encadrement, il est alors possible d'utiliser des polygones à davantage de côtés tels que des pentagones, hexagones etc.

Voir une démonstration interactive.

π peut également être approximé de manière plus calculatoire.

Il apparaît par exemple dans le problème de Bâle qui consiste à calculer la valeur de la série

$$\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \dots$$

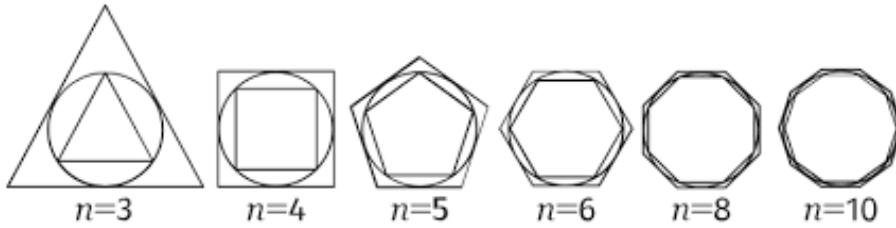


Figure 3: Illustration de la méthode d'achimète en différents polygones

$$= \sum_{n=1}^{+\infty} \frac{1}{n^2}$$

Autrement dit la somme de l'inverse des carrés.

Leonhard Euler répond à cette question en 1735, établissant que cette série vaut précisément $\frac{\pi^2}{6}$

De fait, il donne la première valeur non-triviale de ce qui sera appelée par la suite la fonction zéta de Riemann, au cœur du plus grand problème des mathématiques modernes.

π apparaît également de manière plus surprenante dans certains phénomènes physiques.

Imaginons le problème suivant.

Nous disposons d'une petite aiguille et d'une feuille de papier sur laquelle nous avons tracé des droites à intervalles réguliers.

Supposons que l'aiguille soit de longueur l et que les droites soient séparées d'une distance d ($d \geq l$).

Si l'on décide de lancer l'aiguille sur la feuille, quelle est la probabilité que l'aiguille se place dans une position telle qu'elle coupe l'une des droites ?

Le mathématicien français Georges Louis Lelecrc, Comte de Buffon en a donné la réponse en 1777, établissant que cette probabilité est exactement égale à

$$p = \frac{2}{\pi} \frac{l}{d}$$

Ce qui donne pour $l = d$

$$p = \frac{2}{\pi}$$

Ce résultat signifie que l'on peut obtenir des valeurs approchées de π par l'expérience.

Implémentée sur ordinateur, cette méthode constitue un excellent exemple d'algorithmes randomisé, utilisant l'aléatoire pour déterminer une valeur précise, en l'occurrence π .

Voir une démonstration interactive.

Une autre formule, et par ailleurs l'une des premières, faisant intervenir π a été découverte par le mathématicien indien Madhava au 14-ème / 15-ème siècle avant d'être redécouverte près de 200 ans plus tard en 1673 par le mathématicien et philosophe allemand Leibniz qui en porte désormais le nom. Elle s'énonce comme suit

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots = \sum_{n=0}^{+\infty} \frac{(-1)^n}{2n+1}$$

Autrement dit, $\pi/4$ est égal à la somme des inverses des nombres impaires en faisant alterner les signes.

À noter que cette série converge extrêmement lentement. Pour calculer π avec une précision de 6 décimales, il faut près de deux millions d'itérations.

Démonstration

Quelques prérequis sont nécessaires pour pouvoir comprendre cette démonstration.

Notion de primitive

La primitive est en quelque sorte “l'inverse” de la dérivée.

Soit f une fonction, une primitive de f notée F est une fonction telle que

$$F'(x) = f(x)$$

Par exemple, une primitive de $f(x) = x^2$ pourrait être

$$\frac{x^3}{3}$$

car

$$\left[\frac{x^3}{3} \right]' = x^2$$

Notion d'intégrale

Soit f une fonction.

L'intégrale de a à b de la fonction f correspond à l'aire sous la courbe représentative de f entre a et b .

Elle est notée

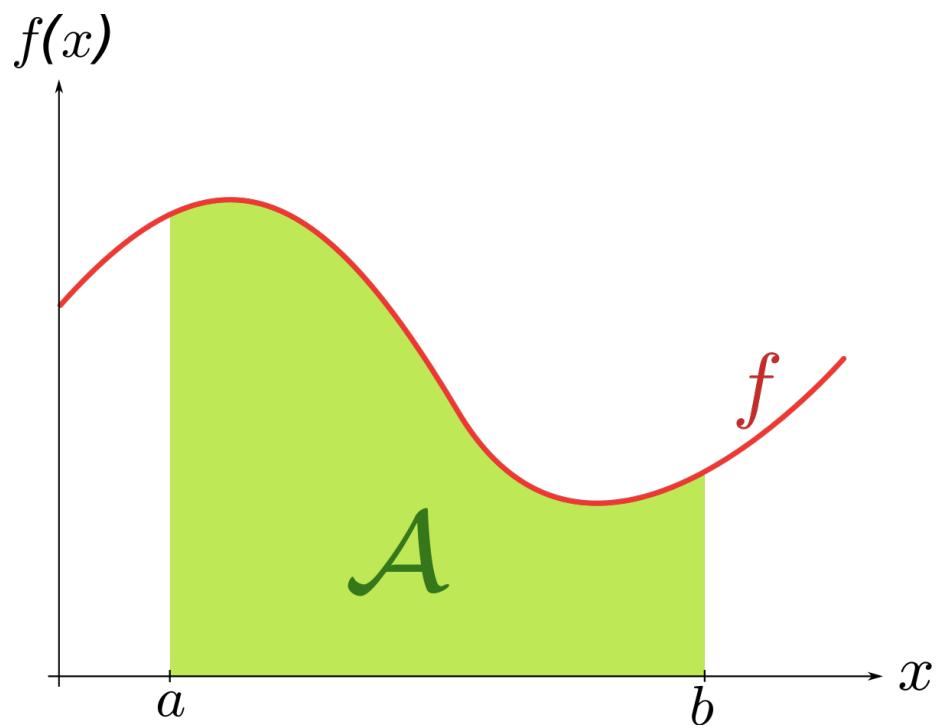


Figure 4: Représentation de l'air sous la courbe d'une fonction f entre deux points a et b

$$\int_a^b f(x)dx$$

qui doit être lu “intégrale de a à b de $f(x)$ dx ”.

Cette intégrale est alors égale, d’après le théorème fondamental de l’analyse, à

$$F(b) - F(a)$$

où F est une primitive de f .

Par exemple, l’aire entre 0 et 1 de la fonction $f(x) = x^2$ est

$$\int_0^1 x^2 dx = \frac{1^3}{3} - \frac{0^3}{3} = \frac{1}{3}$$

Avant de s’engager dans la démonstration, nous allons démontrer un théorème intermédiaire (un lemme) dont nous aurons besoin par la suite.

Ce lemme annonce que, pour tout $n \in N$ et $t \in R$ nous avons l’égalité suivante

$$\frac{1}{1+t^2} = 1 - t^2 + t^4 - \dots + t^{4n} - \frac{t^{4n+2}}{1+t^2}$$

Pour démontrer cela, nous allons partir d’une égalité bien connue concernant les suites géométriques.

Soit (u_n) une suite géométrique de premier terme $u_0 = 1$ et de raison q .

Nous savons que la somme des n premiers termes de la suite vaut

$$S = 1 + q + q^2 + \dots + q^n = \frac{1 - q^{n+1}}{1 - q}$$

De même la somme des $2n$ premiers termes de la suite vaut

$$S = 1 + q + q^2 + \dots + q^{2n} = \frac{1 - q^{2n+1}}{1 - q}$$

Posons alors $q = -t^2$.

Nous avons

$$1 + (-t^2) + (-t^2)^2 + (-t^2)^3 + \dots + (-t^2)^{2n} = \frac{1 - (-t^2)^{2n+1}}{1 - (-t^2)}$$

Que nous pouvons simplifier en

$$1 - t^2 + t^4 - t^6 + \dots + t^{4n} = \frac{1 + t^{4n+2}}{1 - t^2}$$

D'où, en décomposant la fraction de droite

$$1 - t^2 + t^4 - t^6 + \dots + t^{4n} = \frac{1}{1 - t^2} + \frac{t^{4n+2}}{1 - t^2}$$

Et ainsi

$$\frac{1}{1 - t^2} = 1 - t^2 + t^4 - t^6 + \dots + t^{4n} - \frac{t^{4n+2}}{1 - t^2}$$

Ce qu'il fallait démontrer.

Attaquons nous maintenant au résultat en lui-même.

Nous avons donc d'après le lemme précédent :

$$\frac{1}{1 - t^2} = 1 - t^2 + t^4 - t^6 + \dots + t^{4n} - \frac{t^{4n+2}}{1 - t^2}$$

Considérons un alors réel x tel que $0 \leq x \leq 1$.

Étant donné que les deux membres sont égaux, il en est de même de leurs intégrales respectives entre 0 et x . De sorte que

$$\begin{aligned} \int_0^x \frac{1}{1 - t^2} dt &= \int_0^x \left(1 - t^2 + t^4 - t^6 + \dots + t^{4n} - \frac{t^{4n+2}}{1 - t^2} \right) dt \\ &= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots + \frac{x^{4n+1}}{4n+1} - R_n(x) \end{aligned}$$

où

$$R_n(x) = \int_0^x \frac{t^{4n+2}}{1 + t^2} dt$$

Or puisque $t^2 \geq 0$ nous avons

$$1 \leq 1 + t^2$$

De sorte que

$$0 \leq R_n(x) \leq \int_0^x t^{4n+2} dt$$

d'où

$$0 \leq R_n(x) \leq \frac{x^{4n+3}}{4n+3}$$

Mais puisque $0 \leq x \leq 1$ il est clair que

$$\frac{x^{4n+3}}{4n+3} \leq \frac{1}{4n+3}$$

D'où

$$\lim_{n \rightarrow +\infty} \frac{1}{4n+3} = 0$$

implique que

$$\lim_{n \rightarrow +\infty} R_n(x) = 0$$

Et nous obtenons ainsi

$$\int_0^x \frac{1}{1+t^2} dt = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots$$

Mais puisque la dérivée de $\arctan(t)$ est $\frac{1}{1+t^2}$ nous avons par définition d'une primitive

$$\int_0^x \frac{1}{1+t^2} dt = \arctan(x) - \arctan(0) = \arctan(x)$$

dès lors

$$\arctan(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots$$

Et en prenant $x = 1$:

$$\arctan(1) = 1 - \frac{1^3}{3} + \frac{1^5}{5} - \frac{1^7}{7} + \dots$$

Autrement dit

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots$$

QED.

Annexe

Voici quelques ressources pour approfondir le sujet.

Chacune d'entre elle est précédée d'un certain nombre d'étoiles suivant sa complexité, allant d'une étoile (*) pour les plus faciles à trois étoiles (***) pour les plus complexes.

- (**) Une vidéo qui démystifie le nombre d'or en étudiant son histoire. Aux origines du nombre d'or - El Jj
- (*) Un rapide tour d'horizon des propriétés les plus remarquables de π . Carnets de voyages du nombre pi — Mickaël Launay
- (**) Un autre rapide tour d'horizon des propriétés les plus remarquables de π . Pourquoi pi est-il si fou ? — Science4All
- (***) La page Wikipédia consacrée à e . e (nombre) — Wikipedia
- (***) La page Wikipedia consacrée au nombre π Pi — Wikipedia
- (***) La page Wikipedia consacrée à ϕ Nombre d'or — Wikipedia

[TOC]

“La mathématique est l’art de donner le même nom à des choses différentes.” — Henri Poincaré

Notations

Nous adopterons, dans cet atelier, quelques notations relatives à la logique mathématique.

- \Rightarrow signifie “implique”
- \forall signifie “pour tout” ou “quelque soit”
- \exists signifie “il existe”
- $\exists!$ signifie “il existe un unique”
- \vee signifie “ou”
- \wedge signifie “et”
- \neg signifie “non”

Par ailleurs, on appellera *assertion* une affirmation soit *vraie*, soit *fausse*.

En notant, par exemple, A l'assertion “Socrate est un homme”, B l'assertion “Tous les hommes sont mortels” et C l'assertion “Socrate est mortel” on a :

$$(A \wedge B) \Rightarrow C$$

Axiomatisation des mathématiques

Structure des mathématiques

Les mathématiques occupent une place particulière au sein de la grande famille des sciences.

Une science a toujours pour objet la connaissance. Les mathématiques se distinguent des autres sciences dans leur manière d'accéder à cette connaissance.

Une science dite *empirique* (la biologie, la physique, la médecine etc.) recherche la connaissance dans l'expérience. Il s'agit de construire des *modèles* qui puissent prévoir au mieux les phénomènes réels. Pour cela, le scientifique émet des hypothèses quant aux lois qui régissent ces phénomènes et construit des expériences pour tester ces hypothèses. Deux possibilités se présentent alors : les résultats de l'expérience peuvent confirmer les prédictions, auquel cas l'hypothèse est renforcée (mais non prouvée), ou bien l'expérience contredit l'hypothèse qui doit donc être ajustée. Ainsi, la notion de *preuve* ou de *vérité absolue* dans les sciences empiriques n'a pas de sens. Une théorie ne peut être que plus vraisemblable qu'une autre.

Les mathématiques, *a contrario*, reposent sur l'idée d'une vérité absolue. La vérité mathématique est accessible par la preuve, qui consiste en une succession logique d'étapes menant à une conclusion *irréfutable*. Plus précisément, il s'agit de combiner, selon les lois de la logique, des résultats vrais pour aboutir à d'autres résultats vrai. Une question se pose alors : quel est le point de départ ? En effet, si les démonstrations consistent en une combinaison logique de résultats, arrive inévitablement un moment où le mathématicien doit admettre un résultat, une base, à partir de laquelle construire le reste. Ces résultats *admis* comme constituant la brique fondamentale de l'édifice mathématique sont appelés *axiomes*. Les mathématiques peuvent alors être vue comme une pyramide reposant sur un certain nombre d'axiomes desquels sont déduits, par inférence logique, théorèmes, lemmes, propositions et corollaires.

L'enjeu consiste alors à déterminer le “meilleur” système d'axiomes, le meilleur ensemble d'axiomes, pour établir une base solide à l'ensemble des mathématiques.

À noter ici que la démarche d'axiomatisation ne cherche pas à donner un sens aux objets manipulés. Elle distingue en effet la *nature* des objets leurs *propriétés*. Les mathématiques sont ici perçus comme un jeu logique de symboles, les objets définis n'ayant de sens que celui que nous leur donnons.

Axiomes d'Euclide

Proposés dans le livre 1 des *Éléments*, les axiomes d'Euclide constituent le premier système d'axiomes de l'histoire des mathématiques. Euclide se basera par la suite dessus pour construire l'ensemble de sa théorie géométrique.

Les axiomes d'Euclide sont au nombre de cinq :

1. Deux points définissent un et un seul segment.
2. Tout segment entre deux points peut être prolongé en une et une seule droite.
3. Un point et une longueur définissent un et un seul cercle.
4. Tous les angles droits sont égaux.
5. Par un point, il passe toujours une et une seule droite parallèle à une droite donnée.

Ces cinq propositions, évidente au premier abord, suffisent à construire l'ensemble de la théorie géométrique classique.

Le dernier postulat (dit postulat des parallèles) a fait l'objet de nombreux débats dans au sein de la communauté mathématiques. Les mathématiciens se sont en effet longtemps interrogé sur le caractère *axiomatique* de ce dernier : n'est-il pas possible de le démontrer à partir des autres axiomes ? Auquel cas il s'agirait d'un théorème et non d'un axiome. Cette question a donné naissance aux géométries non-euclidiennes, des géométries dont le système d'axiomes n'admet pas le postulat des parallèles (nous aurons l'occasion d'explorer davantage ce sujet dans un prochain atelier). De cette manière les mathématicien ont pu prouver qu'il s'agissait bien d'un axiome et non d'un théorème.

Axiomes de Peano

La géométrie euclidienne, aussi élégante soit-elle, montre rapidement certaines faiblesses. Limitée au cadre de la géométrie, elle ne permet pas de formaliser les nouvelles branches des mathématiques qui se développent alors, que ce soit l'analyse, l'algèbre ou encore l'arithmétique.

Dans cette ambition d'élargir le champ de l'axiomatisation, le mathématicien italien Giuseppe Peano répertoria, en 1889, les propriétés structurelles de l'ensemble des entiers naturels \mathbb{N} afin de donner une construction axiomatique de l'arithmétique.

L'arithmétique de Peano repose sur deux idées fondamentales : l'objet 0 et le concept de successeur. Les entiers naturels sont en effet définis par ceci qu'ils *succèdent* tous (à l'exception de 0) à un autre entier naturel. Une fois le 0 défini, ils peuvent ainsi tous être décrit par une relation de succession par rapport à un précédent entier.

Les axiomes de l'arithmétique de Peano sont les suivants :

- $\forall x, \neg(S(x) = 0)$ (aucun entier ne possède 0 pour successeur ou, autrement dit, 0 n'a pas de prédécesseur)
- $\forall x, \exists y, \neg(x = 0) \Rightarrow (S(y) = x)$ (tout entier non nul est le successeur d'un autre entier)
- $\forall x, \forall y, (S(y) = S(x) \Rightarrow x = y)$ (deux entiers de même successeur sont égaux)

Une fois définis les objets de l'arithmétique de Peano (les entiers), il s'agit de définir des opérations sur ces derniers, là encore à partir d'axiomes. L'addition dans un premier temps :

- $\forall x, (x + 0 = x)$ (0 est un élément *neutre* pour l'addition)
- $\forall x, \forall y, (x + S(y) = S(x + y))$ (l'addition d'un successeur est le successeur de l'addition)

À noter ici que l'addition est définie de façon récursive. Additionner deux entiers revient à dérouler une procédure aboutissant *in fine* au résultat. De la même manière, il est possible de définir la multiplication à partir de l'addition et de façon récursive :

- $\forall x, (x \times 0 = 0)$ (0 est un élément *absorbant* pour la multiplication)
- $\forall x, \forall y, (x \times S(y) = x \times y + x)$ (multiplier x par un successeur revient à ajouter x à la multiplication)

Le dernier axiome, enfin, et sans doute le plus important, est celui permettant de raisonner par récurrence. Il se définit comme suit :

- $(\phi(0) \wedge (\forall x, \phi(x) \Rightarrow \phi(S(x)))) \Rightarrow \forall x, \phi(x)$

Autrement dit, si une propriété est vérifiée en 0 et que sa vérité pour un certain rang *implique* celle du successeur, alors elle est vraie pour tout entier naturel.

Combinés aux axiomes de la logique, il est alors possible, à partir de ces quelques axiomes, de réécrire (non sans peine néanmoins) toute l'arithmétique de façon rigoureuse.

Démontrons, pour illustrer ces notions, l'égalité $1 + 1 = 2$ ou, selon les notations de l'arithmétique de Peano :

$$S(0) + S(0) = S(S(0))$$

D'après l'axiome n°5 (qui définit l'addition par récurrence) :

$$S(0) + S(0) = S(S(0) + 0)$$

Puis d'après axiome n°4 (qui définit le 0 comme élément neutre de l'addition) :

$$S(S(0) + 0) = S(S(0))$$

Enfin, par transitivité de l'égalité (qui est un axiome de la logique) :

$$S(0) + S(0) = S(S(0))$$

Ainsi, l'arithmétique de Peano permet de décrire l'ensemble de l'arithmétique dans un système d'une étonnante simplicité. Cependant, il présente la même limite soulevée dans le cadre des axiomes d'Euclide : il se limite à l'étude de l'arithmétique.

C'est dans ce contexte que naît au sein de la communauté mathématique du XXe siècle la volonté de construire un système d'axiomes universel, pouvant établir une base solide à l'ensemble des mathématiques.

Axiomes ZFC

ZFC, nommé ainsi en hommage aux mathématiciens Ernst Zermelo et Abraham Fraenkel qui en sont à l'origine, est un système d'axiomes ayant pour ambition de décrire toutes les mathématiques. En particulier, ZFC contient l'arithmétique de Peano et la géométrie euclidienne, mais permet également de décrire l'analyse, les probabilités etc.

Développé au XXe siècle, ZFC a été le point de départ des travaux du groupe Bourbaki, décidé à reconstruire toutes les mathématiques sur des fondements nouveaux, donnant naissance aux mathématiques "modernes".

Le système ZFC repose entièrement sur le concept d'ensemble. Tout objet mathématique se ramène, dans ZFC, à un ensemble. De même, toute opération sur ces objets, se rapport à des opérations ensemblistes d'inclusion, d'appartenance, d'union et d'intersection.

ZFC est aujourd'hui admis, grâce aux travaux de Bourbaki, comme le système d'axiomes de référence.

Théorèmes de Gödel

Complétude et cohérence d'un système

Mais ZFC est-il parfait ? Sommes-nous réellement parvenus à établir des bases inébranlables à toutes les mathématiques ?

Pour répondre à cette question, définissons dans un premier temps les caractéristiques nécessaires à un système d'axiomes *parfait*.

La première caractéristique est la *complétude* du système. Étant donné une certaine affirmation A formulée dans le système en question, il doit être possible de démontrer A (A est alors dite *démontrable*) ou de réfuter A , c'est à dire de démontrer $\neg A$ (A est alors dite *réfutable*) au sein du système d'axiomes. Aucune assertion ne doit être *indécidable*, c'est à dire ni démontrable, ni réfutable. Un système d'axiomes *parfait* doit être en capacité de démontrer ou de réfuter toutes les assertions formulées dans son langage.

La seconde caractéristique est la *cohérence* du système. Un système est *incohérent* lorsqu'une proposition peut à la fois être démontrée et réfutée, qu'elle

peut être à la fois *vraie* et *fausse*. Un tel système est nécessairement inconsistant et ne peut donc être utilisé pour établir une base solide à quelque théorie que ce soit. *A contrario* un système est dit cohérent si toute affirmation *A* est soit *vraie*, soit *fausse* mais jamais les deux à la fois.

Intuitivement, il semble que l'incohérence d'un système apparaît clairement dans son système d'axiomes. Mais de subtiles paradoxes peuvent rendre incohérente une théorie pourtant joliment construite. L'ancêtre de ZFC, une théorie axiomatique générale également basée sur des ensembles, s'est par exemple vue entièrement détruite par un paradoxe soulevé par Bertrand Russel. En effet, étant donné une propriété *P*, ce système autorisait de construire l'ensemble *E* de tous les objets vérifiant cette propriété *P*. Il est par exemple possible de définir \mathbb{R}^+ comme l'ensemble réels supérieurs ou égaux à 0. Mais un problème surgit lorsque l'on considère des ensembles d'ensembles et en particulier "l'ensemble des ensembles qui ne se contiennent pas eux mêmes". *E* appartient-il à lui-même ? Si *E* n'appartient pas à lui-même, alors il respecte la propriété et doit donc appartenir à lui-même. De la même manière, si *E* appartient à lui-même, alors il doit être supprimé puisqu'il ne vérifie plus la propriété. Autrement dit :

$$E \in E \Leftrightarrow E \notin E$$

Une contradiction. Ce système d'axiomes était donc *incohérent*.

Ainsi, le système d'axiomes parfait se doit d'être à la fois *cohérent*, *complet* et capable de décrire l'ensemble des mathématiques.

Théorèmes de Gödel

Kurt Gödel, mathématicien et logicien autrichien, publie ainsi en 1931 dans son article *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme* (« Sur les propositions formellement indécidables des Principia Mathematica et des systèmes apparentés »), ses théorèmes d'incomplétude, établissant qu'un système d'axiomes parfait ne peut exister.

Le premier théorème d'incomplétude de Gödel affirme que tout système d'axiomes suffisamment puissant pour représenter l'arithmétique est nécessairement incomplet ou, autrement dit, pourra formuler des affirmations *indécidables*.

Le second théorème d'incomplétude de Gödel affirme quand à lui que tout système d'axiomes suffisamment puissant pour représenter l'arithmétique ne peut démontrer sa propre cohérence. Un système d'axiomes peut démontrer la cohérence d'un autre système, ou d'une restriction de lui-même, mais pas sa propre cohérence. Étant donné que le système d'axiomes parfait est censé représenter la totalité des mathématiques, il apparaît absurde d'envisager de se placer dans un système plus large pour démontrer la cohérence du premier.

Ainsi, la complétude est fondamentalement impossible, quelque soit le système choisis, et la cohérence indémontrable, contraignant les mathématiciens à travailler sous une épée de Damoclès, ne pouvant qu'espérer que leurs mathématiques soient consistantes.

John Von Neumann, mathématicien américano-hongrois, sans doute l'un des plus importants de l'histoire, notamment pour sa contribution dans la théorisation des ordinateurs (avec son *architecture de Von Neumann*) déclara en découvrant les résultats de Gödel "C'est fini." L'espoir d'une génération de mathématiciens d'établir des bases solides aux mathématiques s'effondrait.

Le cas ZFC

Qu'en est-il alors de ZFC ?

D'une part, il est peu probable que ZFC soit *incohérent*. Bien qu'il soit impossible de prouver sa cohérence, les mathématiciens s'accordent à dire que si ZFC était incohérent, une telle incohérence aurait déjà été découverte, tant ZFC a été éprouvé en tout sens.

L'incomplétude, toutefois, est plus problématique. Les mathématiciens travaillent désormais en connaissance de cause. Certains théorèmes, tels que l'hypothèse du continu (affirmant qu'il n'existe aucun ensemble de cardinal compris entre \mathbb{N} et \mathbb{R}), ne peuvent être *décidés* au sein de ZFC. Les mathématiciens doivent faire preuve d'humilité en acceptant que certains théorèmes n'admettent pas de démonstration.

[TOC]

"Les nuages ne sont pas des sphères, les montagnes ne sont pas des cônes, les rivages ne sont pas des arcs de cercle, l'écorce d'un arbre n'est pas lisse et l'éclair ne trace pas de ligne droite. La nature est complexe et la géométrie fractale rend compte de cette complexité et permet de l'étudier." — Benoît Mandelbrot

Introduction

Au XIX-ième siècle, apparaissent en mathématiques des objets pour le moins... étranges... Infiniment complexes, ils bouleversent la géométrie euclidienne en vigueur jusqu'à lors amenant progressivement au développement d'une nouvelle théorie géométrique, la *géométrie fractale*.

Le but de cet atelier est de présenter les principaux objets fractales rencontrés en mathématiques et de définir le plus clairement possible ce terme. Cette étude nous donnera l'occasion d'appréhender une nouvelle branche des mathématiques, l'analyse complexe.

La fonction de Weierstrass

En 1872, le mathématicien Allemand Karl Weierstrass présente à l'académie prussienne des sciences ce qui sera plus tard considéré comme le premier exemple de fractale. La *fonction de Weierstrass* se distingue en ceci qu'elle est continue partout mais dérivables nul part.

Elle est définie comme suit :

$$f(x) = \sum_{n=1}^{+\infty} b^n \cos(a^n x \pi)$$

où a est un entier impair, $b \in [0, 1[$ et $ab > 1 + \frac{3}{2}\pi$

Outre ces considérations purement *analytiques*, son graphe présente d'étranges propriétés géométriques.

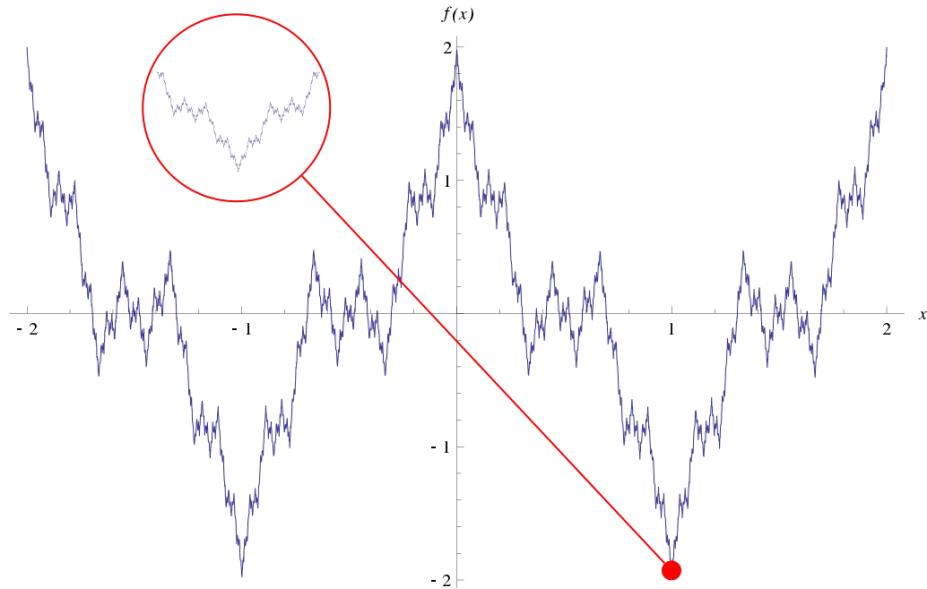


Figure 5: Représentation graphique de la fonction de Weierstrass

La fonction apparaît en effet infiniment détaillée et présente surtout une remarquable *auto-similarité*. Agrandir le graphe sur un point précis dévoile systématiquement une version plus petite du graphe global et ce à l'infini.

Les poussières de Cantor

Moins d'une décennie plus tard, le mathématicien Allemand Georg Cantor présente une autre construction autrement plus simple mais aux propriétés similaires : *les poussières de Cantor*.

Considérons le segment $C_0 = [0, 1]$.

À la première étape nous divisons notre segment en trois parties égales ($[0, \frac{1}{3}], [\frac{1}{3}, \frac{2}{3}], [\frac{2}{3}, 1]$) et supprimons la partie centrale. La figure ainsi obtenue correspond au segment $C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$

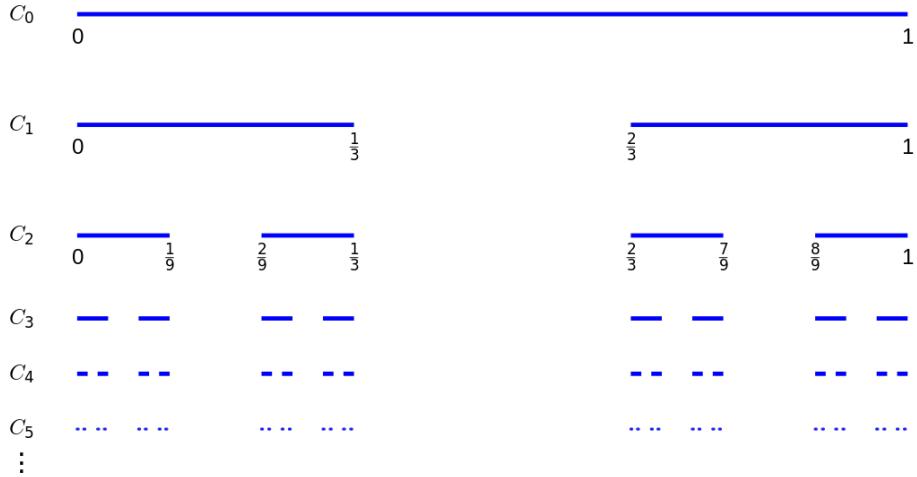


Figure 6: Étapes de construction des poussières de Cantor

De façon similaire, à la deuxième étape, nous divisons chacun des segments obtenus en trois parties et retirons celles du milieu, pour finalement aboutir au segment C_2 à 4 morceaux.

De manière générale, pour passer du segment C_k au segment C_{k+1} , nous divisions en 3 les 2^k morceaux obtenus à l'étape C_k et retirons les parties du milieu, obtenant ainsi 2^{k+1} morceaux.

Il s'agit là d'une construction dite “itérative” où le résultat de chaque étape sert de matériaux à l'étape suivante.

Les poussières de Cantor sont alors obtenues en réalisant cette procédure un nombre infini de fois ou, de manière plus formelle :

$$C = \bigcap_{i=1}^{+\infty} C_i$$

2

Une nouvelle fois, l'ensemble de Cantor présente la propriété d'auto-similarité (ce qui apparaît d'ailleurs clairement dans sa définition).

²Il est également tout à fait possible, et c'est en réalité le cas le plus fréquent dans la littérature mathématique, de définir ζ sur \mathbb{C} .

Le flocon de Von Koch

Le mathématicien suédois Helge von Koch publie en 1904 ce qui sera appelé plus tard la *courbe de Von Koch*.

Cette fractale reprend la construction itérative des poussières de Cantor tout en présentant les propriétés de continuité et de non dérivabilité de la fonction de Weierstrass.

Partons d'un segment V_0 d'une unité de longueur.

Pour passer à l'étape suivante, nous découpons V_0 en trois parties, retirons la partie centrale et plaçons de part et d'autre du trou ainsi formé des segments de même longueurs que la partie retirée de sorte à former un triangle équilatéral. Nous obtenons alors V_1

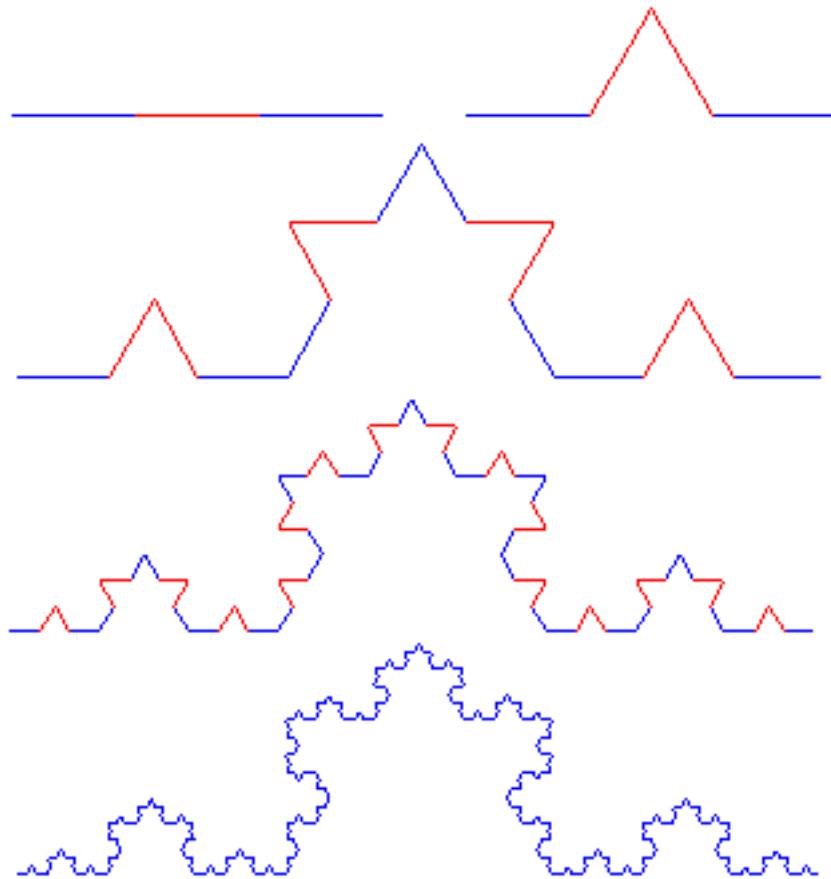


Figure 7: Étapes de construction du flocon de Von Koch

De la même manière, pour passer de l'étape V_1 à V_2 , nous appliquons à chaque segment de V_1 cette même transformation.

De façon générale, en notant f la fonction qui à une certaine étape associe l'étape suivante et en notant f^k la composée k-ième de f ,³ la *courbe de Von Koch* V est :

$$V = \lim_{k \rightarrow +\infty} (f^k(V_0))$$

À nouveau, et par construction, la courbe de Von Koch présente cette propriété d'auto-similarité.

Fractales Aléatoires

Il est également possible de générer des fractales de façon aléatoire / stochastique. Ce type de modèle est en particulier utilisé dans la modélisation de phénomènes physiques où la “régularité” des méthodes précédentes ne permet pas de représenter correctement les réalités du monde réel.

L'une des fractales aléatoires les plus simples consiste en une généralisation d'une fractale déterministe, le “tapis de Sierpinski”. Considérons un carré C_0 de côté 1.

À la première étape, nous découpons C_0 en 9 sous-carrés égaux de côtés $\frac{1}{3}$. Nous fixons alors une probabilité $0 < p < 1$ et choisissons de supprimer chacun des sous-carrés selon cette probabilité. De la même manière, pour passer de C_1 à C_2 , nous découpons chacun des sous-carrés restants en 9 parts égales et répétons pour chacun d'entre eux cette expérience aléatoire.

De manière générale, en notant f la fonction permettant de passer de l'étape C_k à l'étape C_{k+1} , un tapis de Sierpinski associé à la probabilité p correspond alors à :

$$T = \lim_{n \rightarrow +\infty} f^k(C_0)$$

La notion de fractales aléatoires est par exemple particulièrement utile en informatique dans la génération de paysages aléatoires. Pensons par exemple au jeu vidéo minecraft dont la structure du monde peut s'apparenter à un objet fractale.

Définition d'une Fractale

Avec tous ces exemples en tête, il est désormais possible d'approcher une définition du concept de “fractale” comme possédant une ou plusieurs des propriétés suivantes :

³de sorte que $f(V_0) = V_1$, $f(V_1) = f^2(V_0) = V_2$ et, de manière générale $f^k(V_0) = V_k$

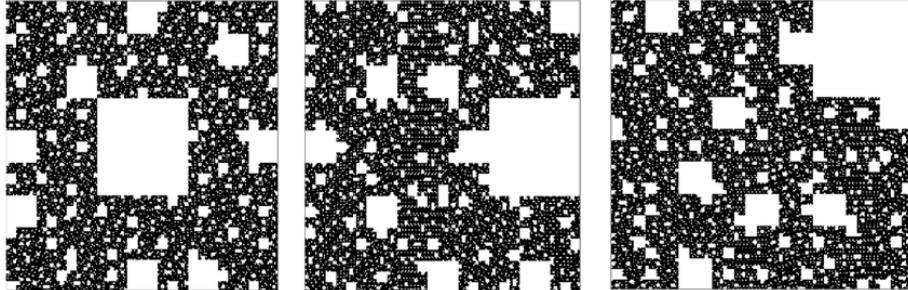


Figure 8: Trois exemples de tapis de Sierpinski aléatoires

- Présence de détails à une échelle arbitrairement petite ;
- Une structure trop irrégulière pour être décrite avec les outils de la géométrie traditionnelle ;
- Propriété d'auto-similarité

Par ailleurs, il est possible de distinguer dans ce large éventail d'objets mathématiques trois “catégories” générales de fractales :

- Fractales aléatoires (telles que le tapis de Sierpinski aléatoire)
- Fractales construites sur des fonctions itérées (telles que le flocon de Von Koch, les poussières de Cantor etc.)
- Fractales construites sur des relations de récurrences (telles que les ensembles de Julia et l'ensemble de Mandelbrot)

Mais existe-t-il une définition plus générale du concept de fractales ?

La dimension de Hausdorff-Besicovitch

Benoît Mandelbrot, pionnier de l'étude des fractales, a fourni une définition convenable et relativement générale du concept de fractale. Nous nous contenterons d'en esquisser les concepts mathématiques sous-jacents.

Nous pouvons parfaitement concevoir intuitivement le concept de dimension dans le cas d'espaces “simples”. Une droite est de dimension 1, un plan de dimension 2, l'espace de dimension 3 etc. Elle peut être définie comme le nombre minimal de coordonnées nécessaires pour identifier un point de l'espace en question.

La dimension de Hausdorff-Besicovitch (ou dimension de Hausdorff) fournit une autre manière d'appréhender le concept de dimension d'un ensemble. Considérons un ensemble U non vide⁴. On note $|U|$ le *diamètre* de U défini comme :

$$|U| = \sup\{|x - y| : x, y \in U\}$$

⁴Il s'agit en réalité d'une partie d'un espace métrique.

i.e la borne supérieure des distances de deux éléments de l'ensemble ou, de façon similaire, la plus grande distance séparant deux éléments de l'ensemble. Il est alors possible de se demander : étant donné un certain ensemble M , combien de sous ensemble de diamètre au plus δ sont nécessaires pour couvrir entièrement M ?

Pour $\delta > 0$, nous notons $N(\delta)$ ce nombre. Il est clair que $N(\delta)$ croît à mesure que δ décroît. Nous nous intéressons alors au comportement de ce nombre lorsque δ tend vers 0.

Dans le cas d'un segment de longueur 1, il est clair que :

$$N(\delta) = \frac{1}{\delta}$$

Dans le cas d'un segment de longueur 2, nous avons cette fois-ci :

$$N(\delta) = \frac{2}{\delta}$$

Lorsque l'on fait tendre δ vers 0, le $N(\delta)$ associé au segment de longueur 2 croît proportionnellement au $N(\delta)$ associé au segment de longueur 1, en l'occurrence $\frac{1}{\delta} = \delta^{-1}$.

La dimension de Hausdorff correspond alors précisément à l'exposant associé au δ , qui semble se confondre, dans les cas simples et réguliers, avec la dimension “classique”.

Elle mesure en réalité le niveau d'intrication de l'ensemble.

Les fractales sont alors caractérisées comme étant les ensemble dont la dimension de Hausdorff excède la dimension “classique”. Le flocon de Von Koch, par exemple, possède une dimension de Hausdorff de $\frac{\ln(4)}{\ln(3)} \approx 1,26$, légèrement au dessus de la dimension 1 associée à une courbe. Dès lors, le flocon de Von Koch peut être considéré, selon la définition de Benoît Mandelbrot, comme une fractale.

Cette définition, en plus d'être délicate à manipuler, présente toutefois certaines limites. Définir le concept de fractale semble être une quête vaine, le paysage fractale étant bien trop large pour être rassemblé sous une unique propriété commune.⁵: La définition généralement admise est alors celle de la section précédente, définissant une fractale comme un ensemble respectant une ou plusieurs propriétés caractéristiques.

⁵Benoît Mandelbrot a lui même écrit dans son ouvrage *Fractales, hasard et finance* “Il est vrai que [mes textes] avaient eu l'imprudence de proposer, pour le concept de fractale, une « définition pour voir », ou « définition tactique ». Ses défauts majeurs, vite apparus, me l'ont fait retirer dès le deuxième tirage.”

Les ensembles de Julia

Les mathématiciens français Pierre Fatou et Gaston Julia développent au début du XXe siècle une nouvelle branche des mathématiques, la dynamique holomorphe, ouvrant la porte à de nouvelles constructions fractales, parmi lesquelles les ensemble de Julia et de Fatou.

Il s'agit de deux ensembles complémentaires définis dans un cadre très large que nous restreindront à un cas particulier.

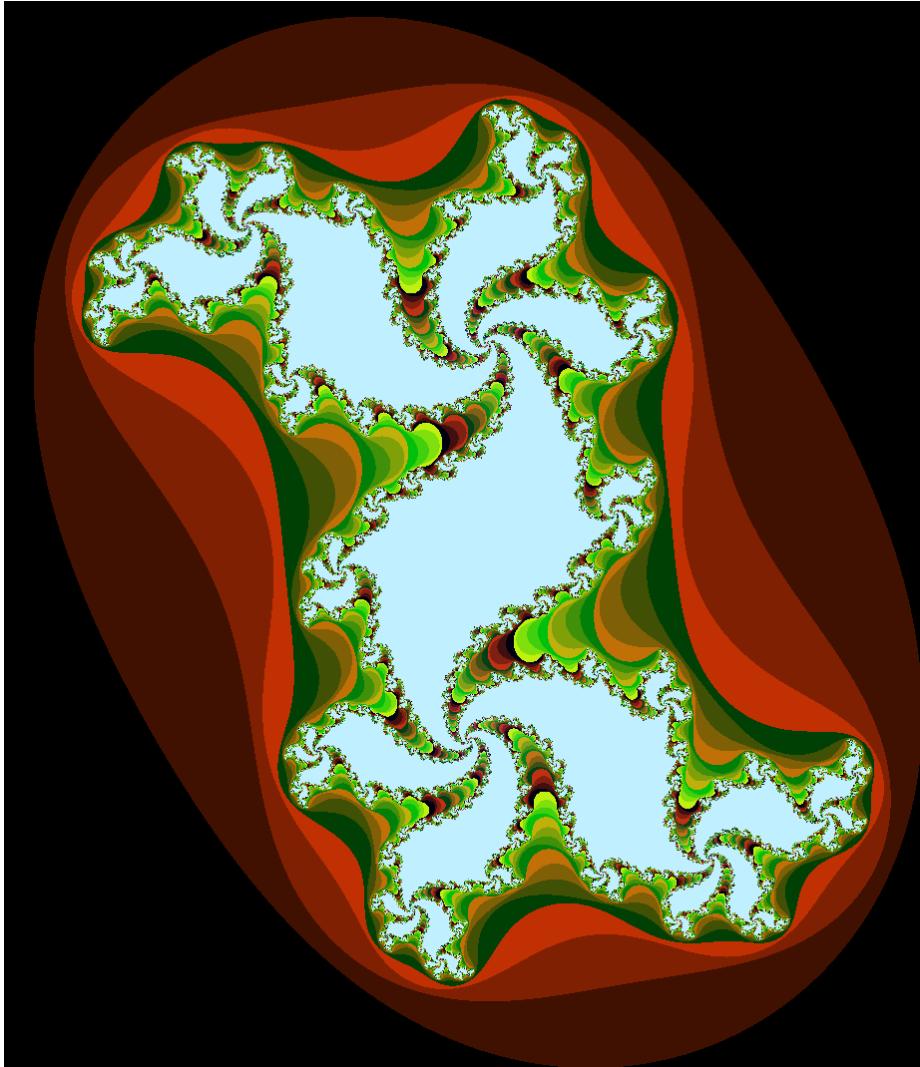
Considérons la suite à valeurs complexes (z_n) telle que $z_0 \in \mathbb{C}$ et

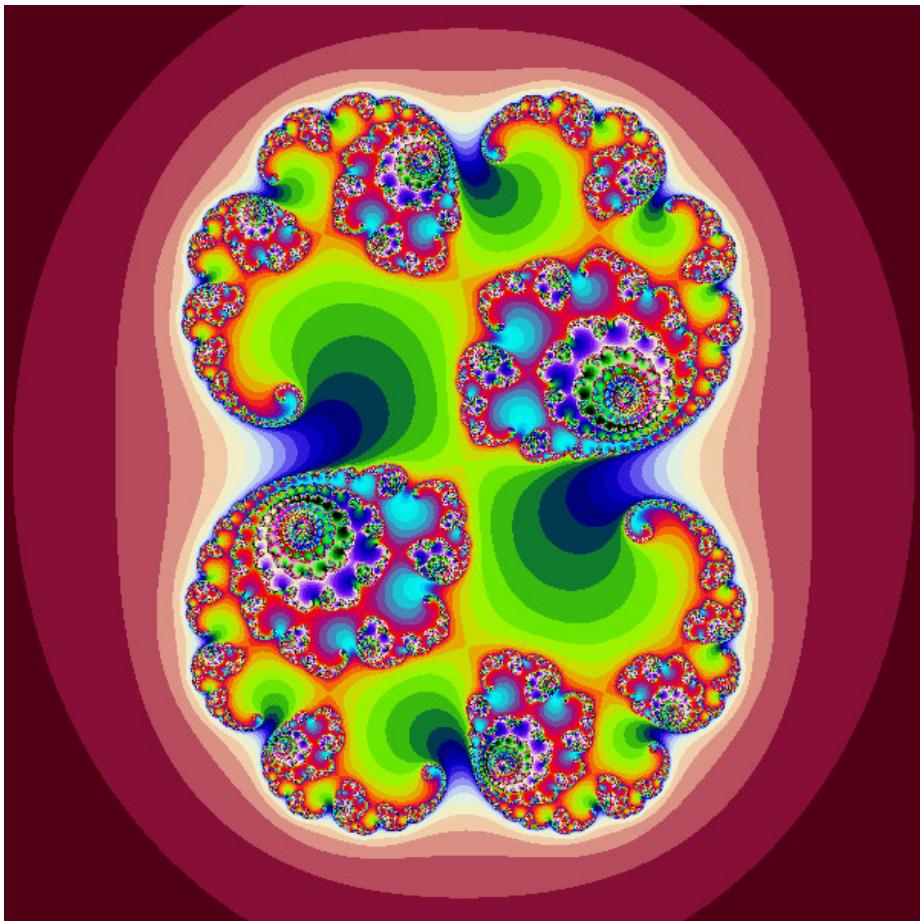
$$z_{n+1} = z_n^2 + c$$

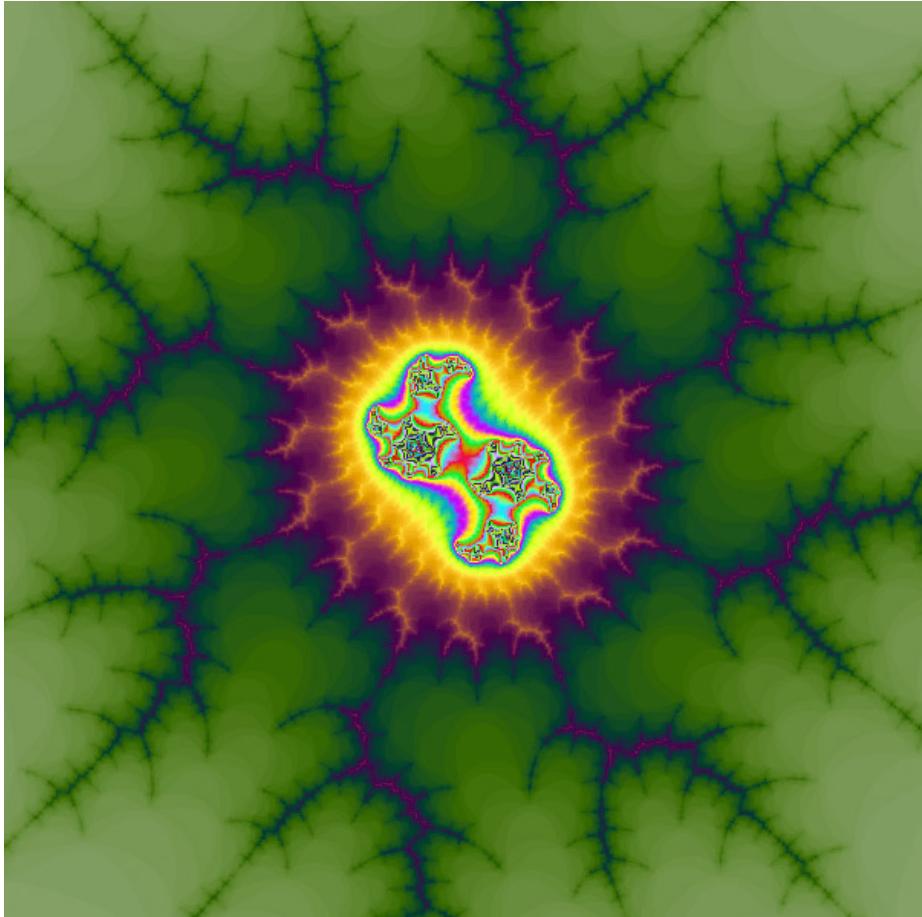
Pour $c \in \mathbb{C}$ fixé, les différentes valeurs de z_0 peuvent donner naissance soit à une suite bornée soit à une suite dont les termes tendent progressivement vers l'infini.

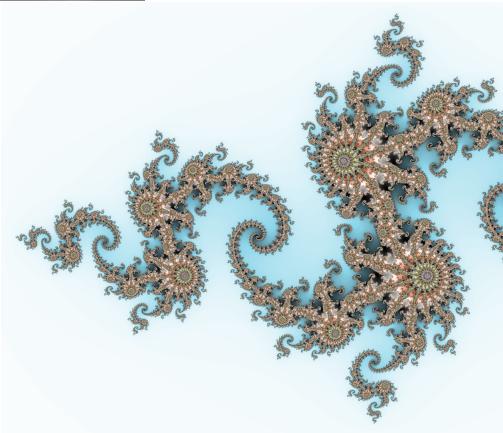
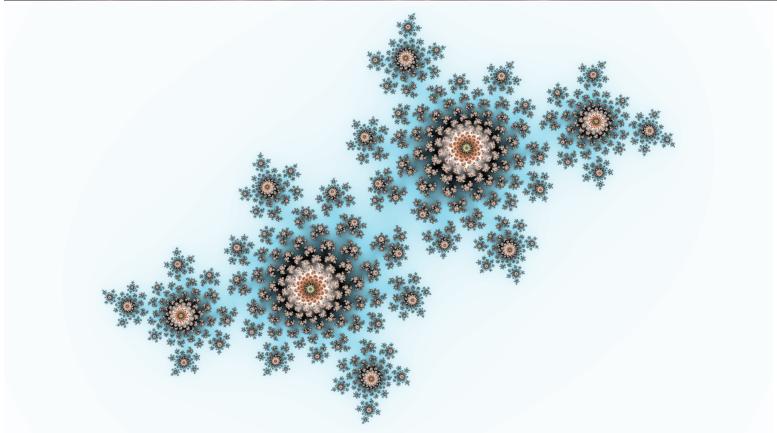
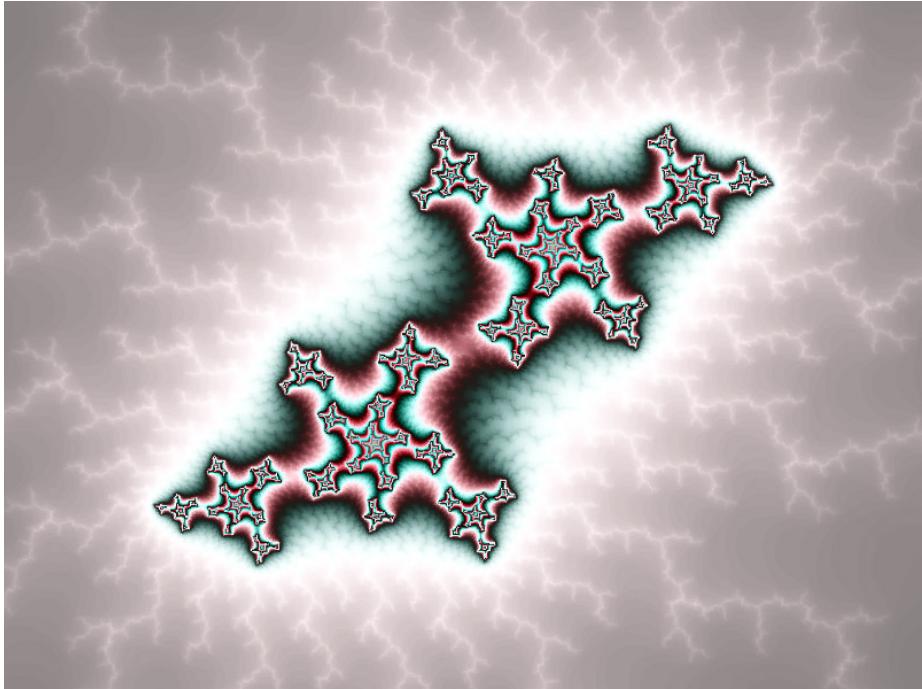
L'ensemble de Julia pour un certain $c \in \mathbb{Z}$ est alors défini comme *la frontière* de l'ensemble des $z_0 \in \mathbb{C}$ tels que la suite (z_n) associée est bornée.

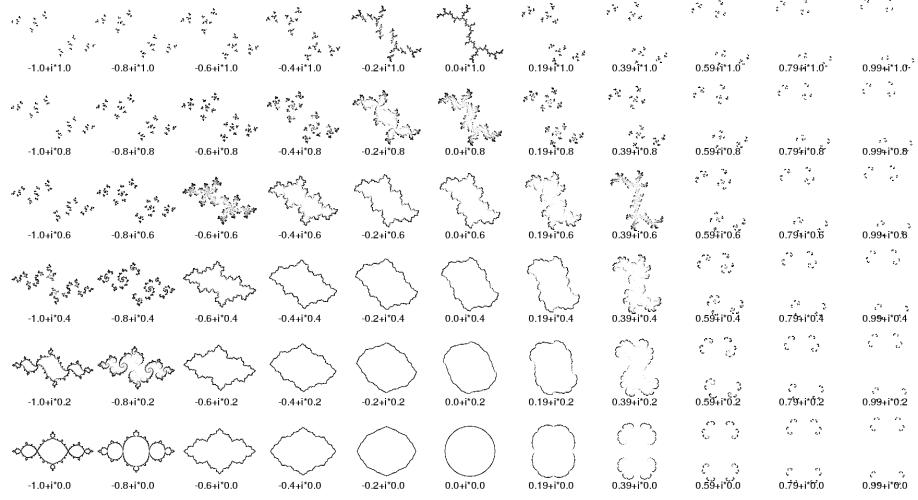
Il est alors possible de représenter cet ensemble dans le plan en coloriant d'une certaine couleur les points de l'ensemble, et d'une autre les points du complémentaire. Mieux encore, nous pouvons étudier la vitesse avec laquelle la suite diverge vers l'infini et colorier les points du complémentaire selon cette valeur.











L'ensemble de Julia dépend de la valeur de c choisie. Nous pourrions dès lors imaginer faire varier ce paramètre, obtenant ainsi un “film”, dont chaque image représente un ensemble de Julia pour un certain paramètre $c \in \mathbb{C}$.

L'ensemble de Julia est une fractale, en ce sens qu'elle présente une structure infiniment détaillée (sans pour autant être auto-similaire).

Ensemble de Mandelbrot

De la notion d'ensemble de Julia naît celle d'ensemble de Mandelbrot, du nom du mathématicien polono-franco-américain Benoît Mandelbrot.

Plutôt que d'observer le comportement de (z_n) lorsque l'on fait varier z_0 , on se propose ici de fixer $z_0 = 0$ et de faire varier le paramètre c . De la même manière que précédemment, l'ensemble de Mandelbrot peut alors être défini comme l'ensemble de toutes les valeurs de c pour lesquelles la suite z_n est bornée, en posant $z_0 = 0$.

Dans ce contexte, chaque point du plan complexe est associé à une fractale de Julia. En particulier, il est possible de démontrer qu'un point $z \in \mathbb{C}$ appartient à l'ensemble de Mandelbrot si et seulement si la fractale de Julia associée est connexe, i.e d'un seul morceau.

À nouveau, il est possible de tracer l'ensemble de Mandelbrot dans le plan, en faisant varier les couleurs des points du complémentaire suivant leur vitesse de divergence. (explorer l'ensemble de mandelbrot)

Contrairement aux ensembles de Julia, l'ensemble de Mandelbrot présente la propriété (ultime) d'auto-similarité. Il est alors possible de retrouver dans l'ensemble de Mandelbrot une version miniature mais strictement identique à l'ensemble global.

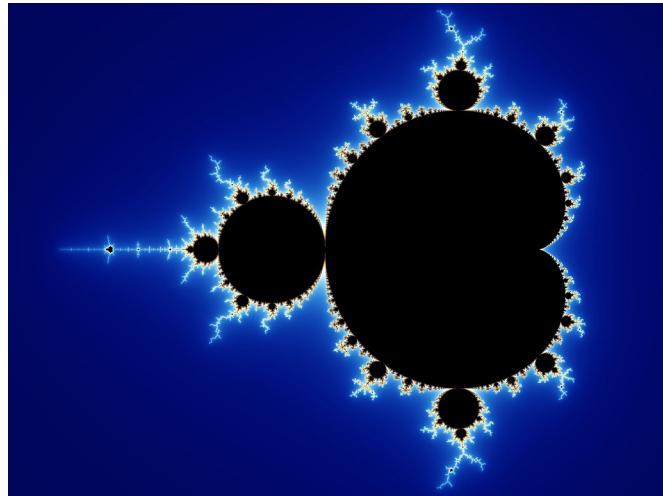
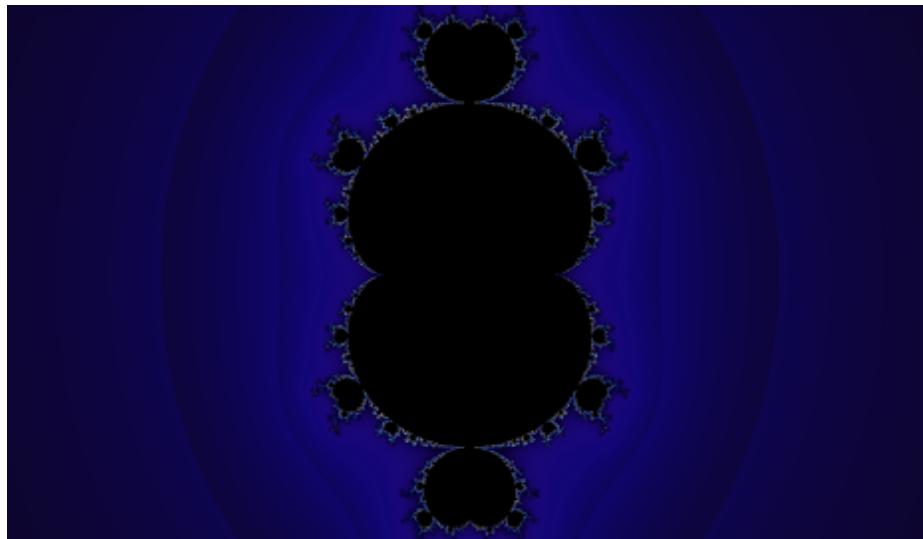
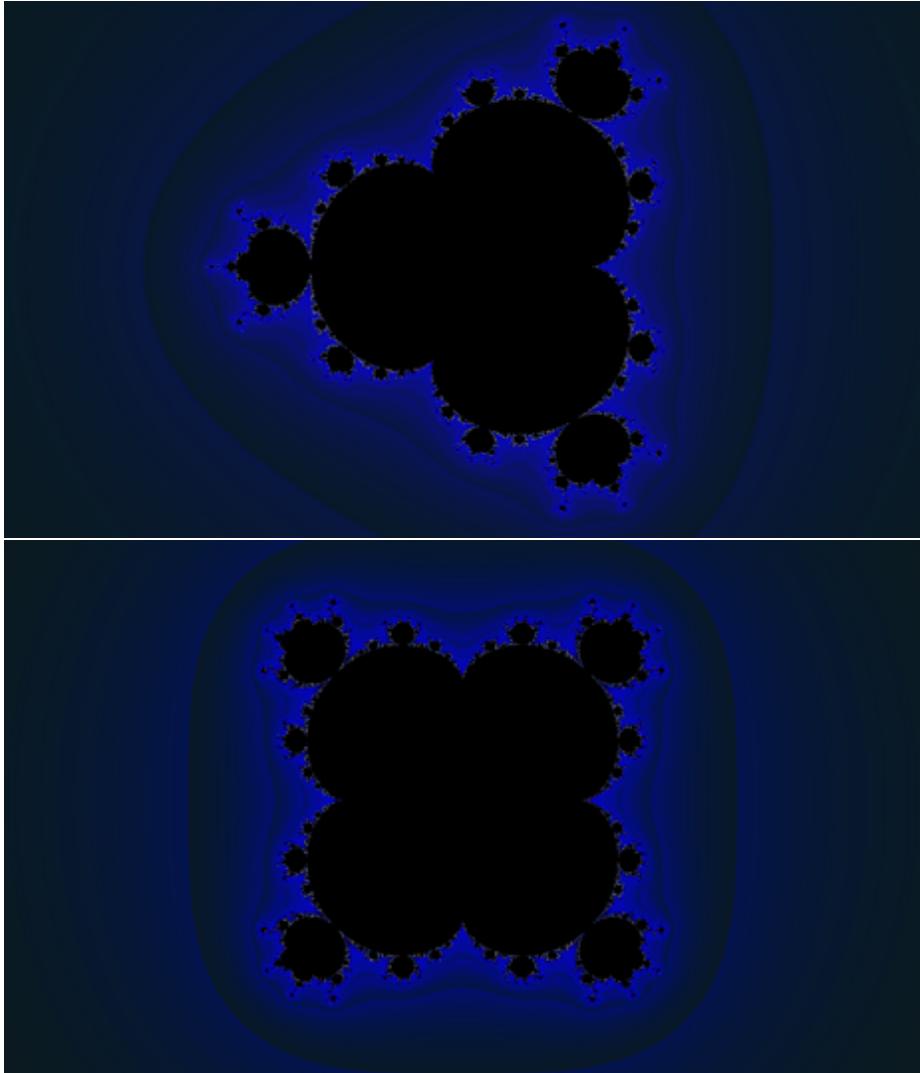
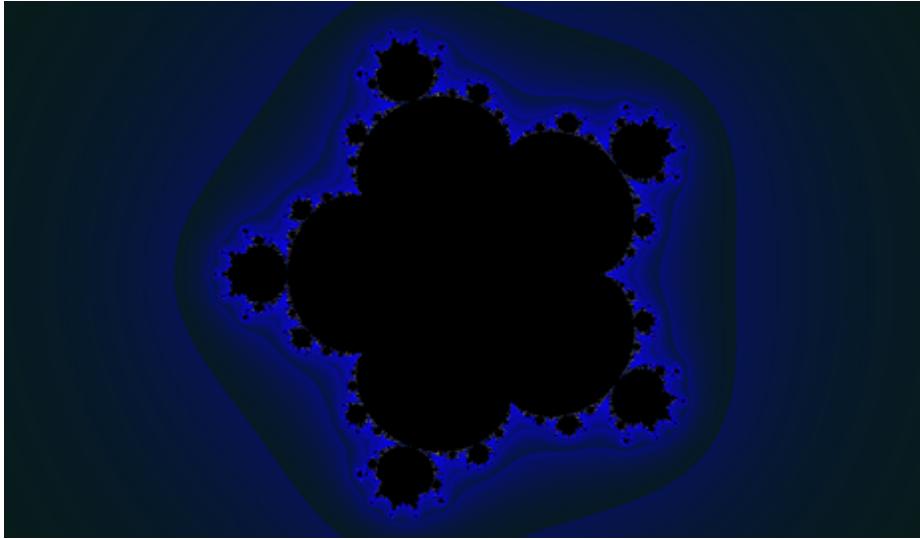


Figure 9: Ensemble de Mandelbrot

La construction de l'ensemble de Mandelbrot peut se généraliser en prenant comme suite support non plus $z^2 + c$ mais $z^d + c$ avec $d > 2$. Nous obtenons alors des “multibrot”.







Annexe

Voici quelques ressources pour approfondir le sujet.

Chacune d'entre elle est précédée d'un certain nombre d'étoiles suivant sa complexité, allant d'une étoile (*) pour les plus faciles à trois étoiles (***) pour les plus complexes :

- (*) Une vidéo d'introduction au concept de fractales Les fractales - Mickaël Launay
- (**) Une vidéo sur les fractales de Julia et l'ensemble de Mandelbrot Deux (deux ?) minutes pour Mandelbrot — El Jj
- (***) La page Wikipedia consacrée aux ensembles de Julia Ensemble de Julia — Wikipedia
- (***) La page Wikipedia consacrée à l'ensemble de Mandelbrot Ensemble de Mandelbrot — Wikipedia
- (**) La page Wikipedia consacrée à la notion de Fractales Fractales — Wikipedia
- (***) Une conférence sur les fractales aléatoires Vincent Beffara, Fractales aléatoires — Institut Henri Poincaré
- (***) Un papier introduisant les concepts mathématiques sous-jacents à la théorie des fractales (en anglais) Introduction to Fractals and Julia Sets — Fergus Cooper

[TOC]

Introduction

La théorie des nombres, branche des mathématiques étudiant les propriétés relatives aux nombres premiers, est généralement étudiée en classe de terminal. Les principaux résultats y sont brièvement exposés, passant sous silence leur démonstration pourtant fondamentale, tant par leur esthétique mathématique que pour leur intérêt dans la compréhension profonde des concepts.

Cet article constitue donc un tour d'horizon des principaux résultats de théorie des nombres, ainsi que de leur démonstration.

Il peut être lu sans pré-requis particuliers.

Rappels

Soit $n \in \mathbb{N}$. n est dit premier si et seulement si n n'est divisible que par 1 et par lui-même.

En particulier, 1 **n'est pas** un nombre premier.

Nous noterons par la suite \mathbb{P} l'ensemble des nombres premiers, de sorte que :

$$\mathbb{P} = \{2, 3, 5, 7, 11, 13, 17, \dots\}$$

On dit de deux entiers qu'ils sont premiers entre eux si leur seul diviseur commun est 1. En particulier, tout nombre premier est premier avec tous les entiers inférieurs (strictement) à lui-même.

Le théorème de Bachet-Bézout affirme que deux entiers a et b sont premiers entre eux si et seulement si il existe deux entiers u et v tels que :

$$au + bv = 1$$

La factorielle d'un certain entier n correspond au produit de cet entier par tous les entiers inférieurs à lui-même. Par exemple, $9! = 9 \times 8 \times \dots \times 2 \times 1$. On admet par convention que $0! = 1$.

Soient k et n deux entiers. On appelle *coefficient binomial* et on note $\binom{n}{k}$ le nombre de parties (non ordonnées) de k éléments dans un ensemble à n éléments.

On a par définition :

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

Ces coefficients binomiaux apparaissent dans la formule du binôme de Newton affirmant que pour tout réels a et b et pour tout entier naturel n :

$$(a + b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}$$

En particulier :

$$(a + 1)^n = \sum_{k=0}^n \binom{n}{k} a^k$$

Théorème fondamental de l'arithmétique

L'intérêt de l'étude des nombres premiers réside en ceci qu'ils sont le ciment des nombres entiers. Le théorème fondamental de l'arithmétique affirme en effet que

Tout entier naturel n peut s'écrire sous forme unique (à l'ordre près) comme produit de nombres premiers.

Autrement dit, les nombres premiers permettent de construire tous les nombres et ce, de manière unique.

Par exemple $20 = 2^2 \cdot 5$ ou encore $65 = 13 \cdot 5$.

La démonstration de ce théorème se décompose en deux parties : démontrer d'une part *l'existence* de cette décomposition et d'autre part *l'unicité*.

Existence

Soit un entier naturel n .

Si n est premier, alors nous n'avons rien à démontrer.

Dans le cas contraire, n possède des diviseurs compris entre 1 et n . Posons m le plus petit de ces diviseurs, il est alors premier. Si ce n'était pas le cas, il existerait un certain entier l tel que :

$$1 < l < m$$

et :

$$l|m$$

Mais puisque $l|m$ et $m|n$ nous aurions en particulier $l|n$, ce qui contredit la définition de m comme étant le plus petit diviseur de n . Donc m est nécessairement premier d'où n est divisible par un nombre premier.

Nous pouvons donc réécrire :

$$n = p_1 n_1, 1 < n_1 < n$$

À nouveau, si n_1 est premier, la démonstration s'achève et, si il ne l'est pas, il est divisible par un nombre premier p_2 de sorte que :

$$n = p_1 n_1 = p_1 p_2 n_2, 1 < n_2 < n_1 < n$$

En répétant l'opération suffisamment de fois, et comme le n_i restant diminue à chaque itération, nous obtenons finalement :

$$n = p_1 p_2 \dots p_n$$

qui correspond bien à la forme recherchée. L'existence est donc prouvée.

Comme les p_i ne sont pas nécessairement distincts les uns des autres, il est possible de réécrire cette forme de façon plus commode :

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_n^{\alpha_n}$$

Nous appellerons cette forme la *forme standard* de l'entier n .

Unicité

Il est clair, d'après, la première partie de la démonstration, que si un nombre premier p divise un produit de facteurs premiers $p_1 \dots p_n$ alors p est nécessairement l'un des p_i .

Appuyons nous sur ce résultat pour démontrer l'unicité de la *forme standard*.

Supposons par l'absurde dans un premier temps qu'il existe deux décompositions en facteurs premiers, de sorte que :

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k} = q_1^{\beta_1} q_2^{\beta_2} \dots q_j^{\beta_j}$$

Puisque les deux membres sont égaux et que tout p_i divise le membre de gauche, alors en particulier tout p_i divise le membre de droite, de sorte que, d'après le lemme précédent, tous les p sont des q et, de la même manière, tous les q sont des p . Nous avons donc $k = j$ et $p_i = q_i$ et ce, pour tout i .

Il reste alors à démontrer que les exposants sont eux aussi identiques, i.e que $\alpha_i = \beta_i$ pour tout i . Supposons par l'absurde que $\alpha_i > \beta_i$. En divisant des deux côtés par $p_i^{\beta_i}$ nous obtenons :

$$p_1^{\alpha_1} \dots p_i^{\alpha_i - \beta_i} \dots p_k^{\alpha_k} = p_1^{\beta_1} \dots p_{i-1}^{\beta_{i-1}} p_{i+1}^{\beta_{i+1}} \dots p_k^{\beta_k}$$

Le membre de gauche est divisible par p_i mais pas le membre de droite. Les deux membres étant égaux, il s'agit là d'une contradiction. Nous aurions obtenu

la même contradiction en supposant $\beta_i > \alpha_i$. Nous avons donc nécessairement $\alpha_i = \beta_i$, ce qui achève la démonstration.

À noter que considérer 1 comme un nombre premier n'aurait pas permis d'avoir une décomposition unique.

Un corollaire clair à ce théorème porte le nom de *lemme d'euclide* et affirme que :

Soient a et b deux entiers et p un nombre premier. Si $p|ab$ alors $p|a$ ou $p|b$.

Infinité des nombres premiers

L'étude des nombres premiers mène rapidement à s'interroger sur leur répartition et en particulier sur leur quantité.

Euclide affirme, dans ses *Éléments* (proposition 20 du livre IX), qu'il en existe une infinité et en propose une démonstration, sans doute l'une des plus importante de l'histoire des mathématiques.

Il existe une infinité de nombres premiers.

Raisonnons pour cela par l'absurde.

Supposons dans un premier temps qu'il en existe un nombre fini, disons un nombre n . Autrement dit :

$$\mathbb{P} = \{p_1, p_2, \dots, p_n\}$$

Posons alors $p = p_1 p_2 \dots p_n + 1$.

p est un entier et donc, d'après le théorème fondamental de l'arithmétique, divisible par au moins un facteur premier p_i .

Ainsi, $p_i | p$ et $p_i | p_1 p_2 \dots p_n + 1$ donc $p_i | p_1 p_2 \dots p_n + 1 - p_1 p_2 \dots p_n = 1$, ce qui est absurde.

Donc quelque soit la liste finie de nombres premiers donnée, il sera toujours possible d'en construire un n'appartenant pas à cette liste. \mathbb{P} elle est donc nécessairement de cardinal (de taille) infini.

Une autre manière de démontrer l'infinité des nombres premiers consiste à approximer leur répartition à l'aide d'une fonction.

Considérons la fonction $\pi(x)$ de \mathbb{N} dans \mathbb{N} qui associe à chaque entier le nombre de nombres premiers inférieurs ou égaux à x .

Par exemple, $\pi(2) = 1$, $\pi(3) = 2$ et $\pi(4) = 2$.

Le théorème des nombres premiers affirme alors que

$$\pi(x) \sim \frac{x}{\ln(x)}$$

Ou autrement dit que

$$\lim_{x \rightarrow \infty} \pi(x) \cdot \frac{\ln(x)}{x} = 1$$

La fonction $\pi(x)$ se comporte donc de façon similaire à $\frac{x}{\ln(x)}$, qui diverge clairement vers l'infini.

Nous ne nous attarderons pas plus sur le sujet.

Lemme de Gauss

Un autre résultat fondamental de théorie des nombres est dû au mathématicien Gauss, mathématicien allemand du XIXe reconnu comme “Le Prince des mathématiques”.

Le lemme de Gauss affirme que :

Soient a, b, c trois entiers. Si $a|bc$ et si a et b sont premiers entre eux alors

Démonstration

Soient a, b et c trois entiers relatifs.

Supposons que $a|bc$ et que a et b sont premiers entre eux.

Puisque $a|bc$, il existe un certain entier k tel que :

$$bc = ka$$

Par ailleurs, d'après le théorème de Bachet-Bézout (ou théorème de Bézout), et puisque a et b sont premiers entre eux, il existe deux entiers u et v tels que :

$$au + bv = 1$$

En multipliant les deux membres de l'égalité par c nous obtenons :

$$cau + cbv = c$$

D'où, sachant que $bc = ka$:

$$cau + kav = c$$

En factorisant par a nous obtenons finalement :

$$a(cu + kv) = c$$

Autrement dit, a divise c (ce qu'il fallait démontrer).

Petit théorème de Fermat

Un autre résultat fondamental en théorie des nombres, concernant cette fois-ci les congruences, est dû à Pierre de Fermat, mathématicien français sans doute le plus important dans cette branche des mathématiques.

L'énoncé du petit théorème de fermat est le suivant :

Soient p un nombre premier et a un entier naturel quelconque. Alors $a^p - a$ est divisible par p .

De plus, si p et a sont premiers entre eux alors $a^{p-1} - 1$ est divisible par p .

Par exemple, $8^3 - 8 = 504$ qui est divisible par 3.

Le petit théorème de Fermat peut également être formulé de la façon suivante, faisant intervenir explicitement son lien avec les congruences :

Soient p un nombre premier et a un entier naturel quelconque. Alors $a^p \equiv a \pmod{p}$.

Et si de plus p et a sont premiers entre eux alors $a^{p-1} \equiv 1 \pmod{p}$

Nous donnerons ici une démonstration due à Euler.

Démontrons dans un premier temps un lemme dont nous aurons besoin par la suite pour démontrer le petit théorème de Fermat.

Soient p un nombre premier et $k \in \{1, \dots, p-1\}$ alors $p| \binom{p}{k}$.

En effet :

$$\binom{p}{k} = \frac{p!}{k!(p-k)!} = \frac{p}{k} \times \frac{(p-1)!}{(k-1)!((p-1)-(k-1))!} = \frac{p}{k} \binom{p-1}{k-1}$$

D'où :

$$p \binom{p-1}{k-1} = k \binom{p}{k}$$

Donc p divise $k \binom{p}{k}$.

De plus, p et k étant premiers entre eux, nous avons, d'après le lemme de Gauss, $p| \binom{p}{k}$, ce qu'il fallait démontrer.

Démontrons maintenant le petit théorème de Fermat. Nous raisonnons pour cela par récurrence.

On cherche donc à prouver que pour tout $a \in \{1, \dots, p-1\}$, $a^p \equiv a \pmod{p}$. Il est en effet suffisant de prouver cette affirmation pour $a \in \{1, \dots, p-1\}$ car nous raisonnons ici par congruence.

Initialisation : On a $0^p \equiv 0 \pmod{p}$, donc la propriété est initialisée.

Héritage: Soit $a \in \{1, \dots, p-2\}$, supposons que $a^p \equiv a \pmod{p}$. Montrons que cette assertion est encore vraie pour $a+1$.

D'après le binôme de Newton,

$$(a+1)^p = \sum_{k=0}^p \binom{p}{k} a^k$$

Extrayons alors respectivement le premier et dernier terme de la somme ($k=0$ et $k=p$),

$$(a+1)^p = a^p + 1 + \sum_{k=1}^{p-1} \binom{p}{k} a^k$$

En passant aux congruences nous avons donc (puisque la somme est multiple de p et d'après l'hypothèse de récurrence) :

$$(a+1)^p \equiv a+1+0 \pmod{p}$$

Ou autrement dit :

$$(a+1)^p \equiv a+1 \pmod{p}$$

Ce qui achève la démonstration.

De plus, si a et p sont premiers entre eux alors $p|a^p-a$ implique que $p|a(a^{p-1}-1)$ et d'après le théorème de Gauss $p|a^{p-1}-1$ ou autrement dit $a^{p-1} \equiv 1 \pmod{p}$, qui correspond au corollaire présenté ci-dessus.

Le crible d'Ératosthène

Lorsque l'on cherche à lister les nombres premiers, la méthode consistant à les considérer un par un apparaît rapidement laborieuse et inefficace.

Il est toutefois très facile de construire la liste des nombres premiers inférieurs à un certain entier N en utilisant un algorithme connu sous le nom de “Crible d'Ératosthène” (mathématicien grec du 2ème siècle avant J.C.).

Fixons donc un certain $N \in \mathbb{N}$ et notons tous les entiers inférieurs ou égaux à N .

$$2, 3, 4, 5, 6, \dots, N$$

Le crible d'Ératosthène consiste à supprimer successivement les multiples pour ne garder que les nombres premiers.

Il se déroule comme suit :

- Entourer 2 et supprimer tous ses multiples (4, 6, 8, ...)
- Entourer 3 et supprimer tous ses multiples restants (9, 15, 21, 27, ...)
- Entourer 5 et supprimer tous ses multiples restants (25, 35, 55, 65, ...)

De manière générale, la n -ième étape consiste à entourer le dernier nombre à gauche non encore entouré et à supprimer tous ses multiples restants.

Une fois arrivé à \sqrt{N} , les nombres restants sont les nombres premiers inférieurs ou égaux à N .

Appliquons cette méthode au cas $N = 20$.

Comme $\sqrt{20} \approx 4.7$, il nous suffira de répéter l'opération jusqu'à ce que le dernier nombre restant à gauche soit supérieur à 5.

Écrivons donc la liste des entiers inférieurs ou égaux à N :

$$2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20$$

Nous supprimons donc tous les multiples de 2 (excepté lui-même) :

$$2, 3, 5, 7, 9, 11, 13, 15, 17, 19$$

Puis de 3 :

$$2, 3, 5, 7, 11, 13, 17, 19$$

Le dernier nombre restant à gauche étant 5 ($> \sqrt{20}$), la procédure est terminée.

Les nombres obtenus sont les nombres premiers inférieurs ou égaux à 20.

Introduction à la fonction zêta

La fonction zêta (ζ) est une fonction définie pour tout réel ⁶ x par :

$$\zeta(x) = 1 + \frac{1}{2^x} + \frac{1}{3^x} + \frac{1}{4^x} + \dots$$

ou de manière condensée :

$$\zeta(x) = \sum_{n=1}^{+\infty} \frac{1}{n^x}$$

⁶Il est également tout à fait possible, et c'est en réalité le cas le plus fréquent dans la littérature mathématique, de définir ζ sur \mathbb{C} .

Par exemple, $\zeta(1) = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots = +\infty$ ou $\zeta(2) = 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \dots = \frac{\pi^2}{6}$

Cette fonction revêt d'une importance toute particulière pour son lien avec les nombres premiers.

En effet :

$$\zeta(x) = \prod_{p \in \mathbb{P}} \frac{1}{1 - p^{-x}}$$

Autrement dit, un produit infini faisant intervenir l'ensemble des nombres premiers peut s'évaluer comme image de la fonction zêta.

Démonstration

Soit x un nombre réel.

On a :

$$\zeta(x) = 1 + \frac{1}{2^x} + \frac{1}{3^x} + \frac{1}{4^x} + \frac{1}{5^x} + \frac{1}{6^x} + \frac{1}{7^x} + \frac{1}{8^x} + \frac{1}{9^x} + \dots$$

En divisant par 2^x des deux côtés de l'égalité nous obtenons :

$$\frac{\zeta(x)}{2^x} = \frac{1}{2^x} \left(1 + \frac{1}{2^x} + \frac{1}{3^x} + \frac{1}{4^x} + \frac{1}{5^x} + \frac{1}{6^x} + \frac{1}{7^x} + \frac{1}{8^x} + \frac{1}{9^x} + \dots \right)$$

d'où :

$$\frac{\zeta(x)}{2^x} = \frac{1}{2^x} + \frac{1}{4^x} + \frac{1}{6^x} + \frac{1}{8^x} + \dots$$

En soustrayant la deuxième égalité à la première nous obtenons :

$$\left(1 - \frac{1}{2^x} \right) \zeta(x) = 1 + \frac{1}{3^x} + \frac{1}{5^x} + \frac{1}{7^x} + \frac{1}{9^x} \dots$$

Observons ici que les dénominateurs pairs (multiples de 2) ont été supprimés.

Répétons l'opération en divisant l'égalité obtenue par 3^x :

$$\frac{1}{3^x} \left(1 - \frac{1}{2^x} \right) \zeta(x) = \frac{1}{3^x} + \frac{1}{9^x} + \frac{1}{15^x} + \frac{1}{21^x} + \frac{1}{27^x} \dots$$

En soustrayant à nouveau cette égalité à l'égalité de départ nous obtenons :

$$\left(1 - \frac{1}{3^x}\right) \left(1 - \frac{1}{2^x}\right) \zeta(x) = 1 + \frac{1}{5^x} + \frac{1}{7^x} + \frac{1}{11^x} + \frac{1}{13^x} \dots$$

Nous avons cette fois-ci supprimés les dénominateurs multiples de 3.

Il apparaît alors clair que cette procédure est celle du crible d'Ératosthène.

En répétant le processus à l'infini nous obtenons finalement :

$$\zeta(x) \left(1 - \frac{1}{2^x}\right) \left(1 - \frac{1}{3^x}\right) \left(1 - \frac{1}{5^x}\right) \dots = 1$$

D'où nous pouvons extraire ζ :

$$\zeta(x) = \frac{1}{\left(1 - \frac{1}{2^x}\right) \left(1 - \frac{1}{3^x}\right) \left(1 - \frac{1}{5^x}\right) \dots}$$

Soit de façon condensée

$$\zeta(x) = \prod_{p \in \mathbb{P}} \frac{1}{1 - p^{-x}}$$

Annexe

Voici quelques ressources pour approfondir le sujet.

Chacune d'entre elle est précédée d'un certain nombre d'étoiles suivant sa complexité, allant d'une étoile (*) pour les plus faciles à trois étoiles (***) pour les plus complexes :

- (***): Une vidéo d'introduction à la fonction zêta et à la conjecture de Riemann Devenir RICHE grâce aux maths (La fonction Zeta de Riemann) — Axel Arno
- (***): Un livre de référence en théorie des nombres An Introduction To The Theory Of Numbers — G. H. Hardy

[TOC]

Définitions et architecture générale d'un modèle d'apprentissage supervisé

Le Larousse définit l'intelligence artificielle comme "l'ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine". L'IA est ainsi utilisée dans le cadre de problèmes qui ne peuvent être résolus avec un algorithme simple et "déterministe". Cette définition est volontairement vague et englobe un large éventail de techniques diverses

parmi lesquelles l'apprentissage automatique (en anglais : *machine learning*) défini comme un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'« apprendre ».

Deux types d'apprentissages automatiques existent. **L'apprentissage non supervisé** tout d'abord, qui consiste à laisser la machine apprendre par elle-même. Ce système est notamment utilisé dans le cadre des intelligences artificielles joueuse d'échecs dont on programme les règles et qui apprennent seule à vaincre les plus grands joueurs en s'entraînant contre elles-mêmes, progressant au fur et à mesure des parties. À l'opposé (et c'est ce qui va nous intéresser aujourd'hui) **l'apprentissage supervisé** consiste à assister la machine lors de son apprentissage en lui fournissant des exemples de ce que l'on souhaiterait qu'elle réalise. L'IA s'entraîne alors sur les exemples qui lui sont fournis et tente par elle-même de comprendre comment réaliser le processus qui mène de l'entrée à la sortie.

L'objectif du jour sera le suivant : construire une intelligence artificielle pour détecter si une image correspond à la photo d'un chat ou d'un chien. Nous disposons pour cela d'un “dataset”, c'est-à-dire d'une large base de données d'exemples, sur laquelle l'IA va pouvoir s'entraîner. Nous étudierons aujourd'hui un modèle bien particulier d'IA à apprentissage supervisé : les réseaux de neurones artificiels, modèle le plus répandu aujourd'hui dans le monde de l'IA.

Tout d'abord, avant de présenter les réseaux de neurones artificiels, présentons l'architecture générale d'un modèle d'apprentissage supervisé.

Nous cherchons donc à construire une machine qui, prenant une image en entrée, indique s'il s'agit d'un chat ou d'un chien. Il s'agit d'un **problème de classification** puisque la réponse attendue est la catégorie de l'image (chat / chien). Il s'agit par ailleurs d'un problème de classification **binaire** puisque ces catégories sont au nombre de deux.

La première étape est de transformer les données de manière à les rendre intelligibles par la machine. Nous travaillons dans notre cas avec des images, dont on peut extraire une à une la couleur des pixels de manière à obtenir une liste de réels x_1, x_2, \dots, x_n qui représenteront, mis bout à bout, l'image dans sa totalité. Dans la suite de l'exposé, nous utiliserons la notation X pour faire référence à x_1, \dots, x_n .

La prévision de notre IA lorsqu'une image lui est fournie est alors représentée par une fonction $a(X, W)$ qui dépend non seulement des données en entrée mais également de n paramètres (un associé à chaque variable d'entrée) appelés poids (weight) w_1, w_2, \dots, w_n correspondant là aussi à des nombres réels. Ces poids peuvent être vus comme les curseurs qui influencent la sortie du modèle. L'enjeu sera alors de trouver les bons poids de manière à avoir la prévision la plus juste possible. À noter que nous utiliserons la lettre W pour faire référence à ces poids w_1, \dots, w_n . La fonction a doit retourner un pourcentage correspondant à la probabilité selon le modèle que l'image appartienne à une classe ou à une autre (100% si il s'agit selon elle d'un chien, 70% si il s'agit selon elle probablement

d'un chien, 0% si il s'agit selon d'un chat). À préciser ici que n'ayant aucune information *a priori* sur le problème, les poids sont généralement initialisés aléatoirement.

Pour évaluer la justesse de la prédiction nous disposerons d'une fonction dite **fonction coût** $L(a, Y)$ qui dépend non seulement de la prédiction a du modèle mais aussi de la réponse Y que l'on aurait souhaité obtenir (d'où l'intérêt de s'entraîner sur des exemples "étiquetés", i.e où chaque image est associée à sa catégorie). Cette fonction prendra une valeur d'autant plus grande que l'erreur est importante. Si le modèle prédis un chien à 100% sachant qu'il s'agit d'un chat, l'erreur est importante. *A contrario*, si le modèle prédis qu'il s'agit d'un chien et que cela est confirmé par les données du dataset, l'erreur sera faible. Tout l'enjeu sera alors de trouver les poids W de manière à minimiser la fonction coût. Il s'agit donc d'un **problème d'optimisation**.

Ce processus de minimisation de L est réalisé par un **algorithme d'optimisation** qui se base sur les données fournies par la fonction coût pour adapter correctement les poids. Le plus populaire d'entre eux est l'algorithme de **descente de gradient** que nous allons présenter par la suite.

Ce processus de (1) prévision, (2) évaluation et (3) optimisation doit être répété un certain nombre de fois, il constitue la phase d'entraînement du modèle. Une fois entraîné, nous pourrons donner à notre modèle de nouvelles images qu'il sera alors capable d'analyser. Il procèdera alors à une **généralisation** de ses données d'entraînement.

Le modèle du neurone artificiel

Les modèles d'apprentissage supervisé varient les uns par rapport aux autres aux niveaux du modèle, de la fonction coût et de leur algorithme de d'optimisation. Nous analysons ici le modèle du neurone artificiel, que nous étendrons par la suite au réseau de neurones artificiel.

Au même titre que tous les modèles d'apprentissage supervisé, le neurone prend en entrée un nombre n de variables réelles x_1, x_2, \dots, x_n représentant la donnée fournie.

La prédiction du neurone se fait alors en deux phases.

La première consiste à opérer une simple combinaison linéaire entre les variables x_1, \dots, x_n et les poids w_1, \dots, w_n de notre neurone. Cette opération est symbolisée par la **fonction de combinaison z suivante** :

$$z(X, W) = w_1x_1 + w_2x_2 + \dots + w_nx_n = \sum_{k=1}^n w_kx_k$$

Nous ajoutons en pratique un paramètre supplémentaire b de manière à améliorer les résultats du neurone.

$$z(X, W, b) = w_1x_1 + w_2x_2 + \dots + w_nx_n + b = b + \sum_{k=1}^n w_kx_k$$

Cette fonction fournit alors une valeur d'autant plus grande que le modèle considère que l'image correspond à un chien et d'autant plus faible qu'il considère qu'il s'agit d'un chat. Nous cherchons toutefois à obtenir une probabilité, ce qui nécessite de composer notre fonction de combinaison avec la **fonction d'activation** suivante :

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Représentée par le graphe suivant

Cette fonction, dite **sigmoïde**, présente bien les propriétés recherchées : - Elle tend vers 0 en - (probabilité de 0 pour une valeur très faible) - Elle tend vers 1 en + (probabilité de 1 pour une valeur très élevée) - Elle est à valeur dans $]0, 1[$ (fournit un pourcentage quelque soit z) - Elle est définie sur \mathbb{R} tout entier - Elle est strictement croissante sur \mathbb{R} tout entier (pourcentage proportionnel à la valeur obtenue par la fonction combinaison).

Nous obtenons donc bien un pourcentage, d'autant plus grand que la sortie du neurone est grande, i.e d'autant plus proche de 100% que le neurone considère qu'il s'agit d'un chien.

Ainsi, partant de données X associées aux paramètres (W, b) le neurone fournit la sortie suivante (notée σ) :

$$\sigma = \sigma \circ z(X, W, b) = \frac{1}{1 + e^{-(w_1x_1 + \dots + w_nx_n + b)}}$$

Notre modèle est donc désormais capable de réaliser des prédictions dont nous allons évaluer la pertinence à l'aide de la fonction coût.

Nous disposons pour rappel d'un dataset de m lignes, associant chacune les données de la i ème image $X_{(i)}$ à la sortie souhaitée $y_{(i)}$ (1 pour chien, 0 pour chat). La fonction coût a alors pour rôle d'évaluer la moyenne des écarts, pour chacune des images présentes dans notre dataset, entre $y_{(i)}$ (la sortie que l'on aurait souhaité obtenir) et la sortie $\sigma_{(i)}$ effectivement prédite par le modèle pour cette image. Nous utilisons pour cela la fonction dite **Log Loss** qui s'exprime de la façon suivante :

$$L = -\frac{1}{m} \sum_{i=1}^m y_i \log(\sigma_i) + (1 - y_i) \log(1 - \sigma_i)$$

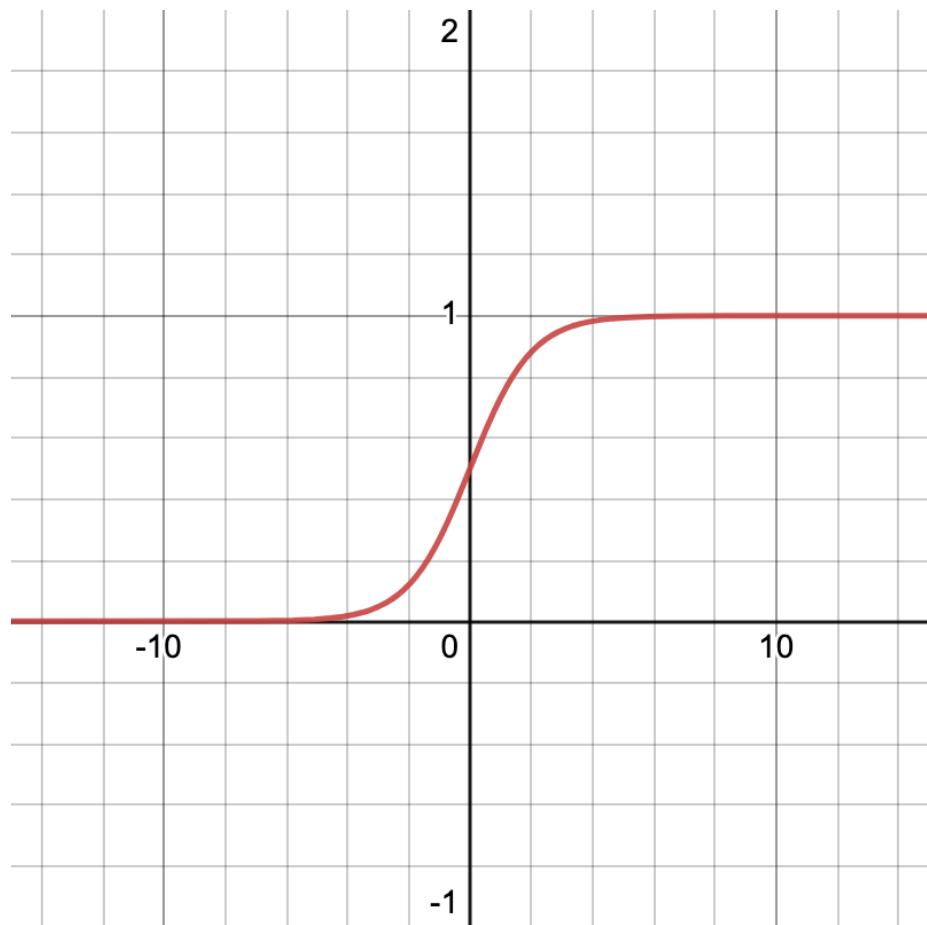


Figure 10: Graphe de la fonction sigmoïde

Le coût L correspond alors à la moyenne des écarts entre la sortie souhaitée et la sortie effectivement obtenue, valeur que l'on évalue dans ce cas à :

$$y_i \log(\sigma_i) + (1 - y_i) \log(1 - \sigma_i)$$

L'utilisation des logarithmes est ici dûe à une contrainte technique qui apparaît dans l'implémentation informatique de l'algorithme. (voir l'annexe pour en savoir plus)

“Il est clair que” la fonction présente les propriétés souhaitées, c'est-à-dire de fournir une valeur d'autant plus grande que l'écart est important et respectivement d'autant plus faible que l'écart est petit.

Nous disposons à ce moment de l'exposé d'une fonction de prédiction et d'une fonction d'évaluation. Il est maintenant temps de mettre en place une procédure pour adapter les poids W de manière à minimiser le coût L . C'est la phase **d'optimisation**.

Nous utiliserons pour cela l'algorithme de **descente de gradient**.

Revenons pour illustrer son fonctionnement à un modèle plus simple. Supposons que nous n'ayons en entrée de notre neurone qu'une seule variable x associée à son unique paramètre w . Nous pouvons alors tracer un graphe représentant le coût L en fonction du paramètre w .

Nous cherchons ici à trouver le paramètre w_{min} qui minimise l'erreur, i.e l'abscisse du minimum de $L(w)$. N'oublions pas que nous ne disposons au départ d'aucune information sur le problème, si bien que le paramètre w est initialisé aléatoirement. L'algorithme de descente de gradient consiste alors à utiliser la dérivée de L en w pour converger progressivement vers le minimum de la fonction. En effet, si la dérivée en w est négative, alors la pente de la tangente à L en w est négative et le w_{min} se trouve à droite de w ($w_{min} > w$). A contrario, si la dérivée en w est positive alors la pente de la tangente à L en w est positive et le w_{min} se trouve alors à gauche de w ($w_{min} < w$). En se basant sur la dérivée de L en w , nous pouvons donc adapter w de manière à le rapprocher de w_{min} . En répétant l'opération un certain nombre de fois, L va progressivement diminuer, découvrant ainsi le paramètre w_{min} (ou du moins une valeur approchée) qui minimise la fonction coût.

Cette idée peut être formalisée comme suit :

$$w_{t+1} = w_t - \frac{\partial L}{\partial w}$$

qu'il faut lire de la façon suivante : “Le poids à l'étape suivante est le poids à l'étape précédente auquel on soustrait la dérivée de L en w ”.

En pratique nous ajoutons un paramètre réel appelé *learning rate* (taux d'apprentissage) et noté α pour contrôler la vitesse de convergence du

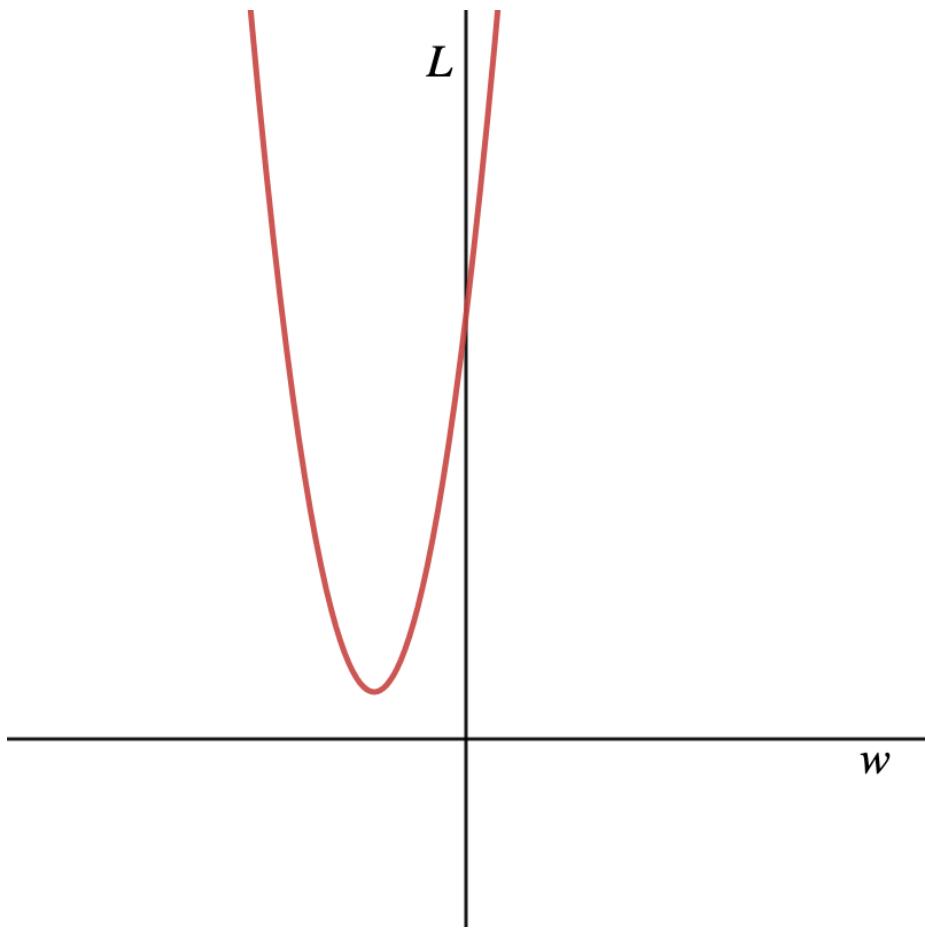


Figure 11: Représentation grpahique de L en fonction de w

paramètre w . Nous n'allons pas ici détailler davantage l'utilité du learning rate, bien que celui-ci soit central en pratique dans l'efficacité de l'entraînement de notre modèle. Dans la plupart des implémentations, α sera compris entre 0,01 et 0,1.

Ainsi, l'équation précédente devient :

$$w_{t+1} = w_t - \alpha \times \frac{\partial L}{\partial w}$$

Notre modèle est enfin terminé. Nous disposons bien d'une fonction de pré-diction, d'une fonction coût et d'un algorithme d'optimisation. En répétant la boucle (1) prédiction, (2) évaluation, (3) optimisation suffisamment de fois et sur suffisamment de données, notre modèle est désormais en capacité d'apprendre et de généraliser par la suite ses apprentissages.

Malgré tout, le modèle consistant en un neurone unique apparaît rapidement trop "limité". La simplicité du modèle ne lui permet pas d'établir des relations complexes entre des ensembles distincts de données, si bien qu'en pratique il ne peut être utilisé que pour un nombre restreint de problèmes "simples". Nous devons donc complexifier notre modèle pour augmenter ses possibilités d'actions.

Le modèle du réseau de neurones artificiels

Nous allons pour cela associer plusieurs neurones en un réseau, d'une façon similaire aux neurones biologiques.

Pour comprendre le fonctionnement d'un réseau de neurones, nous allons tout d'abord étudier un cas simple, celui d'un réseau à deux couches, l'une composée de deux neurones et l'autre d'un unique neurone (réseau 2×1).

L'évaluation d'un entrée X s'effectue selon un processus dit de **forward propagation** (propagation en avant). Les informations sont évaluées couches après couches, chacune d'entre elles utilisant les sorties de la précédente comme entrée.

Les données X sont donc tout d'abord envoyées dans chacun des neurones de la première couche, possédant tous leurs propres paramètres. Ainsi, si nous disposons dans notre première couche de deux neurones 1 et 2, leurs sorties respectives seront, d'après ce qui a été dit précédemment :

$$\sigma^{(1)} = \sigma \circ z^{(1)}(X, W^{(1)}, b^{(1)}) = \sigma \left(w_1^{(1)}x_1 + w_2^{(1)}x_2 + \dots + w_n^{(1)}x_n + b^{(1)} \right)$$

$$\sigma^{(2)} = \sigma \circ z^{(2)}(X, W^{(2)}, b^{(2)}) = \sigma \left(w_1^{(2)}x_1 + w_2^{(2)}x_2 + \dots + w_n^{(2)}x_n + b^{(2)} \right)$$

où $w_i^{(j)}$ désigne le poids associé à la i-ème entrée du j-ème neurone.

Introduisons, avant de continuer une notation dont nous aurons besoin par la suite. Travaillant sur plusieurs couches, il est nécessaire, dans nos équations, de préciser à quelles couches sont associés les différents termes qui interviennent. Nous utiliserons donc la notation suivante $d^{[c]}$ pour désigner la donnée d associée à la c -ème couche. Nous pouvons ainsi réécrire les équations précédentes (associées à la première couche) de la façon suivante :

$$\sigma^{(1)[1]} = \sigma \circ z^{(1)[1]}(X, W^{(1)[1]}, b^{(1)[1]}) = \sigma \left(w_1^{(1)[1]}x_1 + w_2^{(1)[1]}x_2 + \dots + w_n^{(1)[1]}x_n + b^{(1)[1]} \right)$$

$$\sigma^{(2)[1]} = \sigma \circ z^{(2)[1]}(X, W^{(2)[1]}, b^{(2)[1]}) = \sigma \left(w_1^{(2)[1]}x_1 + w_2^{(2)[1]}x_2 + \dots + w_n^{(2)[1]}x_n + b^{(2)[1]} \right)$$

Continuons sur notre lancée en analysant la deuxième couche de notre réseau de neurones 2×1 . Comme indiqué précédemment, le neurone de la deuxième couche ne se base désormais plus sur les données fournies à l'entrée du réseau mais sur les sorties $\sigma^{(1)[1]}$ et $\sigma^{(2)[1]}$ des neurones de la première couche. Sa sortie correspond donc à

$$\sigma^{(1)[2]} = \sigma \circ z^{(1)[2]}(\sigma^{(1)[1]}, \sigma^{(2)[1]}, W^{(1)[2]}, b^{(1)[2]}) = \sigma \left(w_1^{(1)[2]}\sigma^{(1)[1]} + w_2^{(1)[2]}\sigma^{(2)[1]} + b^{(1)[2]} \right)$$

À noter ici que le neurone de la deuxième couche possède autant de paramètre qu'il y a de neurones dans la première couche.

Cette valeur, obtenue à l'issue d'une forward propagation au travers des différentes couches du réseau, correspond à la prédiction du réseau.

Ces équations peuvent facilement se généraliser à n'importe quel réseau de neurones de la façon suivante :

- Pour le i -ème neurone de la première couche :

$$\sigma^{(i)[1]} = \sigma \circ z^{(i)[1]}(X, W^{(i)[1]}, b^{(i)[1]}) = \sigma \left(w_1^{(i)[1]}x_1 + w_2^{(i)[1]}x_2 + \dots + w_n^{(i)[1]}x_n + b^{(i)[1]} \right) = \sigma \left(b^{(i)[1]} + \sum_{k=1}^n w_k^{(i)[1]}x_k \right)$$

- Pour le i -ème neurone de la C -ième couche composée de p neurones ($C > 1$)
:

$$\sigma^{(i)[C]} = \sigma \circ z^{(i)[C]}(S^{[C-1]}, W^{(i)[C]}, b^{(i)[C]}) = \sigma \left(w_1^{(i)[C]}\sigma^{(1)[C-1]} + \dots + w_p^{(i)[C]}\sigma^{(p)[C-1]} + b^{(i)[C]} \right) = \sigma \left(b^{(i)[C]} + \sum_{k=1}^p w_k^{(i)[C]}x_k \right)$$

où $S^{[C-1]}$ correspond à l'ensemble des $\sigma^{(i)[C-1]}$, i.e à l'ensemble des sorties des neurones de la couche précédente.

La sortie du réseau étant un nombre réel, la fonction **Log Loss** est toujours adaptée pour évaluer la pertinence des prédictions du modèle.

Nous devons enfin adapter l'algorithme de descente de gradient au cas du réseau de neurones. Il suffit pour cela de dériver L de façon spécifique par rapport à chacun des paramètres du réseau. Ainsi, le i -ème poids du j -ème neurone de la c -ième couche à l'étape suivante peut être exprimé de la façon suivante :

$$w_i^{(j)[C]}(t+1) = w_i^{(j)[C]} - \alpha \times \frac{\partial L}{\partial w_i^{(j)[C]}}$$

Nous disposons désormais d'une fonction de prédiction, d'évaluation et d'un algorithme d'optimisation. Le modèle de réseau de neurone est donc terminé.

Pour illustrer, nous avons implémenté l'algorithme en Python dans le cadre du problème de la distinction chat / chien sur une image. Les graphiques ci-dessous représentent l'évolution des performances du modèle en fonction du nombre de répétition de la boucle prévision, évaluation, optimisation. Le graphique (1) représente la fonction Log Loss en fonction du nombre d'itération (elle diminue au fur et à mesure des itérations, i.e l'erreur diminue au fur et à mesure de l'entraînement). Le graphique (2) représente quant-à lui la précision du modèle, évaluée comme la proportion de "bonnes" prédictions sur l'ensemble du dataset (elle augmente au fur et à mesure de l'entraînement).

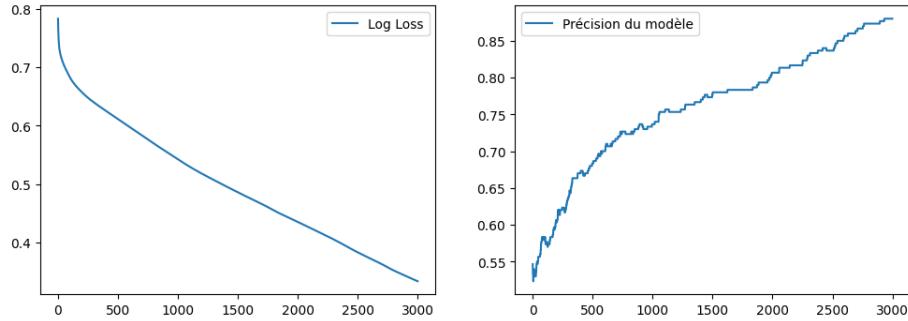


Figure 12: Graphiques représentant respectivement la fonction Log Loss en fonction du nombre d'itération et la précision du modèle en fonction du nombre d'itérations

Annexe

Voici quelques ressources pour approfondir le sujet.

Chacune d'entre elle est précédée d'un certain nombre d'étoiles suivant sa complexité, allant d'une étoile (*) pour les plus faciles à trois étoiles (***) pour les plus complexes.

- (*) Une vidéo sur la manière dont les IA peuvent comprendre le langage ScienceEtonnante — Comment les I.A. font-elles pour comprendre notre langue ?
- (*) Une vidéo pour comprendre le fonctionnement des IA génératrices d'images ScienceEtonnante — Comment ces IA inventent-elles des images ?
- (*) Une vidéo pour comprendre le fonctionnement de ChatGPT ScienceEtonnante — Ce qui se cache derrière le fonctionnement de ChatGPT
- (**) Une excellente série de vidéos sur le fonctionnement et les enjeux présent et futur de l'intelligence artificielle Science4All — L'intelligence artificielle et le machine learning
- (***) Une excellente formation (sur laquelle ce cours est en grande partie basé) sur le fonctionnement des réseaux de neurones artificiels et leur implémentation en Python. MachineLearnia — Formation Deep Learning

[TOC]

Section 1

Cras risus ipsum, faucibus ut, ullamcorper id, varius ac, leo. Pellentesque egestas, neque sit amet convallis pulvinar, justo nulla eleifend augue, ac auctor orci leo non est. Proin faucibus arcu quis ante. Fusce commodo aliquam arcu. In dui magna, posuere eget, vestibulum et, tempor auctor, justo.

Sed hendrerit. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Aenean viverra rhoncus pede.

Cras risus ipsum, faucibus ut, ullamcorper id, varius ac, leo. Pellentesque egestas, neque sit amet convallis pulvinar, justo nulla eleifend augue, ac auctor orci leo non est. Proin faucibus arcu quis ante. Fusce commodo aliquam arcu. In dui magna, posuere eget, vestibulum et, tempor auctor, justo.

Section 2

Cras risus ipsum, faucibus ut, ullamcorper id, varius ac, leo. Pellentesque egestas, neque sit amet convallis pulvinar, justo nulla eleifend augue, ac auctor orci leo non est. Proin faucibus arcu quis ante. Fusce commodo aliquam arcu. In dui magna, posuere eget, vestibulum et, tempor auctor, justo.

[TOC]

Définitions générales

La question du vote est omniprésente dans nos sociétés, que ce soit entre amis pour décider quel film choisir ou à l'échelle d'un pays dans le choix d'un président. De par leur aspect transversal, touchant aussi bien à la politique, à l'économie qu'aux mathématiques, les systèmes de votes font l'objet d'études poussées depuis plusieurs siècles.

Un système de vote, ou mode de scrutin, est un système par lequel agréger les préférences collectives des électeurs pour en déduire une préférence de groupe vis-à-vis d'un certain nombre d'alternatives ou candidats.

Nous nous limiterons ici au cas d'une élection simple aboutissant au choix d'une seule alternative parmi l'ensemble des alternatives possibles (un président parmi les candidats, un restaurant parmi l'ensemble de ceux dans un rayon d'un kilomètre etc.)

Nous négligeons par ailleurs les cas d'égalité entre candidats, dans la mesure où ils sont hautement improbables à l'échelle réelle.

Une façon simple de définir les préférences d'un individu est de définir une relation d'ordre totale sur l'ensemble des alternatives possibles, autrement dit une classification des différentes alternatives possibles.

Dans le cas d'une élection à trois candidats A, B et C, la préférence d'un électeur pourrait par exemple s'exprimer comme "A préféré à B, préféré à C", ou, de la même manière, "B préféré à A, préféré à C".

Nous appellerons cette classification individuelle une "liste de préférence" et nous noterons "A préféré à B" avec la notation usuelle $A > B$.

Nous pouvons alors rassembler l'ensemble de ces listes de préférences dans un "profil de préférences". Imaginons une élection à 3 candidats A, B et C et 10 électeurs parmi lesquels : - 5 préfèrent A à B à C - 3 préfèrent B à A à C - 2 préfèrent C à B à A

Le profil de préférence de l'élection peut alors se représenter de la manière suivante :

5	3	2
<hr/>		
A	B	C
B	A	B
C	C	A
<hr/>		

L'enjeux est alors d'agréger un profil de préférence en une préférence collective.

Panorama des principaux modes de scrutins

Nous disposons pour cela de nombreux modes de scrutin plus ou moins complexes.

Le scrutin majoritaire

Le plus évident d'entre eux est le scrutin majoritaire. Pour définir le gagnant de l'élection, nous comptons le nombre de fois où chaque candidat apparaît comme premier choix des électeurs. Si l'un d'entre eux obtient plus d'une stricte majorité des voies (plus de 50%), il gagne.

Par exemple, cette élection :

6	2	2
<hr/>		
A	B	C
B	A	B
C	C	A

v verrait le candidat A élu. En effet, il y a en tout 8 électeurs, le candidat A apparaît premier choix chez 6 d'entre eux, il s'agit d'une majorité et le candidat A est élu. Mais un problème surgit immédiatement : dans ce système, il n'y a pas toujours de vainqueur. En effet, pour peu que les voies soient correctement réparties, il se peut qu'aucun candidat n'obtienne de majorité. Nous devons donc améliorer ce système de manière à ce qu'il fournisse un vainqueur à chaque élection.

Le scrutin majoritaire modifié

Une autre solution peut alors consister à compter le nombre de fois où chaque candidat apparaît comme premier choix des électeurs et à élire le candidat présentant le plus de voix. Ainsi, le cas suivant :

4	3	3
<hr/>		
A	B	C
B	C	B
C	A	A

pour lequel il n'y aurait pas eu de vainqueur dans le système précédent, mène dans le scrutin majoritaire modifié à l'élection du candidat A. Ce système fournit toujours un vainqueur qui semble représenter la volonté du groupe.

Toutefois ce système présente deux principales limites.

Premièrement, il ne prend en compte que les premiers choix des électeurs, igno-

rant totalement le reste de la liste de préférence. Le fait d'ignorer le reste des préférences mène à des cas où un candidat est élu alors même que celui-ci est dernier sur la liste de préférence d'une majorité stricte d'électeurs. C'est par exemple le cas dans l'élection précédente où A, bien que premier dans 4 listes, est dernier dans les 6 listes restantes. A ne semble pas véritablement représenter la "préférence collective", le choix de B aurait dans ce cas semblé plus judicieux.

Une autre limite est la question de la dépendance aux alternatives non pertinentes. Imaginons une élection où deux partis sont présents : le parti A aux idées de gauche et la partie B aux idées de droite.

Un profil de préférence pourrait alors être le suivant.

6	4
<hr/>	
A	B
B	A

Dans ce contexte, A gagnerait 6 contre 4. Imaginons maintenant qu'un deuxième parti de gauche C se présente. Davantage en accord avec certaines idées de C, certains électeurs de A pourraient se diriger vers C de sorte à obtenir le profile de préférence suivant :

3	4	3
<hr/>	<hr/>	<hr/>
A	B	C
B	A	A

Dans ce contexte, c'est B qui est élu. L'introduction du candidat C a eu pour effet de "diviser" les voies à gauche, provoquant l'élection du candidat de droite, alors même qu'une majorité des électeurs étaient favorable à l'élection d'un candidat de gauche.

Cette absence d'indépendance aux alternatives non pertinentes est une menace sévère au pluralisme politique, fondement d'une démocratie qui fonctionne. Elle entraîne par ailleurs les électeurs à voter utile / stratégique plutôt que selon leurs véritables préférences.

Il nous faut donc améliorer ce système.

Le scrutin à deux tours

Une idée pourrait alors être d'organiser une élections en plusieurs "étapes" / "tours", permettant aux électeurs d'ajuster au fur et à mesure leurs votes pour refléter au mieux leur volonté.

Le manière la plus commune de procéder est d'organiser deux tours : deux candidats sont élus au premier tour selon un scrutin majoritaire modifié et ces derniers s'affrontent dans un scrutin majoritaire au second tour. C'est le système dans les élections présidentielles en France.

Cette manière de procéder, bien que légèrement meilleure que le scrutin majoritaire classique, ne règle pas entièrement le problème. Imaginons une élection entre deux parties de droite D1, D2 et trois parties de gauche G1, G2, G3.

Le profile de préférence suivant :

4	4	3	3	3
D1	D2	G1	G2	G3
D2	D1	G2	G1	G1
G1	G1	G3	G3	G2
G2	G2	D1	D1	D1
G3	G3	D2	D2	D2

Mènerait les deux candidats de droite au second tour (et donc à l'élection d'un candidat de droite), alors même qu'une majorité de la population vote à gauche. C'est pour pallier ce phénomène de division que les partis politiques organisent des primaires (qui ne font en fin de compte que repousser le problème au moment des primaires).

Le système de Hare

Nous pourrions alors pousser l'idée à l'extrême en organisant autant de tour que de candidats. Autrement dit, nous éliminerons à chaque tour le candidat ayant reçu le moins de premier choix des listes de préférences jusqu'à ce qu'il n'en reste qu'un seul. Il s'agit de la méthode de Hare.

Nous pouvons par exemple imaginer l'élection suivante :

4	5	3	2
A	C	B	A
B	B	D	D
C	D	C	C
D	A	A	B

Au premier tour, le candidat D est éliminé puisqu'il ne reçoit aucune voix. Le profile de préférence devient alors :

4	5	3	2
A	C	B	A
B	B	C	C
C	A	A	B

À nouveau, le candidat B est éliminé puisqu'il ne reçoit que 3 voix contre 6 pour A et 5 pour C. Le profile de préférence devient donc :

4	5	3	2
A	C	C	A
C	A	A	C

Dont on peut fusionner les colonnes de manière à obtenir :

6	8
A	C
C	A

Le candidat C gagne au final avec une majorité stricte de 8 contre 6. La force de ce système vient de sa prise en compte non plus seulement des premières préférences des individus mais également du reste de la liste.

Toutefois, en analysant le système en profondeur, une limite apparaît rapidement : le système n'élit pas toujours le vainqueur de Condorcet. Considérons l'élection suivante :

4	6	5
A	B	C
B	A	A
C	C	B

En utilisant le système de Hare, nous éliminerions A au premier tour puis C au second pour finalement obtenir B vainqueur.

Toutefois, en observant les préférences plus en détail un fait saute aux yeux : A est préféré par une majorité d'électeurs à tous les candidats en lice. En effet : - A est préféré à B par 9 électeurs (une majorité) - A est préféré à C par 10 électeurs (une majorité)

A est dit "vainqueur de Condorcet" et il apparaît absurde de ne pas élire A.

Le système de Borda

Nous pourrions également envisager le système suivant : considérons une élection entre trois candidats A, B et C. Chaque fois qu'un candidat apparaît à la première place il reçoit deux points et chaque fois qu'il apparaît à la deuxième place, il reçoit un point. Le candidat élu est alors le candidat ayant reçu le plus de points.

Ainsi dans le profil de préférence suivant :

4	5	7
<hr/>		
C	A	B
B	C	C
A	B	A
<hr/>		

A recevrait $2 \times 5 = 10$ points. B recevrait $1 \times 4 + 2 \times 7 = 18$ points. C recevrait $4 \times 2 + 5 + 7 = 20$ points. C serait donc élu ($20 > 18 > 10$).

Bien que séduisant à première vue, il présente un défaut majeur : les électeurs peuvent avoir intérêt à ne pas voter selon leurs préférences réelles pour favoriser l'élection de leur candidat.

Par exemple, si dans l'élection précédente les 7 électeurs $B > C > A$ avaient votés $B > \mathbf{A} > \mathbf{C}$, le profil aurait alors été :

4	5	7
<hr/>		
C	A	B
B	C	A
A	B	C
<hr/>		

A aurait alors reçu $52 + 7 = 17$ points. B : $4 + 2 \times 7 = 18$ points. C : $4 \times 2 + 5 = 14$ points B serait donc élu. Autrement dit, les électeurs $B > A > C$ ont eu intérêt à mentir sur leurs préférences réelles pour faire élire leur candidat préféré. Ils ont voté utile pour manipuler le résultat du scrutin. C'est là encore un problème.

Le système de Combs

Le système de Combs est similaire au système de Hare, à la seule différence qu'il ne s'agit pas à chaque tour d'éliminer le candidat ayant reçu le moins de première place mais le candidat ayant reçu le plus de dernière place. Est éliminé non plus le moins apprécié mais le plus détesté des candidats.

Dans le profil suivant :

5	5	3	2
A	C	B	A
B	B	D	D
C	D	C	C
D	A	A	B

A serait éliminé puisque dernier chez 8 électeurs :

5	5	3	2
B	C	B	D
C	B	D	C
D	D	C	B

Puis D serait éliminé puisque dernier chez 10 électeurs.

5	5	3	2
B	C	B	C
C	B	C	B

Et enfin C serait éliminé. Le vainqueur de l'élection est donc B.

À noter que malgré ses similitudes avec le système de Hare, tous deux ne produisent pas toujours le même résultat.

Le système de Combs, de part sa construction, présente également certaines limites. Un candidat adulé par une majorité stricte d'électeurs mais détesté par un minorité d'entre eux peut se voir éliminer dès le premier tour, bien que celui-ci semble refléter la volonté du peuple.

Ainsi, aucun des systèmes de votes présentés jusqu'à présent n'est parfait, si bien que nous pouvons nous interroger sur l'existence même d'un mode de scrutin "absolu". Le problème est d'autant plus grand que ces systèmes peuvent, pour un même profil de préférence, produire des résultats différents, ce qui empêche de les départager.

Existe-t-il donc un mode de scrutin parfait ?

Vers un mode de scrutin parfait

Pour répondre à cette question, définissons dans un premier temps ce que nous attendons d'un tel système. À quels critères devrait-il répondre pour être qualifié de "bon" ?

Nous nous plaçons ici dans un cadre plus général. Imaginons que notre système de scrutin ne délivre non plus seulement un unique vainqueur mais une liste de préférences des candidats en lice reflétant la volonté du groupe. Pour une élection entre trois candidats A, B et C, un résultat pourrait - par exemple - être $A > B > C$ si le système considère que le peuple, dans sa globalité, préfère A à B à C.

Critère de Condorcet

Le critère de condorcet peut s'exprimer de la façon suivante : "Si un candidat est préféré en duel à n'importe quel autre candidat alors celui-ci doit être élu. On le nomme alors vainqueur de Condorcet."

Critère de monotonie

Le critère de monotonie s'énonce comme suit : "Si un ou plusieurs électeurs changent leurs préférences de manière à placer un candidat plus haut, alors la préférence globale de groupe ne peut être affectée que par une montée de ce candidat ou par l'absence de changements."

Autrement dit, si un électeur décide de mieux placer un candidat dans sa liste de préférences, celui-ci ne peut pas descendre dans la préférence globale.

Le scrutin à deux tours par exemple ne respecte pas le critère de monotonie. Dans le cas suivant :

11	2	7	4	4
A	B	B	C	C
B	A	C	A	B
C	C	A	B	A

Le second tour opposerait A à B et A gagnerait 15-13. Si toutefois les deux électeurs B > A > C décidaient de modifier leurs votes en plaçant A à la première place, le profil obtenu :

13	7	4	4
A	B	C	C
B	C	A	B
C	A	B	A

verrait B éliminé au premier tour et C gagnerait 15-13. Autrement dit, améliorer la place de A dans les votes lui a porté préjudice.

Critère de Pareto ou Unanimité

Le critère de Pareto (ou critère d'unanimité) s'énonce comme suit : "Si tous les électeurs préfèrent A à B, alors le système ne peut pas dire que la population préfère B à A".

Autrement dit, si la population est unanime sur le fait que A est préférable à B, alors cette préférence doit se manifester au niveau de la préférence globale.

Critère de Strategy-Proofness

Un système est dit strategy proof (ou non manipulable) si un électeur ne peut jamais améliorer les chances de son candidat favori en votant de manière stratégique. Un électeur obtiendra toujours le meilleur résultat en votant selon ses préférences réelles. C'est donc dans l'intérêt de l'électeur de voter selon ses préférences réelles.

Critère D'indépendance aux alternatives non pertinentes

Le critère d'indépendance aux alternatives non pertinentes (ou IIA pour independence of irrelevant alternatives) s'énonce comme suit : "Supposons que l'élection conclut que la population préfère dans l'ensemble A à B, et supposons que certains électeurs venaient à changer leurs votes. Si aucun électeur ne change les positions relatives de A et de B (tous ceux qui plaçaient initialement A au dessus de B continuent de le faire et de la même manière pour ceux qui plaçaient B au dessus de A) alors le système doit toujours considérer que A est préféré à B".

Critère de non-imposition

"Le critère de non-imposition signifie que toutes les configurations possibles de préférences globales peuvent être obtenues, pour peu que l'on choisisse le bon profil de préférence."

En pratique l'indépendance aux alternatives non pertinentes, la monotonicité et la non-imposition impliquent le critère de pareto, si bien que tout théorème valable dans le cas du critère de pareto le sera également dans le cas où l'iiia, la monotonicité et la non-imposition sont simultanément valables.

Ainsi, un système parfait devrait à minima respecter ces critères qui semblent raisonnables.

Le théorème d'impossibilité de Arrow

Mais la réalité nous rattrape et Kenneth Arrow (économiste né en 1921, mort en 2017 prix nobel d'économie) publie en 1951 son théorème d'impossibilité qui s'énonce comme suit :

“Quand il y a trois alternatives ou plus, la seule manière de combiner des préférences individuelles en une préférence de groupe avec unanimité et indépendance aux alternatives non pertinentes est la dictature.”

Nous donnerons ici la démonstration proposée par John Geanakoplos dans son article Three brief proofs of Arrow’s impossibility Theorem. Il est intéressant d’observer qu’il ne s’agit pas là de la preuve originale proposée par Arrow mais d’une re-démonstration. Il est en effet fréquent en mathématiques de “re-démontrer” des théorèmes de manière simple ou plus élégante.

Supposons ici que nous ayons 3 candidats { A, B, C } et N électeurs (le raisonnement pouvant se généraliser à un nombre plus important d’électeurs).

Dans le cadre de la théorie des systèmes de votes, dire d’un électeur qu’il est dictateur de l’élection signifie qu’il dicte par son vote la volonté du groupe. Autrement dit, la volonté du groupe sera la volonté de l’électeur / la préférence globale du groupe sera la préférence individuelle de l’électeur, quelle que soit la répartition des votes. Par analogie, on dit d’un électeur qu’il est dictateur de pair s’il dicte la préférence entre A et B.

Nous cherchons donc ici à démontrer que dans tout système de vote respectant le critère d’unanimité et l’indépendance aux alternatives non pertinentes il existe un électeur n^* tel que n^* est dictateur.

Avant de démontrer le théorème de Arrow en tant que tel, nous allons démontrer un théorème “intermédiaire” que nous utiliserons par la suite. Un tel théorème utilisé comme outil pour en démontrer un autre est appelé lemme. Au même titre que l’homme fabrique des outils pour mieux construire sa maison, le mathématicien fabrique des lemmes pour mieux construire ses théorèmes.

Le lemme en question, dit “lemme extremal” s’énonce comme suit : > “Considérons un candidat b quelconque. Dans n’importe quel profil de préférences où chaque électeur place b en première ou en dernière position dans sa liste de préférence, la préférence collective doit de même placer b en première ou en dernière position”

Autrement dit, un profil de préférence de la forme :

10	4	5	8
<hr/>			
b	b
...
...	b	b	...
<hr/>			

Où b est systématiquement placé dans une position extrême, le système de vote, pour peu qu’il respecte l’iiia et le critère d’unanimité, placera également b dans une position extrême, i.e premier ou dernier.

Pour démontrer cela, nous raisonnons par l'absurde. Supposons donc que ce ne soit pas le cas, qu'il existe un profil de préférences où b est systématiquement placé dans une position extrême mais que ce ne soit pas le cas au niveau de la préférence collective, autrement dit qu'il existe deux alternatives a et c telles que $a > b > c$. Par iia, échanger a et c n'a d'effet ni sur la préférence $a > b$ ni sur la préférence $b > c$. Nous pouvons dès lors échanger les ac de sorte à placer tous les c au-dessus des a. Par unanimité, a étant maintenant toujours préféré à c, la préférence collective considère de même que $a > c$. Mais nous avons toujours (par iia) $a > b$ et $b > c$ donc $a > c$ (par transitivité). Nous avons alors d'une part $a > c$ et d'autre part $c > a$, un résultat absurde qui invalide donc l'hypothèse selon laquelle le théorème serait faux. Le théorème est donc vrai.

Entrons maintenant dans le cœur de la démonstration. Nous partirons d'un profil particulier que nous généraliserons progressivement à l'ensemble des profils possibles.

Imaginons donc un profil de préférence où le candidat b est systématiquement placé en dernière position.

1	2	...	$n-1$	n
...
b	b	b	b	b

(ce tableau représente les listes de préférences des n électeurs, numérotés de 1 à n)

Puisque chaque candidat est unanimement préféré à b, b est nécessairement placé dernier dans la liste de préférences collectives.

Partant de ce profil de préférences, nous allons, candidat par candidat, faire passer b de la dernière à la première position, en laissant le reste des préférences inchangées.

Ainsi au premier tour nous obtiendrons

1	2	...	$n-1$	n
b
...	b	b	b	b

Puis au second tour

1	2	...	$n-1$	n
b	b
...	...	b	b	b

etcetera, etcetera jusqu'au nième tour ou nous obtiendrons finalement

1	2	...	n-1	n
b	b	b	b	b
...

b étant préféré unanimement dans ce profil à n'importe quel autre candidat en lice, il est (par critère d'unanimité) nécessairement vainqueur.

Puisque b reste tout au long du processus dans une position extrême dans les préférences individuelles, il l'est également - d'après le lemme extremal énoncé précédemment - dans la préférence collective. Il existe donc un électeur n^* responsable du "basculement" du candidat b de la dernière à la première position. Autrement dit, le profil suivant (que nous appellerons profil (I))

1	2	...	n^*	...	n-1	n
b	b	b
...	b	b	b	b

b est donné perdant tandis que dans le profil suivant (que nous appellerons profil (II))

1	2	...	n^*	...	n-1	n
b	b	b	b
...	b	b	b

b est donné gagnant.

n^* est donc pivot dans la mesure où son choix va dicter la position de b dans le résultat de l'élection.

Nous allons maintenant montrer que ce n^* est dictateur de pair pour toutes les paires n'impliquant pas b et quel que soit le profil considéré.

Considérons donc un couple de candidats a et c et choisissons l'un d'entre eux, disons a.

Nous construisons alors un profil (III), similaire au profil (II) à la différence que n^* place a au-dessus de b de sorte que $a > b > c$. Puisque les positions relatives de a et c sont les mêmes que dans le profil (I), la préférence collective est (par iia) $a > c$. De même, puisque les positions relatives de b et c sont les mêmes que dans le profil (II), la préférence collective est (par iia) $c > b$. Par transitivité nous avons donc $a > c > b > a$. Puisque la préférence entre a

et b est fixée, les électeurs peuvent placer c où bon leur semble sans modifier la relation ab, dévoilant ainsi l'ensemble des profils possibles (a et b étaient déjà placés de façon arbitraire et c l'est désormais également). Ainsi, quelque soit le profile considéré, n* dicte la préférence globale entre toutes les paires n'impliquant pas c. Il ne reste dès lors plus qu'à prouver que c'est également le cas pour celles impliquant c. Il suffit pour cela de réitérer le raisonnement et d'observer qu'un dictateur de paires impliquant c dicte également des paires dont n* est dictateur, de sorte que ces derniers ne sont en fait qu'un unique dictateur, l'élection présente donc bien un dictateur. (QED)

Nous pouvons toutefois noter deux choses. Premièrement, le théorème de Arrow ne s'applique que dans le cas où la préférence collective s'exprime comme une liste de préférence. Qu'en est-il alors des modes de scrutins qui n'ont pour objectif que d'élire un seul représentant ? Deuxièmement, le critère d'indépendance aux alternatives non pertinentes peut être critiqué puisqu'il ne prend en compte que la position relative des candidats dans les votes et non l'écart qui les sépare.

Le théorème d'impossibilité de Gibbard-Satterthwaite

C'est dans ce contexte que Allan Gibbard et Mark Satterthwaite démontrent séparément dans les années 70 un théorème plus "fort" que le théorème de Arrow, le théorème dit de Gibbard-Satterthwaite (que nous admettrons sans démonstration) :

"Pour trois alternatives ou plus, le seul scrutin qui respecte l'unanimité et le critère de strategy-proofness est la dictature"

Ainsi, il n'existe aucun mode de scrutin ordinal (c'est à dire qui se base sur une liste de préférence des électeurs) qui puisse respecter conjointement tous ces critères.

Vers le jugement majoritaire

Malgré tout, ces théorèmes ne s'appliquent qu'au cas des scrutins ordinaux et nous pouvons parfaitement contourner cette difficulté en mettant en oeuvre un autre système pour exprimer les préférences de l'électorat.

L'une d'entre elle consiste à noter les candidats selon une certaine échelle. Dans une élection à trois candidats A, B et C, un électeur pourrait par exemple attribuer les notes suivantes aux candidats - 3/20 à A - 13/20 à B - 17/20 à C

Deux problèmes surgissent alors immédiatement. Premièrement, chacun possède sa manière de noter. Un excellent candidat pourrait, chez certain électeurs, mériter une note de 15/20, tandis que chez d'autre il s'agirait d'un 18/20. Cela risquerait alors de fausser les résultats et il est donc nécessaire de "normaliser" les notes (ce qui a pour effet de dénaturer les voix). Deuxièmement, un électeur aura tout intérêt à exagérer ses notes pour favoriser ses candidats préférés au détriment des autres.

L'électeur ci-dessus aurait alors tout intérêt à voter : - 0/20 à A - 0/20 à B - 20/20 à C

Il s'agit là encore d'un vote stratégique. Une manière de régler ce problème est alors de considérer non pas la moyenne des notes obtenues mais la répartition de ces dernières, en utilisant par exemple une médiane.

C'est ainsi que naît en 2007 le "jugement majoritaire", un système de vote qui constitue encore aujourd'hui le meilleur système de vote connu. Le jugement majoritaire ne respecte pas tous les critères mais une grande partie d'entre eux et les quelques critères restants sont, en règle générale, bien respectés. Il n'est en particulier pas totalement strategy-proofness.

En conclusion, les systèmes de votes mis en oeuvre dans nos sociétés présentent de nombreuses limites : dilemme du vote utile, indépendance aux alternatives non pertinentes etc. Bien que les théorèmes d'impossibilité de Arrow et de Gibbard-Satterthwaite démontrent qu'un système parfait n'existe pas, nous pouvons largement rendre nos systèmes de votes "moins pire" pour davantage représenter la volonté du peuple et espérer qu'un jour le mot démocratie puisse enfin prendre tout son sens.

Annexe

Voici quelques ressources pour approfondir le sujet.

Chacune d'entre elle est précédée d'un certain nombre d'étoiles suivant sa complexité, allant d'une étoile (*) pour les plus faciles à trois étoiles (***) pour les plus complexes :

- (*) Une excellente série de vidéos sur les mathématiques de la démocratie
La démocratie sous l'angle de la théorie des jeux - Science4All
- (*) Un vidéo de présentation du jugement majoritaire (entre autre) Réformons l'élection présidentielle ! — ScienceEtonnante
- (**) Un papier de l'association mathématique du Québec pour approfondir la théorie des systèmes de vote Mathématiques Électorales — AMQ
- (***) L'article original de la démonstration du théorème de Arrow présentée ci-dessus Three Brief Proofs of Arrow's Impossibility Theorem — John Geanakopolos
- (***) Un livre très complet (et relativement court) sur la théorie des systèmes de votes The Mathematics of Elections and Voting — W.D Wallis