FAQ fr.sci.maths

Les utilisateurs de fr.sci.maths

Version 2.13

Introduction.

`fr.sci.maths` est un groupe de discussion destiné à recueillir les discussions en français concernant les mathématiques. Ce document rassemble les questions qui ont été fréquemment posées dans ce forum.

Nouveautés.

Liste des derières modifications :
- Version 2.13
- 3.1 Article sur les carrés magiques.
- Version 2.12 Modifications mineures.
- Version 2.11
  - 1.2.4 Deux nouvelles démonstration erronnées de l'égalité « 2=1 » ;
  - Correction du problème des polices floues dans la version PDF (utilisation de polices vectorielles plutôt que de polices *bitmaps*)
- Version 2.1
  - 1.9 Les deux échelles ;
  - 1.10 La cuve de vin ;
  - 5.2 Qu'est-ce que le nombre e ;
  - 7.3 L'algorithme de CORDIC sur les calculatrices.

Table des matières

Chapitre 1

Mathématiques récréatives.

1.1 Est-ce que $0,9999\ldots = 1$?

Ce que l'on note $0.9999999999\ldots$ (avec les points de suspension) désigne un *nombre* qui se termine par une infinité de 9. Et donc, est-ce que $0.999\ldots$ est égal à 1 ? La réponse est *oui* !

Voici 5 arguments pour vous en convaincre. Les 3 premiers n'ont absolument aucune rigueur et ne peuvent pas être considérés comme des démonstrations mathématiques, mais ils sont plus simples et assez convaincants.

1.1.1 Trois preuves élémentaires.

On part de : $1/3 = 0,33333\ldots$ On multiplie par 3 des deux côtés : $3 \times (1/3) = 3 \times 0,33333\ldots$ Ce qui donne : $1 = 0,99999\ldots$

On pose $x = 0,99999\ldots$ On multiplie par 10 des deux côtés : $10x = 9,99999\ldots$ On soustrait les deux expressions côté par côté : $10x - x = 9,99999\ldots - 0,99999\ldots = 9,00000\ldots$ Donc $9x = 9$, c'est-à-dire $x = 1$, d'où $0,99999\ldots = 1$.

Un argument très court se déduit du fait suivant : « si 2 nombres réels sont différents, alors il en existe au moins un 3ème entre les deux, différent des deux autres ». (ce troisième nombre peut être la moyenne entre les deux). Or, on ne peut pas intercaler de nombre entre $0,99999\ldots$ et 1 ; ils sont donc égaux.

1.1.2 Démonstration par les séries

Pour les arguments plus rigoureux, il faut commencer par définir proprement ce qu'est $0,99\ldots$

En écrivant $0,99999\ldots = 0,9 + 0,09 + 0,009 + \ldots$, on définit $0,9999\ldots$ comme une série géométrique (c'est-à-dire une somme dont chaque terme est égal au précédent multiplié par une constante, ici 1/10 - on dit que c'est une série géométrique de raison 1/10), et on écrit :

$$0,99999\ldots = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{9}{10^i}$$

On peut facilement montrer que la somme des n premiers termes d'une série géométrique de raison q et de premier terme a vaut :

$$S_n = a \times \frac{1-q^n}{1-q}$$

Cette somme tend vers une limite pour n tendant vers l'infini si et seulement si q est strictement plus petit que 1, et cette limite est alors :

$$S = \frac{a}{1-q}$$

Ici, $a = 0,9$, $q = 1/10$, ce qui est plus petit que 1, donc

$$S = \frac{0,9}{1-1/10} = 0,9 \times \frac{10}{9} = 1$$

Donc $0,99999\ldots = 1$

1.1.3 Démonstration par la limite.

L'argument le plus direct est de vérifier directement, à partir de la définition de la limite, que 1 est la limite pour n tendant vers l'infini de la série

$$S_n = \sum_{i=1}^{n} \frac{9}{10^i}$$

Cela signifie qu'à condition de prendre suffisamment de termes dans la série, on peut s'approcher d'aussi près de 1 que l'on veut (c'est-à-dire rendre la différence $|1 - S_n|$ aussi petite que l'on veut).

Mathématiquement, cette définition de limite s'écrit : $\forall \epsilon > 0$, il existe n_0 tel que pour tout $n \geq n_0$, on a $|1 - S_n| < \epsilon$.

En calculant

$$\left| 1 - \sum_{i=1}^{n} \frac{9}{10^i} \right| = \frac{1}{10^{n+1}}$$

on voit facilement que si n (nombre de termes) est suffisamment grand, alors notre somme peut s'approcher d'aussi près que l'on veut de 1, puisque leur différence, $1/(10^{n+1})$ devient de plus en plus petite quand n augmente.

Pour être plus précis, si on se donne ϵ, la différence maximale que l'on s'autorise, alors il suffit de prendre[1] :

Soit $n_0 > -\log(\epsilon) - 1$. Si $n > n_0$, on aura alors :

$$\left| 1 - \sum_{i=1}^{n} \frac{9}{10^i} \right| = \frac{1}{10^{n+1}} < \epsilon$$

la condition est respectée, donc la limite vaut 1, et $0,99999\ldots = 1$

1.2 J'ai réussi à montrer que $2 = 1$.

Deux petites démonstrations, fausses, bien entendu, mais qui peuvent induire en erreur. N'oublions pas le vieil adage latin :

ex falsus, quod libet[2]

1.2.1 Grâce aux polynômes.

La démonstration (fausse)

Soit a et b deux nombres réels non nuls tels que $a = b$. Alors $a^2 = ab$ (on multiplie par a des deux côtés) D'où $a^2 - b^2 = ab - b^2$ (on soustrait b^2 des deux côtés) D'où $(a - b)(a + b) = b(a - b)$ (on met en évidence $a - b$) D'où $a + b = b$ (on simplifie par $a - b$) D'où $2b = b$ (puisque $a = b$) D'où $2 = 1$ (puisque b est non nul).

Explications

Ici, l'erreur vient de la simplification par $(a - b)$ qui est nul. On a divisé par zéro, ce qui est impossible. Bien souvent, ces démonstrations trouvent leur erreur dans une division par zéro.

1.2.2 Par la dérivée.

La démonstration (fausse)

Soit x appartenant à $\mathbb{R}^*$ On a la relation : $x^2 = x + x + x + \cdots + x$, x fois. On dérive : $2x = 1 + 1 + 1 + 1 + \cdots + 1$, x fois. C'est-à-dire : $2x = x$. Et

[1] log représentant le logarithme en base 10

[2] de quelque chose de faux, on peut trouver n'importe quoi

comme $x \neq 0$, on obtient $2 = 1$.

Explications

L'erreur vient de la définition de la dérivée. « $x^2 = x + x + x + \cdots + x$, x fois »n'a de sens que si x est entier. Or, pour dériver en un point, il faut considérer un voisinage de ce point (grosso-modo un intervalle ouvert contenant ce point) qui, forcément, sera loin de ne contenir que des entiers.

Par exemple, si on essaye d'appliquer cela en $x = 3$:

- Il est exact que $3^2 = 3 + 3 + 3$.
- Par contre, pour x proche de 3 mais $x \neq 3$, $x^2 \neq 3x$
- la dérivée en x d'une fonction ne dépend pas de la valeur de la fonction en x mais de son comportement local et le comportement de x^2 en 3 est très différent de celui de $3x$.
  De plus, si tu dérives $x + \cdots + x$ (x fois), tu ne différencies pas le « x fois », que tu considères donc comme une constante. Quand j'étais au lycée on m'avait posé ce problème et j'avais trouvé un moyen (tordu et absurde) de retomber sur ses pattes, en ajoutant « $x + \cdots + x$ ('dérivée de x' fois) », comme ça on a aussi dérivé le « x fois ».

1.2.3 En utilisant les puissances.

La démonstration (fausse)

$$-1 = (-1)^1 = (-1)^{1/1} = (-1)^{2/2} = ((-1)^2)^{1/2} = 1^{1/2} = 1$$

Explications

L'erreur vient du fait que l'on néglige, ici, la définition de la puissance. En effet, on ne peut pas écrire a^q pour q rationnel et a réel négatif.

Plus précisément, on peut expliquer le phénomène de la manière suivante.

Définition : *Dans un ensemble stable par la loi multiplicative (pour être le plus général possible), on note (pour un élément a de l'ensemble et pour b entier naturel non nul) a^b pour désigner a multiplié b fois par lui-même.*

Définition : *Dans le cas ou on l'on veut mettre un rationnel en exposant, il faut utiliser la définition de la puissance par l'exponentielle : pour a réel strictement positif et b réel, $a^b = \exp(b \ln(a))$.*

On a en fait le droit d'écrire $(-1)^{2/2}$. Mais pas d'utiliser la loi $a^{bd} = (a^b)^d$, car pour utiliser cette loi de composition, il faut, du fait que d est

ici rationnel,prendre la définition avec l'exponentielle, qui interdit à a d'être négatif.

On a bien la loi de composition $a^{bd} = (a^b)^d$ pour la définition 1 et la définition 2, mais on peut l'appliquer (pour a, b et d réels) :

- Selon la définition 1, seulement si b et d entiers naturels.
- Selon la définition 2, seulement si a est strictement positif.

1.2.4 Équivalence de suites

La démonstration (fausse)

Définition : *Soit* (u_n) *et* (v_n) *deux suites numériques. On dit que* (u_n) *et* (v_n) *sont* équivalentes *si et seulement si la suite* $(u_n - v_n)$ *est négligeable devant* (u_n)*. En abrégé,* $(u_n) \sim (v_n)$ *si et seulement si* $\forall \varepsilon > 0$, $\exists N$, $n > N \Rightarrow |u_n - v_n| < \varepsilon |u_n|$.

On peut aisément montrer la propriété suivante :

Propriété : *Si* (u_n) *n'est jamais nulle à partir d'un certain rang,*

$$(u_n) \sim (v_n) \Leftrightarrow \lim_{n\to\infty} \frac{u_n}{v_n} = 1$$

Donc $n \sim n+1$ (car $\lim \frac{n+1}{n} = 1 + \lim \frac{1}{n} = 1$), mais aussi $n+1 \sim n+2$, $n+2 \sim n+3$, ..., $2n-1 \sim 2n$. Par transitivité de cette relation d'équivalence on obtient donc $n \sim 2n$, ce qui veut dire que $\lim \frac{2n}{n} = 1$. Or $\lim \frac{2n}{n} = 2$, donc $2 = 1$.

Explications

L'erreur provient du fait que la transitivité est valable pour un nombre fini d'équivalences mais plus si leur nombre peut être arbitrairement grand. Or ici, on utilise n équivalences ; et simultanément, on fait tendre n vers l'infini.

1.2.5 Par construction fractale

La démonstration (fausse)

Soit les trois points (dans une base orthonormée) $A = P_0(0) = (0, 0)$, $P_0(1) = (1, 1)$, $B = P_0(2) = (2, 0)$.

Une simple application du théorème de Pythagore montre que $Longueur(P_0(0), P_0(1), P_0(2$ $2\sqrt{2}$[3]. On note cette longueur δ_0.

Notons $P_1(1)$ le milieu de $[P_0(0), P_0(1)]$, $P_1(3)$ le milieu de $[P_0(1), P_0(2)]$ et $P_1(2)$ la projection orthogonale de $P_0(1)$ sur $[P_0(0), P_0(1)]$. Pour avoir des notations cohérentes, on pose $P_1(0) = P_0(0)$ et $P_1(4) = P_0(2)$. Cela revient en fait à "plier" en deux le triangle isocèle. On a toujours $\delta_1 = Longueur(P_1(0), P_1(1), P_1(2), P_1(3), P_1(4)) = Longueur(P_0(0), P_0(1), P_0(2)) = 2\sqrt{2}$.

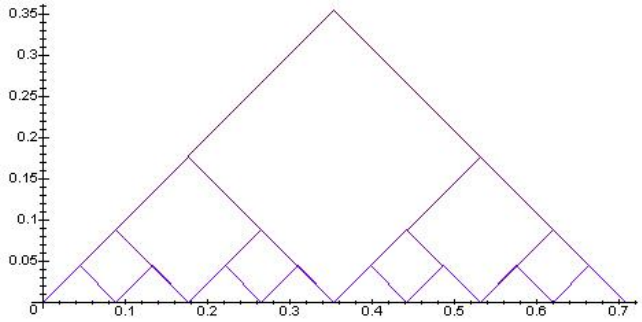

FIG. 1.1 – On réitère le pliage n fois

Plus généralement, à l'étape n, $n \geq 1$, pour tout entier k, $0 \leq k \leq 2^n$:

– si k est impair, $P_n(k)$ est le milieu de $[P_{n-1}(k-1), P_{n-1}(k)]$;

– si k est pair, $P_n(k)$ est la projection orthogonale de $P_{n-1}(\frac{k}{2})$ sur $[A, B]$.

et on note $\delta_n = Longueur(P_n(0), P_n(1), P_n(2), \ldots, P_n(2^n))$. Voir 1.1

Il est clair que quand n tend vers l'infini, les segments formés par les points P_n tendent vers le segment de droite $[A, B]$. Donc la longueur de ce segment tend vers la $Longueur(A, B) = 2$. En abrégé $\lim_{n\to\infty} \delta_n = 2$.

D'autre part, il est évident que $\delta_n = \delta_{n-1} = \ldots = \delta 1 = \delta 0 = 2\sqrt{2}$.

On aboutit donc à $2\sqrt{2} = 2$, soit $sqrt(2) = 1$, ou encore $2 = 1$.

Explications

L'erreur commise vient de l'utilisation d'un théorème intuitivement correct qui serait « la limite de la longueur d'un polygone à n côtés est égal à la longueur de la limite de ce polygone (quand n tend vers l'infini) ». Cet énoncé est complètement *faux*. Ici, ce n'est pas parce que la suite des segments tend

[3] $Longueur(A_1, A_2, \ldots, A_n)$ désigne la distance (euclidienne) entre les points A_1 et A_2, plus la distance entre A_2 et A_3, ... plus la distance entre A_{n-1} et A_n.

vers $[AB]$ (ce qui est effectivement le cas), que la longueur de cette suite de segments tend vers la longueur AB.

Remarque : En fait, cela est du au fait que la longueur entre deux points est définie comme une intégrale, et qu'il n'est pas possible en général d'écrire

$$\lim_{n\to infty} \int f_n = \int \lim_{n\to infty} f_n$$

Il existe cependant des théorèmes, dits d'interversion qui autorise cette égalité quand les focntions f_n vérifient certaines propriétés.

1.3 Zéro puissance zéro égal un $(0^0 = 1)$.

Par convention, les mathématiciens posent que zéro à la puissance zéro est égal à un $(0^0 = 1)$. Mais, si l'on recherche pourquoi une telle chose, on se retrouve face à un grand nombre de problèmes. N'oubliez pas, ce n'est qu'une convention, et il peut être utile de poser $0^0 = 0$.

1.3.1 Approche topologique.

Définition par continuité. On prendra comme définition de la puissance, la formule : pour tous nombres réels x, y (avec $x > 0$) $x^y = \exp(y \ln(x))$. Une approche par continuité pose problème. En effet, on peut choisir trois fonctions $\{x \to x^0\}$; $\{y \to 0^y\}$ ou $\{x \to x^x\}$ pour approcher par continuité (en passant à la limite) la valeur 0^0.

Or l'on a, pour tout nombre réel x non-nul $x^0 = \exp(0 \ln(x)) = 1$. En prolongeant, par continuité, cette fonction quand x tend vers zéro, on trouve : $0^0 = 1$.

Si pour tout nombre réel y non-nul, on prolonge par continuité la fonction $y \to x^y$ pour $x = 0$. Et quand y tend vers zéro, on trouve : $0^0 = 0$.

En outre, pour tout nombre réel x strictement positif, on a

$$x^x = \exp(x \ln(x))$$

On cherchera, alors, la limite en zéro par valeurs positives, notée 0^+. Et l'on a donc :

$$\lim_{x\to 0^+} x^x = \lim_{x\to 0^+} \exp(x \ln(x)) = 1$$

car $\lim_{x\to 0^+} x \ln(x) = 0$ et $\exp(0) = 1$.

La question se pose alors : quelle valeur choisir pour 0^0 ?

Problèmes liés à cette définition. Soit f une fonction continue définie sur un intervalle I contenant 0 et telle que pour tout x appartenant à I, $f(x) > 0$. Soit g une fonction définie sur un intervalle J contenant 0. On supposera également que

$$\lim_{x\to 0^+} f(x) = \lim_{x\to 0^+} g(x) = 0$$

On peut alors écrire pour tout x appartenant à $I \cap J$

$$f(x)^{g(x)} = \exp(g(x)\ln(f(x)))$$

On constatera aisément que l'on est en présence d'une forme indéterminée et donc qu'en choisissant convenablement f et g on peut trouver n'importe quelle valeur réelle positive finie, en passant à la limite par valeurs supérieures.

Par exemple, soit A un réel strictement positif. Il suffit de choisir $f(x) = \exp(-A/x)$ et $g(x) = x + x^2$ On a bien les conditions voulues et on a

$$f(x)^{g(x)} = \exp((x+x^2)\ln(\exp(-A/x)) = \exp(-A(x+x^2)/x)$$

C'est-à-dire : $f(x)^{g(x)} = \exp(-A(1+x))$. Et de là :

$$\lim_{x\to 0^+} f(x)^{g(x)} = \lim_{x\to 0^+} \exp(-A(1+x)) = \exp(-A)$$

Ainsi en choisissant convenablement A, on peut trouver comme limite n'importe qu'elle valeur réelle comprise entre 0 et 1. En gardant f et en prenant $g(x) = -(x+x^2)$ on va trouver une limite supérieure à 1.

On notera, néanmoins, que si l'on choisit une fonction $\{u : \mathbb{R} \to \mathbb{R}^+\}$ telle que sa limite soit nulle quand x tend vers zéro par valeurs positives, on a alors : $u(x)^{u(x)} = \exp(u(x)\ln(u(x)))$ et par composition des limites on trouve :

$$\lim_{x\to 0^+} u(x)^{u(x)} = \lim_{x\to 0^+} \exp(u(x)\ln(u(x))) = \exp(0) = 1$$

En effet, l'on a bien : $\lim_{x\to 0^+} u(x)\ln(u(x)) = \lim_{x\to 0^+} X\ln(X)$ en effectuant le changement de variable $X = u(x)$.

1.3.2 Approche algébrique.

On peut, pour essayer de comprendre pourquoi $0^0 = 1$, revenir à la définition donnée dans les petites classes de la fonction puissance.

Soit maintenant comme définition de la puissance : $x^n = x \times x \times \cdots \times x$ (n fois)

Le nombre réel x est multiplié n fois par lui-même avec n un nombre entier. Alors cette définition nous amène à la relation suivante :

$$\text{Pour tous entiers } n, m \text{ on a : } x^{n+m} = (x^n)(x^m) \tag{1.1}$$

Si on prend $m = 0$ et n différent de zéro, alors on a $x^n = (x^0)(x^n)$. Si x est différent de zéro, alors cela implique que $x^0 = 1$. Il est alors très tentant d'étendre la relation à $x = 0$, et donc : $0^0 = 1$. On notera, cependant, que si l'on pose $0^0 = 0$, alors la relation reste vraie.

De plus la relation (1.1) implique la relation suivante :

$$x^{nm} = x^{nm} \tag{1.2}$$

Encore une fois, si on prend $n = 0$ dans la relation précédente on trouve $x^{0^m} = x^0$. Et il est encore très tentant, pour x non nul, de prendre $x^0 = 1$ et d'étendre cette relation à $x = 0$. Mais on peut tout à fait prendre comme convention $0^0 = 0$, sans que la relation en soit modifiée.

1.3.3 Approche ensembliste.

Il faut d'abord définir qu'est ce qu'on entend par addition, multiplication et puissance.

Qu'est ce que l'addition ? On prend deux ensembles disjoints A et B ayant chacun $|A|$ et $|B|$ éléments (lire 'cardinal' de A et 'cardinal' de B). Et bien l'addition de $|A|$ et $|B|$, c'est ce qu'on obtient en mettant ensemble les éléments de A et B : $|A| + |B| = |A \cup B|$

Qu'est ce que la multiplication ? Si on veut dire $3|A|$, ça veut dire qu'on compte trois fois chaque élément de A. On peut dire ça en disant que pour chaque x dans A on compte $(1, x)$, $(2, x)$ et $(3, x)$.

En clair, on compte les éléments de $\{1, 2, 3\} \times A$. Il est facile alors de voir que ce qu'on entend par multiplication, $|A||B| = |A \times B|$

Qu'est ce alors que l'exponentielle ? Calculer $|A|^n$, c'est donc calculer $|A \times A \times \ldots \times A|$ (n fois), c'est à dire dénombrer tous les n-uplets d'éléments de A.

Pour former un n-uplet, on choisit un premier élément dans A, x_1, puis un deuxième x_2, puis ..., puis un n-ième x_n. Et c'est fait.

C'est à dire qu'on choisit une application (ici notée $\times$) de $\{1, 2 \ldots, n\}$ dans A. On peut alors dire que $|A|^{|B|} = |A^B|$ où A^B est l'ensemble des applications de B dans A. $|A^B|$=nombre d'applications de B dans A.

Et 0^0 dans tout ça ? Fort de ces définitions, on peut se demander ce qu'il se passe quand A et B sont vides. Et bien, il existe exactement une application de l'ensemble vide dans lui-même, c'est l'identité. Il apparaît clairement que $0^0 = 1$.

1.3.4 Conclusion.

On constate donc que l'approche par continuité, bien qu'apparemment la plus simple, ne conduit qu'à des contradictions. On ne peut donc pas définir 0^0 par des fonctions continues et par passage à la limite.

Ainsi, par convention 0^0 est en général égal à 1, parce que cela arrange nombre de formules, notamment celles sur les polynômes.

Mais ce n'est pas une généralité. Il peut en effet être plus utile de poser que $0^0 = 0$ dans certains cas. Nous avons vu que cela n'amène aucune contradiction dans la théorie algébrique. N'oubliez pas : c'est juste une convention d'utilité.

On notera que le forum anglophone `sci.maths` possède également une *Foire Aux Questions* sur le sujet, disponible à l'adresse : `http://www.cs.unb.ca/~alopez-o/math-faq/mathtext/node14.html`.

1.4 Pièces et balance.

1.4.1 Énoncé.

Enoncé : *On vous donne 12 pièces qui paraissent identiques, dont l'une est contrefaite et a un poids légèrement différent des autres (vous ne savez pas si la pièce est plus lourde ou plus légère).*

On vous donne une balance de Roberval, qui vous permet de mettre le même nombre de pièces de chaque côté et d'observer quel côté (s'il y en a un) est plus lourd.

Comment identifier la pièce contrefaite et déterminer si elle est plus lourde ou plus légère, en 3 pesées ?

1.4.2 Solution.

Martin Gardner[4] a donné une jolie solution à ce problème. Supposez que vous avez le droit à P pesées. écrivez les 3^P chaînes possibles de longueur P ayant pour caractères "0", "1" et "2". éliminez les 3 chaînes comportant uniquement un caractère répété P fois.

Pour chaque chaîne, trouvez le premier caractère différent du caractère le précédant. Considérez ce couple de caractères. Si ce couple n'est pas 01, 12 ou 20, éliminez cette chaîne. En d'autres termes, seules les chaînes de la forme $0*01.*$, $1*12.*$ ou $2*20.*$ (expressions rationnelles) sont acceptées.

Il doit vous rester $(3^P-3)/2$ chaînes. C'est le nombre de pièces que vous pouvez contrôler en P pesées. Associez donc chaque pièce à une chaîne de P caractères.

Effectuez P pesées comme suit : Pour la pesée I, mettez d'un côté toutes les pièces ayant un 0 dans la chaîne en position I, et mettez de l'autre côté toutes les pièces ayant un 2 dans la chaîne en position I.

Si le côté avec les 0 en position I est plus lourd, écrivez un 0. Si c'est l'autre côté qui est plus lourd, écrivez un 2. Sinon, écrivez un 1.

Après P pesées, vous avez écrit une chaîne de P caractères. Si votre chaîne correspond à une des pièces, alors c'est cette pièce qui est contrefaite, et elle est plus lourde. Sinon, changez chaque 2 en 0 et chaque 0 en 2 dans votre chaîne. Votre chaîne correspondra alors à l'une des pièces, et cette pièce est plus légère que les autres.

Notez que si vous devez seulement identifier la pièce contrefaite, mais pas déterminer si elle est plus lourde ou plus légère, vous pouvez contrôler $(3^P-3)/2+1$ pièces. étiquetez la pièce supplémentaire par la chaîne contenant uniquement des 1, et utilisez la méthode ci-dessus.

Notez aussi que vous pouvez contrôler $(3^P-3)/2+1$ pièces si vous devez déterminer si la pièce contrefaite est plus lourde ou plus légère, pourvu que vous ayez une pièce de référence, dont vous savez qu'elle a le poids correct. Vous faites ceci en étiquetant la pièce supplémentaire par la chaîne contenant uniquement des 2. Cette pièce est placée toujours du même côté, et ce plateau contient une pièce de plus que l'autre. Alors, placez la pièce de référence de l'autre côté, à chaque pesée.

[4]Traduction de logic/weighing/balance des archives de rec.puzzles
http://alabanza.com/kabacoff/Inter-Links/puzzles.html

Il est très facile de prouver que ceci marche, une fois que vous avez remarqué que la méthode de construction des chaînes assure qu'à chaque position, 1/3 des chaînes ont un 0, 1/3 ont un 1, et 1/3 ont un 2, et que si une chaîne est dans la liste, alors celle obtenue en remplaçant chaque 0 par un 2 et chaque 2 par un 0 n'y est pas.

Si vous savez déjà que la pièce contrefaite est plus lourde (ou plus légère), vous pouvez contrôler 3^P pièces. Avec P pesées, il ne peut y avoir que 3^P combinaisons d'équilibre, plateau de gauche plus lourd et plateau de droite plus lourd.

L'algorithme est dans ce cas : Partagez les pièces en 3 groupes de même taille A, B et C. Pesez A avec B. Si un plateau tombe, il contient la pièce lourde, sinon cette pièce est dans le groupe C. Si la taille de votre groupe est 1 vous avez trouvé la pièce, sinon faites une recurrence sur le groupe contenant la pièce lourde.

1.5 Les âges du capitaine.

1.5.1 Énoncé.

Enoncé : *Le capitaine dit à son fils : « La cabine n°1 abrite M. Dupont et ses deux filles. Le produit de leurs trois âges est* 2450 *et la somme de leurs trois âges est égale à 4 fois le tien. Peux-tu trouver les âges des trois passagers ? »Après un instant, le fils répond : « Non, il me manque une donnée. »Le capitaine ajoute alors : « Je suis plus âgé que M. Dupont. »Le fils du capitaine en déduit aussitôt les trois réponses. Quel est l'âge du capitaine ? de son fils ? de M. Dupont ? Quels sont les âges des deux filles ?*

1.5.2 Une solution.

Étant donné que le produit des âges vaut 2450, c'est donc que les âges des voyageurs sont des diviseurs de 2450. Or $2450 = 1 \times 2 \times 5 \times 5 \times 7 \times 7$ (décomposition en nombres premiers) On a alors comme âges possibles : 1, 2, 5, 7, 10, 14, 25, 35, 49, 50, 70, 98, 175, 245, 350, 490, 1225, ou 2450.

Il semble absurde de supposer que l'âge d'un des passagers puisse excéder 174 ans (quoi que). Ainsi, les âges possibles sont réduits aux douze premiers diviseurs.

On a donc qu'un nombre fini de triplets possible[5] : $\{98, 25, 1\}$, $\{98, 5, 5\}$, $\{70, 35, 1\}$, $\{70, 7, 5\}$, $\{50, 49, 1\}$, $\{50, 7, 7\}$, $\{49, 25, 2\}$, $\{49, 10, 5\}\ldots$

Il suffit de faire la somme de chacun des triplets. Or le fils du capitaine dit ne pas avoir assez d'indices pour trouver avec les sommes, c'est donc qu'il existe deux sommes identiques. En effet, les triplets $\{50, 7, 7\}$ et $\{49, 10, 5\}$ ont la même somme (64 ans).

On en déduit l'âge du fils qui est de $64/4 = 16$ ans.

De plus comme le capitaine est plus âgé que M. Dupont, on déduit que M. Dupont n'a que (sic !) 49 ans De là ses filles ont 10 et 5 ans. On peut également dire que le capitaine a 50 ans.

Donc

- M. Dupont a 49 ans,
- les deux filles ont 5 et 10 ans,
- Le capitaine a 50 ans et
- Le fils a 16 ans.

1.6 Quel est le nombre qui continue cette suite : 2, 12, 1112...

...1112.

La construction de la suite se fait comme suit : il faut lire à haute voix les chiffres qui la composent. On part de "2", on lit un "2", on écrit 12. Puis, on lit un "1", un "2" on écrit 1112. On lit trois "1", un "2", on écrit 3112.

Lemme : *On peut montrer par contradiction que le nombre de signes distincts qui composent cette suite se limite aux chiffres 1, 2 et 3.*

Démonstration.
En effet, supposons qu'à une ligne on trouve le chiffre 4, suivi par exemple du chiffre 1. Cela signifie qu'à la ligne précédente, on avait la suite ...1111... C'est à dire à la ligne d'avant ...11..., que l'on aurait dû traduire par 21 et non 1111 comme cela a été fait. D'où contradiction. Il ne peut donc pas y avoir de chiffre supérieur strictement à 3.

□

Dans la suite, Nous nommerons u_n la valeur trouvée à l'étape n et l'on prendra la convention suivante : $u_0 = 2$, $u_1 = 12$, $u_3 = 1112$, etc.

[5] je ne les écrit pas tous

Lemme : *Cette suite ne peut pas se stabiliser, et même elle tend vers l'infini.*

Démonstration.

Nous savons calculer u_{n+1} en fonction de u_n mais aussi, en toute logique, u_{n-1} en fonction de u_n (cela paraît évident) qui n'a qu'une seule valeur possible en fonction de u_n.

Ainsi, supposons qu'il existe $m > n$ tels que $u_m = u_n$ alors, d'après l'observation précédente, $u_{m-1} = u_{n-1}$ et finalement par une récurrence évidente, $u_{m-n} = u_0 = 2$. Or $m - n > 0$ donc d'après la méthode de construction de la liste, u_{m-n} a un nombre pair de chiffres d'où une contradiction.

Donc quels que soient m et n, $u_m \neq u_n$; soit v_n la suite définie de la façon suivante :
si il existe k tel que $u_k = n$ alors $v_n = k$. Sinon, $v_n = 0$ Quel que soit $A \in N$, $N = \max(v_0 \dots v_N) \Rightarrow (n > N \Rightarrow u_n > A)$. Donc u_n tend vers l'infini.

□

Lemme : *On peut même montrer que le nombre de 1 diverge.*

Démonstration.

Je nomme groupe une suite de chiffres faisant partie d'une autre suite : non décalé (gn) si il commence par un chiffre de rang impair groupe décalé (gd) si il commence par un chiffre de rang pair (1232 est un gn de 221232 et un gd de 2221232).

Si X et Y sont deux groupes, je note $X \to Y$ si une partie de X engendre Y en remontant dans les termes de la suite (engendrer sera toujours employé dans ce sens). Exemple :

$$\begin{array}{ll} 1112 & (\text{gn}) \to \ 12 \\ 2322 & (\text{gn}) \to 3322 \\ 2322 & (\text{gd}) \to \ 222 \end{array}$$

(en effet, les chiffres peuvent être groupés par deux lorsque l'on remonte dans la suite). On appellera un doublet, une gn de deux chiffres.

Deux doublets consécutifs ne peuvent pas se terminer par le même chiffre (on appellera cela une incompatibilité de répétition ir) le groupe 333 ne peut pas exister car il contient forcément un doublet 33 et donc engendrera forcément 333 ce qui par récurrence arrive à une contradiction évidente.

le doublet 33 ne peut donc pas exister. Un groupe contenant un doublet 33 provoquera une incompatibilité 3 ($i3$).

Cherchons les gn de quatre chiffres ne contenant pas de 1 et ne provoquant pas d'incompatibilité (ie : pouvant exister dans la suite) : je ne regarde pas les ir et les $i3$ triviales :

2223 →	2233	supposé compatible
2322 →	3322	supposé compatible
2332 →	33222	i3 si gn et ir si gd incompatible
3223 →	22233	supposé compatible

Tous les autres sont incompatibles.

Cherchons les *gn* de huit chiffres ne contenant pas de 1 et ne provoquant pas d'incompatibilité (il sont constitués des *gn* de quatre chiffres compatibles) :

22232223 →	22332233	i3 si gn donc gd
→	3322233	i3 dans tous les cas
22232322 →	ir	
22233223 →	223322233	i3 dans tous les cas
23222223 →	ir	
23222322 →	33223322	i3 si gn donc gd
→	22233222	i3 si gd et ir si gn
23223223 →	ir	
32232223 →	222332233	i3 si gd donc gn
→	223322233	i3 dans tous les cas
32232322 →	ir	
32233223 →	2223322233	i3 dans tous les cas

Donc tout *gn* de 8 chiffres contient au moins un 1. Cette suite tendant vers l'infini, le nombre de *gn* de 8 chiffres (même sans recouvrement) tend vers l'infini. Donc le nombre de 1 aussi.

□

Lemme : *On constate que la fin du code est stabilisée dès la 6ème itération pour ses signes finaux.*

On note D la transformation telle que $u_{n+1} = D(u_n)$.

On note $|X|$ le nombre de chiffres du groupe X.

Lemme 1 : *pour $n \geq 6$, $u(n)$ se termine par* 222112 *si n est pair, et par* 322112 *si n est impair.*

Démonstration par récurrence. C'est vrai pour $n = 6$. Soit $n > 6$, supposons la propriété vérifiée pour n et montrons la pour $n + 1$. Si n est pair, u_n se termine par 222112. Le chiffre précédent, s'il existe, n'est pas un 2 (puisqu'on n'a pas 4 chiffres identiques consécutifs).

Donc u_n se termine par un bloc de trois 2, puis un bloc de deux 1, puis un 2, et donc u_{n+1} se termine par 322112 ; $n + 1$ étant impair, c'est précisément ce qu'on attendait.

Si n est impair, u_n se termine par 322112. Peu importe si le chiffre précédent est un 3 ou non, seuls les trois derniers blocs nous intéressent, et u_{n+1} se termine donc par 222112 ; $n + 1$ étant pair, c'est ce qu'on attendait.

Lemme 2 : *Soit x_n la suite définie pour $n \geq 6$ par $x_6 = 222112$, $x_7 = 322112$, $x_8 = 3222112$, $x_9 = 3322112$, $x_10 = 23222112$, et pour $n \geq 10$, x_{n+1} est égal à $D(x_n)$ privé de son premier chiffre.*

1. Pour tout $n \geq 6$, x_n est un suffixe de u_n.
2. Pour tout $n \geq 6$, x_n est un suffixe de x_{n+2}.
3. Pour $n \geq 11$, $|x_n|$ est impair et $|x_{n+2}| \geq |x_n| + 2$.
4. Pour tout n, $|x_n| \geq n - 2$.

Démonstrations 1. Par récurrence sur n, d'après la définition de u_n. Noter qu'il est important d'enlever le premier chiffre de $D(x_n)$ pour construire x_{n+1}, car le chiffre précédant x_n dans u_n peut être identique au premier chiffre de x_n, ce qui modifie la longueur du premier bloc ; en revanche, le second chiffre de $D(x_n)$, qui indique la nature du premier bloc de x_n, peut être conservé car il apparaît bien à cet endroit dans u_{n+1}. Au passage, on peut noter que ce chiffre est toujours un 2.

2. Par récurrence sur n. On le vérifie à la main pour $n < 10$. Pour $n \geq 10$, si x_n est un suffixe de x_{n+2}, alors x_{n+1} est un suffixe strict de $D(x_{n+2})$, donc c'est un suffixe de x_{n+3}.

3. Pour $n \geq 11$, la longueur de x_n est impaire par construction. On montre $|x_{n+2}| \geq |x_n| + 2$ par récurrence. C'est vrai pour $n = 11$ (et même $n = 9$ et $n = 10$). Supposons que $|x_{n+2}| \geq |x_n| + 2$ pour un certain $n \geq 11$. Comme x_n est un suffixe de x_{n+2}, on peut écrire $x_{n+2} = wx_n$, avec $|w| \geq 2$.

L'avant dernier chiffre de w ne peut être égal au premier chiffre de x_n, car il s'agit de deux chiffres consécutifs en position impaire qui sont toujours distincts dans une image par D.

Par conséquent $D(x_{n+2})$ contient au moins un bloc de plus que $D(x_n)$ et donc $|D(x_{n+2})| \geq |D(x_n)| + 2$, d'où $|x(n + 3)| \geq |x_{n+1}| + 2$.

4. Se déduit facilement des valeurs initiales et de (3).

Théorème.

Pour tout $l \geq 0$, il existe un entier n_0 tel que pour tout $n > n_0$, le suffixe de longueur l de u_n coïncide avec le suffixe de longueur l de u_{n+2}.

C'est vrai d'après le lemme 1 pour $l \leq 6$, en prenant $n_0 = 6$. Pour l supérieur à 6, on prend $n_0 = l + 2$. Alors pour $n > n_0$, le suffixe de longueur l de u_n est un suffixe de x_n, puisque $|x_n| \geq n - 2 \geq l$ d'après le 4. du lemme 2.

Il coïncide donc avec le suffixe de longueur l de u_{n+2}, puisque x_n est un suffixe de x_{n+2} qui est un suffixe de u_{n+2}. (d'après les 2. et 1. du lemme 2.)

On peut formaliser ce résultat en définissant une topologie convenable sur l'ensemble des mots finis et infinis sur l'alphabet $\{1, 2, 3\}$. On peut alors montrer que la suite u_n a deux valeurs d'adhérence, qui sont deux mots infinis à gauche :

...13122112132113111211132231121113213221231221132221222112112

et

...21133112311322132112311312132211121311222113321132211221121332211 2

Exercice : faire la même chose en regardant cette fois le début des mots (ce qui est plus habituel que de regarder la fin, d'ailleurs !). Cette fois, il y a 3 valeurs d'adhérence.

Lemme : *Enfin, on peut montrer que la proportion de* 1 *a une limite finie, qui est un nombre algébrique de degré* 71.

L'idée de Conway est d'identifier un certain ensemble fini de blocs (qu'il nomme selon les éléments chimiques), de sorte que si x est l'un de ces blocs, $D(x)$ s'exprime comme concaténation de plusieurs blocs élémentaires, et de plus qu'il n'y ait jamais d'interaction entre blocs consécutifs, de sorte que $D(xy) = D(x)D(y)$.

D agit alors comme une substitution (i.e. un morphisme de monoïde) sur les mots formés de symboles de blocs élémentaires, et le nombre d'occurrences de chacun des blocs élémentaires peut être exprimé à l'aide de puissances de la matrice de la substitution.

D'où finalement une densité qui s'exprime en fonction des valeurs propres de cette matrice et est donc un nombre algébrique.

Lire à ce sujet : [Con87], [**?**] et [Del97].

Références L'Encyclopédie électronique des Suites Entières

1.7 Probabilité que 2 personnes soient nées le même jour.

On réunit dans une pièce n personnes. On veut determiner, d'une part, quelle est la probabilité que deux de ces personnes soient nées le même jour et d'autre part, à partir de combien de personnes il y a plus d'une chance sur deux que l'événement cherché soit réalisé.

Enfin, on s'intéressera aux différentes méthodes de calcul et à ce qu'il se passe quand les dates de naissance ne sont pas équiprobables.

1.7.1 Toutes les dates de naissance sont équiprobables.

On supposera dans un premier temps que les dates d'anniversaire ont la même probabilité d'apparition. La bonne solution, comme souvent en probabilités, consiste à calculer la probabilité de l'événement complémentaire. C'est à dire qu'on s'intéresse à la probabilité qu'il n'y ait aucune coïncidence des dates d'anniversaire dans un groupe de n personnes.

En effet, la probabilité d'un événement additionnée à celle de son complémentaire est égale à 1. Donc la probabilité de l'événement est égale à un moins la probabilité de son complémentaire.

Le nombre total de cas possibles est 365 à la puissance n noté 365^n. Le nombre de cas favorables est le nombre de choix ordonnés de n dates parmi 365, soit : $\frac{365!}{(365-n)!}$ Donc, la probabilité qu'il n'y ait aucune coïncidence de dates d'anniversaire est :

$$\frac{365!}{(365-n)! \times 365^n} = \prod_{i=0}^{n-1} \frac{365-i}{365} = \prod_{i=0}^{n-1} 1 - \frac{i}{365}$$

La probabilité $\mathbb{P}_n$ qu'il y ait au moins une coïncidence est donc :

$$\mathbb{P}_n = 1 - \prod_{i=0}^{n-1} 1 - \frac{i}{365}$$

Il existe d'autres solutions, par exemple, la suivante. On nomme A, B etc. les personnes de l'assistance. Alors, il y a une coïncidence si A a le même anniversaire que B, ou que $C, \ldots$ Ou encore si B a le même anniversaire que C, etc. Le problème de cette méthode, c'est qu'il faut ensuite calculer la probabilité d'une union d'événements qui ne sont pas disjoints. Il faut faire la somme des probabilités des évènements, en enlever la somme des intersections 2 à 2, ajouter la somme des intersections 3 à 3, etc. (c'est la formule de Poincaré). C'est long et pénible.

A partir de quelle valeur de n cette probabilité dépasse-t-elle $1/2$ **?** La seule solution est de calculer la probabilité pour différentes valeurs de n, et de rechercher la première valeur de $\mathbb{P}_n$ dépassant $1/2$. Cette valeur est 23. En effet, on trouve par le calcul :

$$\begin{aligned} \mathbb{P}_{22} &= 0.4756953077 \\ \mathbb{P}_{23} &= 0.5072972343 \end{aligned}$$

Les différentes méthodes de calcul pratique. Si on veut calculer brutalement le nombre : $\frac{365!}{(365-n)!\times 365^n}$ avec une machine à calculer, on obtient un dépassement de capacité. Il faut donc réfléchir un peu pour faire le calcul, en tenant compte des possibilités informatiques.

Si on dispose d'un logiciel de calcul mathématique, tel que Maple ou Mathématica, le problème de dépassement de capacité disparaît, et on utilise n'importe laquelle des expressions ci-dessus.

Si on dispose d'une calculatrice programmable, il est possible de programmer, avec une boucle, le calcul de l'expression :

$$\prod_{i=0}^{n-1} 1 - \frac{i}{365}$$

Comme cette programmation dépend de la calculatrice et du langage utilisés, il est difficile d'en dire plus.

Si l'on dispose d'un tableur, l'expression : $\prod_{i=0}^{n-1} 1-\frac{i}{365}$ est facile à calculer en faisant un tableau de taille n, avec 3 colonnes : une colonne pour i, une colonne pour $1-\frac{i}{365}$, une colonne pour le produit.

Si l'on n'a qu'une calculatrice scientifique, il est nécessaire d'utiliser une approximation. Il faut remarquer que

$$\prod_{i=0}^{n-1} 1 - \frac{i}{365} = \exp(\sum_{i=0}^{n-1}(\log(1 - \frac{i}{365})$$

Or, si n est beaucoup plus petit que 365, on a : $\log(1-\frac{i}{365}) \sim -\frac{i}{365}$. Et comme la somme des n premiers entiers est égal à $\frac{n(n+1)}{2}$.

On a donc :

$$\prod_{i=0}^{n-1} 1 - \frac{i}{365} \exp(-n\frac{n-1}{730})$$

Cette approximation est relativement précise. Elle donne les valeurs suivantes pour $\mathbb{P}_{22}$ et $\mathbb{P}_{23}$, ce qui donne une réponse juste pour la question (2) malgré la très forte proximité de $\mathbb{P}_{23}$ avec $1/2$:

$$\begin{aligned}\mathbb{P}_{22} &\sim 0.4689381108\\ \mathbb{P}_{23} &\sim 0.5000017522\end{aligned}$$

1.7.2 Probabilités inégales pour les dates de naissance.

Nous avons jusque là supposé que toutes les dates de naissance étaient de même probabilité. Que se passe-t-il si on se passe de cette hypothèse ?

Les probabilités de coïncidence d'anniversaire sont augmentées. Cela paraît assez naturel puisque, si ces probabilités sont très concentrées, par exemple sur une seule date dans l'année, la probabilité de coïncidence se rapproche de 1.

La difficulté de cette question tient au fait que l'on ne peut pas faire varier n'importe comment les probabilités des différentes dates de naissance. En effet, il faut que la somme de toutes ces probabilités fasse 1.

Notons p_i la probabilité de naissance le jour i et A l'ensemble des jours de l'année. Alors la probabilité de non-coïncidence est :

$$\sum_{\substack{S\in A\\ |S|=n}} \prod_{i\in S} p_i$$

On s'intéresse aux jours 1 et 2. Les ensembles de cardinal n inclus dans A peuvent être classés en 3 catégories :

- les ensembles ne contenant ni 1 ni 2,
- les ensembles contenant 1 ou 2, mais pas les deux,
- les ensembles contenant 1 et 2.

On note A' l'ensemble A, moins les éléments 1 et 2. La somme ci-dessus peut être réécrite :

$$\sum_{\substack{S\in A'\\ |S|=n}} \prod_{i\in S} p_i + (p_1+p_2) \sum_{\substack{S\in A'\\ |S|=n-1}} \prod_{i\in S} p_i p_1 p_2 \sum_{\substack{S\in A'\\ |S|=n-2}} \prod_{i\in S} p_i$$

Supposons que p_1 et p_2 soient différents, et montrons que l'on peut augmenter la probabilité de non-coïncidence. On remplace p_1 et p_2 par $\frac{p_1+p_2}{2}$.

Le premier des 3 termes ci-dessus n'est pas modifié, puisque ni p_1 ni p_2 n'y apparaissent. Le second non plus, car il ne dépend que de p_1+p_2. En revanche, le troisième terme est augmenté, car on remplace p_1p_2 par $(\frac{p_1+p_2}{2})^2$ qui est plus grand.

Donc, si les p_i sont différents, la probabilité de non-coïncidence n'est pas maximale.

Par la suite, on démontre qu'une fonction continue sur un ensemble fermé borné atteint son maximum. Or, l'ensemble des vecteurs formés de 365 probabilités, dont la somme fait 1, est un ensemble fermé borné. Donc le maximum de la probabilité de coïncidence est atteint. Il ne peut être atteint que si tous les pi sont égaux, qui est donc le maximum. Donc, la probabilité de non coïncidence est maximale si les probabilités de jours de naissance sont égales.

Par conséquent, si les probabilités sont égales, la probabilité de coïncidence d'anniversaire est minimale.

1.8 Somme et produit de deux entiers.

1.8.1 Énoncé.

Enoncé : *Un professeur de mathématiques donne un problème à résoudre à ses deux meilleurs élèves, Pierre et Sophie. Il donne à Pierre le produit de deux nombres entiers compris (au sens large) entre* 2 *et* 100*, et à Sophie la somme des deux mêmes nombres, puis il leur demande s'ils peuvent déterminer quels étaient les nombres de départ.*

> *Pierre : « Non, je ne peux pas trouver ces deux nombres. »*
> *Sophie : « Je le savais. »*
> *Pierre : « Dans ce cas, je connais les deux nombres. »*
> *Sophie : « Alors moi aussi. »*

Sachant que les deux élèves sont d'excellents logiciens et que leurs quatre déclarations étaient rigoureusement exactes, saurez-vous être aussi futés qu'eux, et trouver les deux nombres choisis par le professeur ?

Commentaires. L'énoncé tel qu'il est présenté ici est le plus proche de ce qui est en général posé dans `fr.sci.maths`. Malheureusement, il lui manque des précisions importantes sur ce que le prof de maths dit effectivement aux élèves.

D'abord, il n'est pas précisé que Pierre sait que Sophie a la somme, et que Sophie sait que Pierre a le produit. Bon, d'accord, c'est « évident ». Mais il y a un autre point qui semble « évident »alors qu'il est souvent source d'erreurs de raisonnement.

Cet autre point, c'est que l'on ne sait pas si Pierre et Sophie connaissent la valeur minimum (2) et la valeur maximum (100) des nombres de départ. Pour ce qui est de la valeur minimum, il faut qu'ils la connaissent ; sinon, le problème est impossible (par exemple, s'ils pensent que les nombres commencent à 1 au lieu de 2, alors la seule solution serait 1 et 4, qui ne rentre

pas dans le cadre de l'énoncé). En ce qui concerne la valeur maximum, les deux cas sont possibles (soit ils connaissent la limite, soit ils ne la connaissent pas), ce qui donne deux problèmes différents, intéressants tous les deux.

1.8.2 Solutions.

Puisqu'il y a deux interprétations possibles de l'énoncé, il y a aussi deux solutions distinctes. La première suppose que Pierre et Sophie savent que les nombres sont compris entre 2 et 100, alors que la seconde suppose qu'ils savent seulement que les nombres sont supérieurs ou égaux à 2. Néanmoins, la méthode utilisée est la même dans les deux cas, à savoir prendre chacune des 4 affirmations dans l'ordre, et en déduire des informations précieuses sur les sommes S et produits P possibles.

Solution dans le premier cas (valeur maximum connue, égale à 100).

Pierre : « Non, je ne peux pas trouver ces deux nombres. »

Ceci signifie que le produit P peut se décomposer d'au moins deux manières différentes en produit de deux nombres compris entre 2 et 100. Par exemple, on pourrait avoir $P = 75$, car ce produit se décompose en 3×25 ou 5×15, mais on ne peut pas avoir $P = 77$ car alors la décomposition serait unique : 7×11.

Voici une liste de valeurs que nous pouvons d'ores et déjà éliminer :

- toute valeur inférieure à 4 ou supérieure à 10000
- le produit de deux nombres premiers (par exemple $77 = 7 \times 11$)
- le cube d'un nombre premier (par exemple $125 = 5 \times 25$)
- le double du carré d'un premier plus grand que 10 (par exemple, $242 = 2 \times 11 \times 11 = 11 \times 22$: la décomposition en 2×121 est impossible)
- un multiple strict d'un nombre premier plus grand que 50 (par exemple $318 = 6 \times 53$)
- le produit du carré d'un premier plus grand que 10 par un nombre premier (par exemple $242 = 2 \times 11 \times 11 = 11 \times 22$; la décomposition en 2×121 est impossible)
- et bien d'autres...

Sophie : « Je le savais. »

Ceci signifie que la somme S ne peut pas s'écrire comme somme de deux nombres dont le produit aurait été éliminé dans l'étape précédente.

Par exemple, la somme 11 convient car tous les produits possibles sont *non uniques* :

$$\begin{aligned} 11 &= 2+9 \\ 2\times 9 = 18 &= 3\times 6 \\ 11 &= 3+8 \\ 3\times 8 = 24 &= 2\times 12 = 4\times 6 \\ 11 &= 4+7 \\ 4\times 7 = 28 &= 2\times 14 \\ 11 &= 5+6 \\ 5\times 6 = 30 &= 2\times 15 = 3\times 10 \end{aligned}$$

En revanche, la somme 13 ne convient pas car : $13 = 2 + 11$; $2 \times 11 = 22$ (pas d'autre décomposition)

Par conséquent, on peut commencer par éliminer toutes les sommes de deux nombres premiers. Vous pouvez vérifier que cela élimine déjà toutes les sommes paires (ceci a été conjecturé par Goldbach dans le cas général, et vérifié par ordinateur sur beaucoup plus de nombres que ce dont on a besoin pour résoudre ce problème). Pour ce qui est des sommes impaires, on élimine celles qui sont égales à un nombre premier plus 2 : $5 = 3 + 2$, $7 = 5 + 2$, $9 = 7 + 2 =$, $13 = 11 + 2$, etc.

Après ce premier débroussaillage, il nous reste les sommes qui sont égales à un nombre composé impair plus 2 : 11 ($3 \times 3 + 2$), 17 ($3 \times 5 + 2$), 23 ($3 \times 7 + 2$), 27 ($5 \times 5 + 2$), 29, 35, 37, 41, 47, 51, 53, 57, 59, etc.

Nous pouvons aussi supprimer toutes les sommes S à partir de 57, puisque si $57 \leq S \leq 153$, on peut écrire $S = 53 + n$, avec $4 \leq n \leq 100$, si $155 \leq S \leq 197$, on peut écrire $S = 97 + n$, avec $58 \leq n \leq 100$, si $S = 199$, on peut écrire $S = 100 + 99$. Dans chacun de ces trois cas, le produit P correspondant (soit $53n$, soit $97n$, soit 100×99) a une décomposition unique.

On peut enfin supprimer la somme $S = 51 = 17 + 34$, car le produit $P = 17 \times 34$ n'a pas d'autre décomposition.

Voici donc la liste exhaustive des sommes possibles à cette étape du raisonnement, avec pour chaque somme la liste des produits possibles.

11	18, 24, 28, 30
17	30, 42, 52, 60, 66, 70, 72
23	42, 60, 76, 90, 102, 112, 120, 126, 130, 132
27	50, 72, 92, 110, 126, 140, 152, 162, 170, 176, 180, 182
29	54, 78, 100, 120, 138, 154, 168, 180, 190, 198, 204, 208, 210
35	66, 96, 124, 150, 174, 196, 216, 234, 250, 264, 276, 286 294, 300, 304, 306
37	70, 102, 132, 160, 186, 210, 232, 252, 270, 286, 300, 312 322, 330, 336, 340, 342

41	78, 114, 148, 180, 210, 238, 264, 288, 310, 330, 348, 364, 378 390, 400, 408, 414, 418, 420
47	90, 132, 172, 210, 246, 280, 312, 342, 370, 396, 420, 442, 462 480, 496, 510, 522, 532, 540, 546, 550, 552
53	102, 150, 196, 240, 282, 322, 360, 396, 430, 462, 492, 520, 546 570, 592, 612, 630, 646, 660, 672, 682, 690, 696, 700, 702

Pierre : « Dans ce cas, je connais les deux nombres. »

Pour que Pierre puisse faire cette affirmation, il faut que le produit P se trouve une fois et une seule dans la liste que nous venons d'écrire. Cela élimine donc les produits $P = 30$ ($S = 11$ ou 17), $P = 42$ ($S = 17$ ou 23), etc. Il reste :

11	18, 24, 28
17	52
23	76, 112, 130
27	50, 92, 110, 140, 152, 162, 170, 176, 182
29	54, 100, 138, 154, 168, 190, 198, 204, 208
35	96, 124, 174, 216, 234, 250, 276, 294, 304, 306
37	160, 186, 232, 252, 270, 336, 340
41	114, 148, 238, 288, 310, 348, 364, 378, 390, 400, 408, 414, 418
47	172, 246, 280, 370, 442, 480, 496, 510, 522, 532, 540, 550, 552
53	240, 282, 360, 430, 492, 520, 570, 592, 612, 630, 646, 660, 672 682, 690, 696, 700, 702

Sophie : « Alors moi aussi. »

Pour que Sophie puisse dire cela, il faut qu'il ne reste plus qu'un seul produit correspondant à la somme qu'elle connaît. Ceci n'est réalisé que si la somme est 17, auquel cas le produit est 52. Les nombres de départ sont donc 4 et 13.

Solution dans le second cas (valeur maximum inconnue).

Pierre : « Non, je ne peux pas trouver ces deux nombres. »

Ceci signifie que le produit n'est pas le carré ou le cube d'un nombre premier, ni le produit de deux nombres premiers. Nous ne pouvons rien en déduire de plus pour le moment.

Sophie : « Je le savais. »

Comme dans le premier cas, nous pouvons éliminer toute somme paire et toute somme d'un nombre premier avec 2, et il nous reste les sommes égales à un nombre composé impair plus 2. Contrairement au premier cas, nous ne pouvons éliminer aucune autre somme. La liste (incomplète) des sommes et produits possibles est la suivante :

11	$18, 24, 28, 30$
17	$30, 42, 52, 60, 66, 70, 72$
23	$42, 60, 76, 90, 102, 112, 120, 126, 130, 132$
27	$50, 72, 92, 110, 126, 140, 152, 162, 170, 176, 180, 182$
29	$54, 78, 100, 120, 138, 154, 168, 180, 190, 198, 204, 208$ 210
35	$66, 96, 124, 150, 174, 196, 216, 234, 250, 264, 276, 286$ $294, 300, \ldots$
37	$70, 102, 132, 160, 186, 210, 232, 252, 270, 286, 300, 312$ $322, 330, \ldots$
41	$78, 114, 148, 180, 210, 238, 264, 288, 310, 330, 348, 364$ $378, 390, \ldots$
47	$90, 132, 172, 210, 246, 280, 312, 342, 370, 396, 420, 442$ $462, 480, \ldots$
...	

Pierre : « Dans ce cas, je connais les deux nombres. »

Comme tout-à-l'heure, nous éliminons les produits P qui se trouvent plus d'une fois dans la liste. Il reste (liste exhaustive pour toutes les sommes inférieures à 200) :

11	$18, 24, 28$
17	52
23	$76, 112$
27	$50, 92, 140, 152, 176$
29	$54, 100, 208$
35	$96, 124, 216, 304$
37	$160, 232, 336$
41	$148, 288, 400$
47	$172, 280, 496$
51	$98, 144, 188, 308, 344, 608, 620, 644$
53	$520, 592$
57	$212, 260, 392, 656, 800$

59	220, 688
65	244
67	192, 472, 1116
71	268, 448, 880
77	292, 832, 976
79	228, 568, 1504
83	316, 1072, 1216
87	332, 632, 836, 1136, 1340, 1472, 1880
89	1168
93	356, 1040, 1856, 1952
95	1264, 1984
97	712, 1296
101	388, 1144, 2368, 2440
107	412, 2752
113	436, 1552
117	452, 872, 1616, 3392
119	1648, 2728
121	904, 2848
123	242, 476, 1712, 3440, 3776
125	484, 1744, 3904
127	1776
131	384, 508, 3784, 4288
135	266, 524, 896, 1016, 1904, 2996, 3296, 3956, 4544
137	4672
143	556, 2032, 4120, 5056
145	1096, 2176, 3616
147	290, 1112, 1496, 2096, 2432, 3680, 4280, 5312
155	604, 2224
157	1192, 3712
161	628, 2320, 6208, 6424
163	4192
167	652, 2416, 5080, 6592
171	338, 668, 1304, 2480, 2888, 3020, 3404, 3888, 4448, 5240, 5504 6848, 6968, 7100, 7304
173	2512, 6976
177	692, 3140, 7232
179	2608

185	724
187	1432, 7552
189	374, 1448, 2768, 2924, 5024, 5624, 7208, 7448, 7808
191	8128
197	772, 2896

Sophie : « Alors moi aussi. »

Ici encore, il doit rester un seul produit sur la ligne de la somme correspondant à ce qu'a Sophie. Là encore, les nombres 4 et 13 sont solution ($S = 17$, $P = 52$), mais ce ne sont plus les seuls. Les autres solutions sont : 4 et 61 ($S = 65$, $P = 244$), 16 et 73 ($S = 89$, $P = 1168$), 64 et 73 ($S = 137$, $P = 4672$). Bien évidemment, on ne tiendra pas compte des cas où l'un des nombres est plus grand que 100, par exemple pour $S = 127$ et $P = 1776$: les nombres seraient 16 et 111.

1.9 Les deux échelles.

1.9.1 Énoncé.

Enoncé : *Deux échelles se croisent dans une cour rectangulaire. Les échelles mesurent respectivement 10 et 14m. Vues de côté, elles se croisent à 5m du sol.*
On suppose les échelles droites et sans épaisseur. Quelle est la largeur de la cour ?

1.9.2 Solution.

On appelle :
- BF l'échelle qui va du haut à gauche au bas à droite (B en haut à gauche) ;
- EA l'échelle qui va du bas à gauche au haut à droite (E en bas)
- I le point d'intersection des deux échelles,
- H la projection orthogonale de I sur le sol.

On pose :

$$\begin{array}{cccccccccccc} AF & = & x & ; & EF & = & y & ; & BE & = & z & ; \\ AE & = & a & ; & BF & = & b & ; & IH & = & c & ; \end{array}$$

On applique le théorème de Pythagore dans le triangle FAE rectangle en F. $AF^2 + EF^2 = AE^2$. C'est-à-dire : $x^2 + y^2 = a^2$.

On applique le théorème de Pythagore dans le triangle FEB rectangle en E. $EF^2 + BE^2 = BF^2$. C'est-à-dire : $y^2 + z^2 = b^2$.

On applique le théorème de Thalès avec les parallèles (BE), (IH) et (AF). $\frac{c}{x} + \frac{c}{z} = \frac{EH}{EF} + \frac{HF}{EF} = \frac{EH+HF}{EF} = 1$, d'où $\frac{c}{z} = \frac{x-c}{x}$

Il vient donc : $x^2 - z^2 = a^2 - b^2$ et $z = c \times \frac{x}{x-c}$ Et puis : $x^2 - (c\frac{x}{x-c})^2 = a^2 - b^2$ ce qui conduit à

$$x^4 - 2cx^3 - (a^2 - b^2)x^2 + 2c(a^2 - b^2)x - c^2(a^2 - b^2) = 0$$

Cette équation de degré 4 se résout par radicaux ou de manière approchée avec tout outil de calcul. On a ensuite : $y = \sqrt{a^2 - x^2}$.

On a ici : $a = 14$, $b = 10$, $c = 5$, d'où l'équation :

$$x^4 - 10x^3 - 96x^2 + 960x - 2400 = 0$$

Et donc : $x = 12.78405$ avec un outil de calcul. Puis $y = \sqrt{a^2 - x^2} = 5.706842$.

1.10 La cuve de vin.

1.10.1 Énoncé.

Enoncé : *Du vin est placé dans une cuve cylindre horizontale de longueur L est de rayon R. On mesure la hauteur h de vin dans la cuve, et on veut en déduire le volume v de vin.*

1.10.2 Une solution.

On peut s'en sortir avec un peu de calcul intégral. Si l'on considère un axe vertical dont l'origine est placée à la hauteur du centre du cylindre, alors la section du cylindre à la hauteur z a pour aire

$$a(z) = 2\sqrt{R^2 - z^2}L$$

Par conséquent, on a : $v = \int_{-R}^{h-R} a(z)$

L'intégrale se calcule par le changement de variable $z = R\sin t$. On a alors successivement, en notant $x = \frac{h}{R} - 1$ ($x = -1$ lorsque la cuve est vide, $x = 0$ à mi-hauteur et $x = 1$ quand elle est pleine) :

$$\begin{aligned} v &= \int_{-\pi/2}^{\arcsin x} 2LR^2 \cos^2 t \\ v &= LR^2 \int_{-\pi/2}^{\arcsin x}, 1 + \cos 2t \\ v &= LR^2[\arcsin x + \sin 2 \arcsin x - (-\pi/2 + \sin(-\pi)] \end{aligned}$$

Finalement, si l'on appelle $v_0 = \pi R^2 L$ le volume total de la cuve, on obtient le taux de remplissage :

$$\frac{v}{v_0} = \frac{\arcsin x + x\sqrt{1-x^2} + \frac{\pi}{2}}{\pi}$$

ou plus simplement

$$\frac{v}{v_0} = \frac{x\sqrt{1-x^2} + \arccos(-x)}{\pi}$$

Voici quelques valeurs :

x	-1.00	-0.75	-0.50	-0.25	0.00	0.25	0.50	0.75	1.00
$v/v0$	0%	7%	20%	34%	50%	66%	80%	93%	100%

À noter que la formule ne s'inverse pas simplement (i.e. il n'y a pas d'expression élémentaire de h en fonction de v), donc si c'est le niveau de vin que l'on cherche, le plus simple est de résoudre l'équation numériquement. On trouve par exemple :

$v/v0$	10%	20%	30%	40%	50%	60%	70%	80%	90%
x	-0.69	-0.49	-0.32	-0.16	-0.00	0.16	0.32	0.49	0.69

Chapitre 2

Les ensembles et les nombres

2.1 Sommes usuelles

2.1.1 Somme des x^k

Soit a et b deux entiers naturels. Soit $x \in \mathbb{R}$, $x \neq 1$.

$$(x-1)\sum_{k=a}^{b} x^k = x^b - x^a$$

Cette formule se démontre facilement par récurrence sur b.

Remarque : En fait cette formule est vraie sur tout anneau, même s'il n'est pas commutatif.

2.1.2 Somme des puissances des premiers entiers.

Dans tout ce qui suit, n est un entier naturel.

On cherche une expression de la somme $\sum_{i=0}^{n} i^k$, pour différentes valeurs de k.

$k = 1$: Somme des premiers entiers.

La formule est due à Léonard Euler. Il suffit d'écrire :

$$\begin{array}{rcccccc} S = & 1 & + & 2 & +\cdots+ & n \\ S = & n & +(n-1)+ & \cdots & + & 1 \\ \hline 2S = & (n+1)+ & (n+1)+ & \cdots & +(n+1) & \end{array}$$

C'est-à-dire : $S = \frac{n(n+1)}{2}$

k=2

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

Cette formule se démontre sans difficulté par récurrence.

$k > 1$: Calcul général

Il faut alors définir les nombres de Bernoulli (voir aussi 2.2) que l'on notera B_t. On peut les définir comme suit : les B_t sont les nombres de Bernoulli définis (par exemple) par la série génératrice :

$$\sum_{t=0}^{\infty} B_t \frac{x^t}{t!} = \frac{x}{\exp(x) - 1}$$

Alors on peut écrire :

$$\sum_{i=0}^{k} i^k = \frac{1}{k+1} \sum_{i=0}^{k} B_i C_{k+1}^t (n+1)^{k+1-t}$$

Remarque. Les C_{k+1}^t représentent les coefficients du binôme de Newton, c'est-à-dire le nombre de combinaisons de t éléments d'un ensemble à $k+1$ éléments.

La définition des nombres de Bernoulli n'est pas tout à fait standardisée : il y traîne chez certains auteurs des facteurs $(-1)^t$; chez d'autres les indices t deviennent $t/2$, et j'en passe. Il convient donc de toujours bien regarder la définition adoptée.

Avec celle-ci, on a :

$$\begin{aligned}
B_0 &= 1 \\
B_1 &= -\tfrac{1}{2} \\
B_2 &= \tfrac{1}{6} \\
B_4 &= -\tfrac{1}{30} \\
B_6 &= \tfrac{1}{42} \\
B_8 &= -\tfrac{1}{30} \\
B_{10} &= \tfrac{5}{66} \\
B_{12} &= -\tfrac{691}{2730} \\
B_{14} &= \tfrac{7}{6} \\
B_3 &= B_5 = B_7 = \cdots = 0
\end{aligned}$$

On connaît bien sûr des techniques de calcul rapide des nombres de Bernoulli, la plupart récurrentes.

À propos de formules explicites pour calculer rapidement ces nombres, on dispose tout de même du théorème de von Staudt - Clausen qui dit que $B_{2k} + \sum 1/p$ étendue aux nombre premiers p tels que $(p-1)$ divise $(2k)$ est un entier.

Sachant par ailleurs que, pour $k > 0$, on a

$$B_{2k} = (-1)^{k-1} 2(2k)! \frac{\zeta(2k)}{(2\pi)^{2k}}$$

ζ étant bien sûr la fonction de Riemann, on peut programmer le problème par un procédé hybride sans utiliser de récurrence (d'abord approcher, en multi-précision, puis obtenir la valeur rationnelle exacte grâce à von Staudt).

On a alors comme valeurs pour k variant de 2 à 9 (inclus) :

$\sum_{i=1}^{n} i^2 = \frac{1}{6} n(n+1)(2n+1)$
$\sum_{i=1}^{n} i^3 = \frac{1}{4} n^2(n+1)^2$
$\sum_{i=1}^{n} i^4 = \frac{1}{30} n(2n+1)(n+1)(3n^2+3n-1)$
$\sum_{i=1}^{n} i^5 = \frac{1}{12} n^2(2n^2+2n-1)(n+1)^2$
$\sum_{i=1}^{n} i^6 = \frac{1}{42} n(2n+1)(n+1)(3n^3+6n^3-3n+1)$
$\sum_{i=1}^{n} i^7 = \frac{1}{24} n^2(3n^4+6n^3-n^2-4n+2)(n+1)^2$
$\sum_{i=1}^{n} i^8 = \frac{1}{90} n(2n+1)(n+1)(5n^6+15n^5+5n^4-15n^3-n^2+9n-3)$
$\sum_{i=1}^{n} i^9 = \frac{1}{20} n^2(n^2+n-1)(2n^4+4n^3-n^2-3(n+3)(n+1)^2$

On peut également dévélopper une approche algébrique assez détaillée. On montre dans un premier temps que cette somme $S(r,n)$ peut s'écrire $\sum_{i=0}^{r} H(r,i) n^{i+1}$.

Le fait qu'un tel polynôme existe découle de l'observation suivante : Dans le sous espace vectoriel des polynômes de degrés inférieurs ou égaux à r, les u_i (i variant de 0 à r) définis par : $u_i = (1+X)^{i+1} - X^{i+1}$ constituent une base.

Si l'on nomme $(H(r,i); i = 0, \ldots, r)$ le système des coordonnées de $(1+X)^r$ on retrouve que $H(r,i) n^{i+1} = \sum_{p=1}^{n} p^r$

L'utilisation de cette base permet de déduire que les $H(r,i)$ sont solution du système $Mv = w$.

où :

- C_y^x désignant le coefficient binômial bien connu...
- M est la matrice triangulaire supérieure dont le terme $m(l,c)$ vaut C_{l-1}^c pour c entre 1 et $r+1$ et l entre 1 et c par exemple dans le cas $r = 2$, la matrice est :

$$\begin{pmatrix} C_0^1 & C_0^2 & C_0^3 \\ 0 & C_1^2 & C_1^3 \\ 0 & 0 & C_2^3 \end{pmatrix}$$

– v est le vecteur dont les composantes sont $H(r,0)$ à $H(r,r)$
– w est le vecteur dont les composantes sont C_0^r à C_r^r.

Le pivot de Gauss donne rapidement les premières valeurs :

$$\begin{array}{ll} H(r,r) & = \frac{1}{r+1} \\ H(r,r-1) & = \frac{1}{2} \quad \text{(constant... c'est marrant)} \\ H(r,r-2) & = \frac{r}{12} \\ H(r,r-3) & = 0 \\ H(r,r-4) & = \frac{C_r^3}{120} \\ H(r,r-5) & = 0 \\ H(r,r-6) & = \frac{C_5^3}{252} \\ H(r,r-7) & = 0 \end{array}$$

Il y a là de quoi traiter rapidement jusqu'au cas $r = 7$ mais... En dérivant l'égalité polynômiale somme des $u_i = (1+X)^r$ on trouve $H(r,i) = \frac{r}{i+1} H(r-1, i-1)$

On en déduit $H(r,i) = \frac{C_i^r}{i+1} H(r-i, 0)$.

Cette relation limite les recherches aux $H(k,0)$ mais, dans le système initial, c'est celui que le pivot de Gauss donne en dernier...

La relation donnant $H(r,i)$ en fonction de $H(r,i+1), \ldots, H(r,r)$ devient

$$H(r,0) = 1 - \sum_{j=0}^{r-1} \frac{C_j^{r+1}}{r+1} H(j,0)$$

et l'on a $H(0,0) = 1$.

Cette dernière relation permet de programmer rapidement le calcul des coefficients dans les cas où r est assez grand...

Par ailleurs, on a remarqué qu'il y avait pas mal de 0... En fait, on a : pour tout k : $H(2k+3, 0) = 0$. Cela peut se montrer de plusieurs façons... 'égalité polynômiale avec $X = 1$ puis $X = -1$ donne :

$$\sum_{\substack{j=0 \\ j \text{ pair}}}^{k} H(k, k-j) = \sum_{\substack{j=0 \\ j \text{ impair}}}^{k} H(k, k-j) = 1/2$$

en bricolant un peu avec les termes connus on trouve que

$$\sum_{\substack{j=0 \\ j \text{ impair}}} C_j^{k+1} H(j,0)$$

est nulle. cette égalité est valable pour tout k (au moins 3...) et l'on sait que $H(3,0)$ est nul... donc, de proche en proche...

2.2 Les nombres et les polynômes de Bernoulli.

Cet article fournit des informations sur les nombres de Bernouilli ainsi que quelques considérations sur les polynômes attribués au même mathématicien. Les démonstrations ne sont pas réellement faites mais des pistes sont fournies.

Les polynômes de Bernoulli jouent un rôle central dans la formule d'Euler-Mac Laurin qui a de nombreuses applications en analyse numérique (accélération de la convergence de certaines séries numériques, intégration numérique... entre autres)

Je conseille à ceux qui veulent s'y plonger de le faire avec une feuille de papier pour noter les choses au fur et à mesure...

... après avoir imprimé l'ensemble.

En effet le format texte constitue vite un frein à la compréhension...

Habituellement on définit les polynômes de Bernoulli B_n par $B_0(X) = 1$ et $B_n(1) = B_n(0)$ pour $n \leq 2$. Alors $B_(n+1)$ a pour dérivé B_n et l'on pose alors $b_n = B_n(0)$ On prouve facilement que B_n est ainsi parfaparagraphDémonstration.nt défini... Les nombres de Bernoulli sont les $i!b_i$ Notons que des définitions variantes existent.

On en déduit les propriétés suivantes :

1. $B_n(X) = \sum_{j=0}^{n} b_{n-j} \frac{X^j}{j!}$
2. Pour tout n et tout X : $B_n(1-X) = (-1)^n B_n(X)$
3. Pour tout $p > 0$, tout n, tout X, on a

$$B_n(X) = p^{n-1} \sum_{j=0}^{p-1} B_n(\frac{x+j}{p})$$

4. $B_(n+1)(X+1) - B_(n+1)(X) = (X^n)/n!$
5. $1^n + 2^n + \ldots M^n = n!(B_{n+1}(m+1) - B_{n+1}(0))$

Preuves :

1. Taylor pour les polynômes...
2. On veut que B_n soit égal à un certain polynôme... On montre que ce polynôme vérifie la « définition »de B_n...
3. pareil qu'au 2.
4. Par récurrence...
5. avec 4.

Les premiers nombres de Beroulli sont $1\,;\ -\frac{1}{2}\,;\ \frac{1}{6}\,;\ 0\,;\ \frac{-1}{30}\,;\ 0\,;\ \frac{1}{42}\,;\ \ldots$ De $B_n(1) = B_n(0)$ on déduit $b_n = -\sum_{j=1}^{n} \frac{b_(n-j)}{(j+1)!}$ égalité qui permet de trouver une formule de récurrence pour les nombres de Bernoulli.

2.3 Expression par radicaux des racines d'un polynôme de degré n.

2.3.1 Historique.

Antiquité. La notion moderne d'équation n'émerge, en fait, qu'assez tardivement dans l'Histoire, et ce que l'on appelle « algèbre »dans l'Antiquité se limite dans une large mesure à la résolution de problèmes de degré n, c'est-à-dire de problèmes numériques concrets visant à déterminer une certaine quantité, qui pour nous dépend algébriquement des données.

À cet égard, les mathématiciens Mésopotamiens sont particulièrement avancés. Les Babyloniens, par exemple, sans pour autant dégager de « formule générale », disposent de méthodes systématiques de résolution des problèmes de degré 1 et 2, dont certains mettent même en jeu des systèmes, linéaires (ou non). Dans quelques cas particuliers, ils résolvent même des problèmes de degré 3 et 4.

Par comparaison, l'algèbre égyptienne de la même époque (début du II^e millénaire avant notre ère) peut paraître assez rudimentaire. Pénalisés par un système de numération inadapté et des notations lourdes (pour les fractions par exemple), les égyptiens résolvent au cas par cas des problèmes du premier degré uniquement, et cela par des méthodes qui nous sembleraient de peu de rigueur.

Les Grecs eux-mêmes, à cause peut-être de la fameuse crise des irrationnels, se méfient, un peu, de l'algèbre et l'ont peu fait progresser. Ni les Pythagoriciens (plus préoccupés des entiers), ni les successeurs d'Euclide (qui se consacrent avant tout à la géométrie) ne s'y sont beaucoup intéressés. Le dixième livre des *éléments*constitue néanmoins le fondement de nombreuses recherches algébriques du Moyen-âge.

Exception éclatante : Diophante d'Alexandrie, mathématicien du III^e siècle après J.-C., dont les *Arithmétiques*constituent peut-être le premier traité « d'algèbre classique ». Il y introduit en effet la notion d'équation algébrique, c'est-à-dire la relation entre les puissances successives d'un nombre inconnu (arithmos) qu'il s'agit de déterminer par transformations successives de la relation. Sa démarche déductive est certes en recul par rapport à la méthode axiomatique d'Euclide, mais il se permet ainsi de considérer les

fractions et les irrationnels comme des nombres à part entière, ce qui renforce la généralité de ses méthodes.

Du IV^e au XIV^e siècle S'inspirant peut-être de la numération chinoise, les Indiens inventent un système décimal de position comportant le zéro et les relatifs négatifs dès le IV^e ou le V^e siècle après J.-C. et qui permet des notations algébriques bien plus élégantes.

Ainsi, au VII^e siècle, le mathématicien Brahmagupta, dans son traité *Brahmasphutasiddhanta*énonce-t-il des règles générales de transformation des expressions algébriques, contenant éventuellement des quantités négatives ou nulles, et donne explicitement la solution de l'équation générale de degré 2. Au XII^e siècle, Bhaskara (à ne pas confondre avec son homonyme contemporain de Brahmagupta) généralise ces méthodes, qu'il étend à des équations particulières de degré supérieur à 2. Il tient compte, en outre, de la seconde racine des équations de degré 2.

L'algèbre arabe fait, en quelque sorte, la synthèse des mathématiques grecques et indiennes, et constitue le sujet de prédilection des mathématiciens arabes. Au IX^e siècle, al-Khwarizmi remarque que la transformation des équations constitue une théorie à part entière, dont il décrit les principes dans le *Kitab al jabr wa-l-muqabla* dont l'algèbre tire son nom. Il reprend les méthodes de Diophante et la numération indienne, qu'il contribue à populariser. Néanmoins, il est encore gêné par les nombres négatifs, ce qui n'est pas le cas de son principal successeur, Abu Kamil.

Forts des progrès de l'algèbre arabe vers l'abstraction, al-Karaji à la fin du X^e siècle et al-Samaw'al au XIIe siècle développent une puissante arithmétique des polynômes et des fractions rationnelles : multiplication, division, et même extraction de racines et une sorte de développement limité en $O(1/x^n)$. Dès le XIe siècle, l'équation cubique suscite par ailleurs un vif intérêt. Le persan Umar al-Khayyam donne notamment de nombreux éléments d'une étude géométrique, voire en des termes modernes « analytique », du problème.

Les travaux de Léonard de Pise (le célèbre Fibonacci), au début du $XIII^e$ siècle diffusent en Europe le savoir algébrique arabe. Son *Liber Abaci* constitue la source principale des nombreuses recherches de ses successeurs. De plus, il propose avec l'empereur Frédéric II des sortes de « défis scientifiques »sous la forme de problèmes réunis dans le « Liber Quadratorum »et comprenant la résolution de plusieurs équations de degré 3.

La Renaissance. à la Renaissance, les mathématiques, et surtout l'algèbre, se développent rapidement en Italie, sur la base de l'héritage gréco- arabe.

Les premiers progrès s'effectuent sur le terrain du symbolisme, de plus en plus concis et suggestif. Nicolas Chuquet et Luca Pacioli présentent sous une forme concise les résultats classiques. C'est celui-ci qui introduit la notation « cossique »des équations algébriques. Jusqu'au XVIIe siècle, beaucoup s'attachent à perfectionner un symbolisme, qui atteint à peu près sa forme actuelle avec Descartes...

Mais le grand apport des mathématiciens italiens à l'algèbre est la résolution par radicaux des équations de degré 3 et 4à la toute fin du XV^e siècle, Scipione dal Ferro parvient à l'expression par radicaux des racines de l'équation cubique sans terme en x^2 (ce qui est équivalent à la résolution complète, mais il semblerait qu'il ne le savait pas). Quoi qu'il en soit, dans une tradition médiévale un peu surannée, il choisit de garder sa découverte secrète. Il la confie à sa mort à son élève Fior qui ne la divulgue pas. En 1535, Tartaglia, établi à Venise comme professeur de mathématiques, propose une méthode de résolution des équations cubiques sans terme en x, mais Fior lui en conteste la priorité. Ce genre de querelles se réglaient en des défis. Fior met Tartaglia au défi de résoudre l'équation sans terme en x^2, et celui-ci y parvient, assurant sa victoire.

Quelques années plus tard, un médecin et mathématicien milanais, Cardan, vient trouver Tartaglia pour obtenir l'autorisation de publier ses formules dans sa grande somme mathématique, *l'Ars Magna*[Car45]. Tartaglia refuse, mais devant l'insistance de Cardan, il consent à lui exposer sa méthode, avec la promesse qu'elle ne sera pas publiée. Malgré tout, les fameuses « formules de Cardan »apparaissent bien dans *l'Ars Magna*, et une violente querelle s'ensuit qui ne prend fin qu'en 1548. On trouve également dans le traité de Cardan la solution de l'équation générale de degré 4 que l'on peut attribuer avec certitude à l'élève de Cardan, Ferrari (auquel on pense en fait devoir un grand nombre des résultats publiés par Cardan)...

Une particularité de la méthode de Tartaglia est de faire intervenir, au cours du calcul, des racines carrés de nombres négatifs, ce qu'il avait du mal à prendre en considération. Le premier à avoir véritablement admis les complexes en tant que nombres, plutôt qu'artifices calculatoires, est Bombelli. Il présente les règles générales de calcul sur les complexes et toutes les récents progrès de l'algèbre peu avant sa mort, dans *Algebra, parta maggiore dell'arithmetica*[Bom72].

Vers l'algèbre moderne. Après le XVI^e siècle, les mathématiciens semblent se désintéresser de l'algèbre, pour se consacrer plutôt à la géométrie et à la toute jeune analyse.

Les diverses tentatives de résolution de l'équation de degré 5 sont in-

fructueuses, de même que les essais de démonstration de la « conjecture »de Girard, selon laquelle toute équation algébrique de degré n admet exactement n racines complexes distinctes ou confondues. Étrangement, le rigoureux Descartes semble avoir considéré ce résultat comme évident.

En tous les cas, l'algèbre pure fait peu de progrès jusqu'à la seconde moitié du $XVII^e$ siècle. Il y a cependant des recherches liées à l'analyse, sur l'approximation des racines par exemple (recherches de Rolle, de Newton). Le tournant se situe néanmoins aux environs de 1770, lorsque Lagrange et Vandermonde entament des recherches sur la Théorie des Substitutions.

Lagrange saisit, en particulier, l'importance de la notion de permutations sur la famille des racines d'une équation algébrique, et développe avec Waring l'idée selon laquelle, si la conjecture de Girard est vraie, alors les coefficients d'une équation algébrique sont au signe près les fonctions symétriques élémentaires des racines (Viète, puis Girard avaient remarqué ce résultat longtemps auparavant, dans quelques cas particuliers). Ce résultat lui permet de présenter une méthode élégante de résolution des équations de degré 3 et 4. Vandermonde étudie les fonctions invariantes par permutations circulaires, et en déduit les solutions par radicaux de certaines équations particulières (au groupe de Galois cyclique) telles que $x^{11} - 1 = 0$.

Une nouvelle étape est franchie avec la première démonstration rigoureuse de la conjecture de Girard, publiée par Gauss en 1799. D'Alembert s'y était essayé avant lui, mais sa démonstration était incomplète. Gauss perçoit par ailleurs l'importance du *groupement des opérations* (selon l'expression de Galois) et dans ses recherches sur les formes quadratiques et sur l'arithmétique modulaire se dégagent déjà les concepts qui fondent l'algèbre moderne. Il développe de plus les idées de Vandermonde, et montre que le polygône à dix-sept côtés est constructible à la règle et au compas.

Ruffini, appliqué à l'étude de la théorie des substitutions, effectue des recherches sur les valeurs prises par une fonction de cinq variables par toutes les permutations de ces variables. Il parvient à des résultats, généralisés par Cauchy, qui l'amènent à conclure, en 1813, à l'impossibilité de résoudre l'équation de degré 5 par radicaux.

Ses arguments ne suffisent pas pour une démonstration rigoureuse, cependant Abel apporte des arguments plus probants sur ce point, et parvient, aux alentours de 1826, à l'impossibilité de la résolution par radicaux de l'équation générale de degré premier supérieur ou égal à 5.

Cependant, son raisonnement présente des difficultés, et il maîtrise mal ses méthodes, qui ne se formalisent correctement que dans le cadre de la théorie des corps (laquelle n'émerge que bien plus tard). Galois est le premier à adopter une méthode complètement générale, en introduisant la notion de

groupe d'une équation (l'ensemble des permutations[1]des racines conservant les relations algébriques entre celles-ci).

Dans des mémoires successifs rédigés à partir de 1830, il dégage le critère général de résolubilité par radicaux d'une équation algébrique. Ainsi Galois clôt-il définitivement la question essentielle de l'algèbre classique, tout en posant, plus encore que Gauss, les jalons de l'algèbre moderne.

2.3.2 Démonstrations.

Soit $P_n(x) = x^n + \sum_{k=0}^{n} a_k x^k$ un polynôme de degré n. Les a_k sont soit des réels soit des complexes. On cherche à exprimer les racines de P_n en fonction des a_k. C'est ce que l'on appelle une résolution par radicaux.

Notes :

- Les racines des nombres négatifs ou complexes sont "admises"
- Je ne traite que les cas où les polynomes sont unitaire (leur coefficient de plus haut degré est 1. Si ce n'est pas le cas il suffit de diviser le polynôme par son coefficient de plus haut degré et d'appliquer la méthode que je développe plus bas).

Le polynôme est de degré $n = 1$. On part de l'équation : $x + b = 0$. La solution est bien évidement $x = -b$.

Le polynôme est de degré $n = 2$. On part de l'équation : $x^2 + ax + b = 0$. Deux possibiltés se présentent : (i) $a = 0$ et (ii) $a \neq 0$

(i) $a = 0$. L'équation s'écrit $x^2 = -b$ Les deux solutions sont donc : $x_1 = \sqrt{-b}$ et $x_2 = -\sqrt{-b}$

(ii) $a \neq 0$. L'équation s'ecrit $x^2 + ax + b = 0$ Cette équation peut s'écrire : $x^2 + ax + (1/4)a^2 + b - (1/4)a^2 = 0$ Or l'on a : $(x + a/2)^2 = x^2 + ax + (1/4)a^2$ Donc on peut écrire dans la première équation : $(x + a/2)^2 = (a^2 - 4b)/4$

Ce qui revient au cas (i). Donc l'équation s'écrit : $x + a/2 = \sqrt{(1/4)a^2 - b}$ ou $x + a/2 = -\sqrt{(1/4)a^2 - b}$

On en déduit les solutions de l'equation : $x_1 = -a/2 + \sqrt{(1/4)a^2 - b}$ et $x_2 = -a/2 - \sqrt{(1/4)a^2 - b}$

[1]Pour ceux qui liraient Galois, il est intéressant de noter qu'il appelle *permutation* une certaine disposition de n lettres, et emploie le terme de substitution pour désigner l'opération de changement de cette disposition.

Le polynôme est de degré $n = 3$. On part de l'équation : $x^3 + ax^2 + bx + c = 0$. On effectue un changement de variable $x = z - a/3$. On obtient alors une équation du type : $z^3 + pz + q = 0$ Avec : $p = b - (1/3)a^2$ et $q = (2/27)a^3 - (1/3)ab + c$ [2]. Pour l'équation en z, deux cas sont possibles : (i) $p = 0$, (ii) $p \neq 0$.

(i) $p = 0$. L'équation s'écrit donc $z^3 = -q$ Cette équation a trois solutions dans $\mathbb{C}$: $z_1 = \sqrt[3]{(-q)}$, $z_2 = jz_1$ et $z_3 = (j^2)z1$. Où $j = \frac{-1+i\sqrt{3}}{2}$

(ii) $p \neq 0$. L'équation est $z^3+pz+q = 0$. On effectue un autre changement de variable z = u + v. Avec u non-nul. Et l'équation s'écrit : $u^3 + v^3 + q + (3uv + p)(u + v) = 0$ on s'intéresse alors au système suivant[3] :

$$\left\{ \begin{array}{rcl} u^3 + v^3 + q & = & 0 \\ 3uv + p & = & 0 \end{array} \right. \qquad [S]$$

Le système [S] est équivalent à :

$$\left\{ \begin{array}{rcl} u^6 + qu^3 - (1/27)p^3 & = & 0 \\ v & = & -p/(3u) \end{array} \right.$$

Encore (!) un changement de variable dans la première équation. On pose $y = u^3$, et celle-ci devient : $y^2 + qy - (1/27)p^3$. De là, une solution est donc : $y = -q/2 + \sqrt{((1/2)q^2 + (1/27)p^3)}$

Donc, il ne reste plus qu'à trouver les solutions de $u^3 = y$. C'est le cas (i). On a donc comme solutions :

$$\left\{ \begin{array}{rcl c rcl} u_1 & = & \sqrt[3]{y} & \text{et} & v_1 & = & -p/(3u_1) \\ u_2 & = & ju_1 & \text{et} & v_2 & = & jv_2 \\ u_3 & = & j^2u_1 & \text{et} & v_3 & = & (j^2)v_3 \end{array} \right.$$

De là, on a $z_1 = u_1 + v_1$, $z_2 = u_2 + v_2$ et $z_3 = u_3 + v_3$. Ce qui nous donne les solutions pour $x\ldots$

Le polynôme est de degré $n = 4$. On part de l'équation $x^4+ax^3+bx^2+cx+d = 0$. On effectue le changement de variable $x = z - a/4$. On obtient une équation réduite de la forme : $z^4+pz^2+qz+r = 0$ Avec $p = b-(3/8)a^2$; $q = c - ab/2 + (1/8)a^3$ et $r = d - ac/4 + (1/16)ba^2 - (3/256)a^4$

On a deux cas pour l'équation en z : (i) $q = 0$ et (ii) $q \neq 0$.

[2] Une fois que l'on a une solution z_0 de l'équation en z, alors $x_0 = z_0 - a/3$ est solution de l'équation en x.

[3] si (u_0, v_0) est solution du sytème [S], on remarque que $z_0 = u_0 + v_0$ est solution de l'équation 3.

(i) $q = 0$. L'équation s'écrit $z^4 + pz^2 + r = 0$. C'est ce que l'on appelle une équation bicarrée. On pose $y = z^2$ et l'equation devient $y^2 + py + r = 0$ Les solutions sont donc : $y_1 = -p/2 + \sqrt{(1/4)p^2 - r}$ et $y_2 = -p/2 - \sqrt{(1/4)p^2 - r}$

De là les valeurs de z sont : $z_1 = sqrt(y_1)$; $z_2 = -\sqrt{(y_1)}$; $z_3 = sqrt(y_2)$ et $z_4 = -sqrt(y_2)$.

(ii) $q \neq 0$. L'équation s'écrit $z^4 + pz^2 + qz + r = 0$. On pose alors $2P - Q^2 = p$; $-2QR = q$ et $P^2 - R^2 = r$. On a alors $(z^2 + P)^2 - (Qz + R)^2 = 0$. Ce qui est une autre façon d'écrire $z^4 + pz^2 + qz + r = 0$.

Si l'on arrive à determiner le triplet (P_0, Q_0, R_0) alors trouver les solutions de l'équation réduite revient à résoudre :

$$\left\{ \begin{array}{rcl} z^2 + P_0 + Q_0 z + R_0 & = & 0 \quad \text{ou} \\ z^2 + P_0 - Q_0 z - R_0 & = & 0 \end{array} \right.$$

On peut donc trouver z. Il reste donc à determiner P, Q et R. C'est à dire à résoudre le système

$$\left\{ \begin{array}{rcl} 2P - Q^2 & = & p \\ -2QR & = & q \\ P^2 - R^2 & = & r \end{array} \right. \qquad [S]$$

Ce système revient à :

$$\left\{ \begin{array}{rcl} Q^2 & = & q^2/(4P^2 - r) \\ R^2 & = & P^2 - r \\ QR & = & -q/2 \end{array} \right.$$

Ce qui revient à résoudre l'équation (en P) suivante : $p^3 - (p/2)P^2 - rP + pr/2 - (1/8)q^2 = 0$ De là, on trouve (pas si) facilement P_0. Et grâce au système [S] on peut lui associer un couple (Q_0, R_0) et donc trouver $z \ldots$ (ouf !)

Le polynôme est de degré $n > 4$. Au XIXè siècle, Abel a montré que la résolution par radicaux de l'équation du cinquième degré était impossible dans le cas général. Indépendamment, Galois a généralisé cette démonstration à l'ensemble des cas où n est supérieur ou égal à 5.

Cette démonstration demande un recours à la théorie de Galois, elle ne sera pas developpée ici. Cette théorie est très bien développée dans *Équations algébriques et théorie de Galois* [Mut] et, en anglais, dans *Galois Theory* [Ste].

Chapitre 3

Algèbre linéaire

3.1 Carrés magiques

Cet article présente une méthode générale pour construire des carrés magiques. Pour un historique, voir [Kan].

3.1.1 Espace O_n

On appelle $\mathcal{M}_n$ l'ensemble des matrices carrés de dimension (n, n) (n entier naturel). On appelle O_n les matrices A de $\mathcal{M}_n$ dont la somme des termes horizontaux et verticaux sont égales, et l'on note cette somme $s(A)$.

O_n est une sous-algèbre de $\mathcal{M}_n$

Démonstration.

On commence par montrer que O_n est un sous-espace vectoriel. Il suffit de vérifier la stabilité par somme et produit. On montre donc que quelque soit λ, quelque soient A et B de O_n, $\lambda A + B$ appartienne à O_n.

Par exemple, une base évidente de O_2 est formée de

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ et } \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Il faut enfin vérifier la stabilité par la multiplication avec un réel.

□

O_n est de dimension n

Dans Rn rapporté à sa base canonique, f est l'endomorphisme de matrice A. Le sous-espace vectoriel G engendré par le vecteur $(1, 1, 1, 1...1)$ et l'hyperplan H orthogonal au vecteur $(1, 1, 1, 1...1)$ sont stables par f. La première stabilité traduit la linéarité des lignes et la seconde celle des colonnes de A.

L'intersection de G et H est le vecteur nul, donc F et G supplémentaires. Donc O_n est de dimension $n = (n-1) + 1 = n$.

3.1.2 Matrices magiques

A est dite carrée magique si elle appartient à O_n et que la somme des termes des deux diagonales valent $s(A)$.

Base de l'espace vectoriel

On vérifie immédiatement qu'il s'agit d'un sous-espace vectoriel.

L'ensemble des matrices carrées magiques antisymétriques (${}^tA = -A$) est

$$\left\{ \begin{pmatrix} 0 & a & -a \\ -a & 0 & a \\ a & -a & 0 \end{pmatrix}, a \in \mathbb{R} \right\}$$

L'ensemble des matrices carrées magiques symétriques (${}^tA = A$) avec $s(A) = 0$ est

$$\left\{ \begin{pmatrix} b & -b & 0 \\ -b & 0 & b \\ 0 & b & -b \end{pmatrix}, b \in \mathbb{R} \right\}$$

Toute matrice A peut s'écrire $A = A_s + A_a$ où A_s est symétrique et A_a est antisymétrique : $A_s = \frac{1}{2}(A + {}^t A)$; $A_a = \frac{1}{2}(A - {}^t A)$.

Puisque A magique $\Leftrightarrow {}^tA$ magique, A magique si et seulement si A_s et A_a sont magiques.

Exemple en dimension$n = 3$

Dans le cas de la dimension $n = 3$, on déduit que B, C, J forment une base :

$$B = \begin{pmatrix} 1 & -2 & 1 \\ 0 & 0 & 0 \\ -1 & 2 & -1 \end{pmatrix} \quad C = \begin{pmatrix} 1 & 0 & -1 \\ -2 & 0 & 2 \\ 1 & 0 & -1 \end{pmatrix} \quad J = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

Toute matrice de la forme $xB + yC + zJ$, (x, y, z) entiers relatifs, sera magique.

Par exemple, fabriquons un carré magique. Au hasard, prenons $x = 1$ et $y = 2$ (pour le moment, $z = 0$).

$B + 2\,C$ vaut

$$\begin{pmatrix} 3 & -2 & -1 \\ -4 & 0 & 4 \\ 1 & 2 & -3 \end{pmatrix}$$

Et pour ne pas avoir de cases contenant un nombre négatif on ajoute $4\,J$ ($z = 4$), ce qui donne

$$\begin{pmatrix} 7 & 2 & 3 \\ 0 & 4 & 8 \\ 5 & 6 & 1 \end{pmatrix}$$

3.1.3 Annexe

Algorithme pour des matrices carrées de taille (n, n) où n est impair.

```
i:=1;
j:=(n+1)/2
Pour k=1..n^2 Faire
| A[i,j]:=k;
| Si i=1 Alors
| | i':=n
| Sinon
| | i':=i-1
| Finsi
| Si j=n Alors
| | j':=1
| Sinon
| | j':=j+1
| Finsi
|
| Si A[i',j'] !=0 Alors
| | j':=j
| | Si i=n Alors
| | | i':=1
| | Sinon
| | | i':=i+1
| | Finsi
| Finsi
```

```
| i,j:=i',j'
Fait
```

Télécharger l'algorithme pour Maple V'R4

Chapitre 4

Théorie des nombres.

4.1 Le petit théorème de Fermat.

Enoncé : *Si p est un nombre premier, et n un entier quelconque non divisible par p, alors le reste de la division de n^{p-1} par p est égal à 1.*

Par exemple, si on prend $p = 1999$ qui est premier, et $n = 1665$, année de la mort de Fermat, qui n'est pas divisible par 1999, alors le théorème dit que le reste de la division de 1665^{1998} par 1999 est égal à 1.

Ce théorème a sans aucun doute été démontré par Fermat. (Il me semble qu'on n'a pas retrouvé la démonstration de Fermat, mais Euler en a publié une dès le $XVIII^e$ siècle).

Démonstration.

Soit p un nombre premier.

Si $k < p$, $k!(p-k)!C_p^k = p!$, donc p divise $k!(p-k)!C_p^k$. Comme $\forall j \leq k$, $j \wedge p = 1$ et $\forall j \leq (p-k)$, $j \wedge p = 1$, on a $p|C_p^k$.

Cela prouve le lemme suivant : Si $0 < k < p$, $C_p^k = \frac{p!}{k! \times (p-k)!}$ est multiple de p.

Ainsi, le binôme de Newton donne tout de suite : $\forall x \in \mathbb{N}$, $(1+x)^p \equiv 1+x^p \mod p$.

Première façon Soit n un entier quelconque. Sommons pour x variant de 0 à $n-1$, il reste $n^p \equiv n \mod p$.

Deuxième façon, par récurence Pour tout n, notons $H(n)$ l'hypothèse « $n^p \equiv n \mod p$ ».

Pour n=0, c'est évident.

Soit n un entier quelconque. Supposons $H(n)$ et montrons $H(n+1)$. $(1+n)^p \equiv 1 + n^p \mod p$ d'après le binôme de Newton. Or $n^p \equiv n \mod p$ d'après $H(n)$. Ainsi $(1+n)^p \equiv 1 + n \mod [p]$ ce qui prouve $H(n)$ et achève la démonstration.

□

4.2 Le grand théorème de Fermat.

Le "grand" ou "dernier" théorème de Fermat dit ceci : alors qu'il existe des carrés qui sont somme de deux carrés (par exemple $5^2 = 3^2 + 4^2$, $13^2 = 12^2 + 5^2$), il n'existe aucun cube (non nul) qui soit somme de deux cubes (non nuls). De même, il n'existe aucune puissance quatrième qui soit somme de deux puissances quatrièmes (non nulles), etc.

Enoncé : *En général, pour $n > 2$, il n'existe aucun triplet $\{x, y, z\}$ d'entiers non nuls, tels que $x^n + y^n = z^n$.*

En abrégé, pour $n > 2$ l'équation $x^n + y^n = z^n$ n'a pas de solution entière non triviale.

4.2.1 Histoire liée au théorème.

Contrairement au petit théorème, il s'agit d'un résultat extrêmement difficile, dont Fermat n'a pas publié de démonstration, et qu'il n'a probablement pas démontré. Fermat n'a même jamais affirmé publiquement l'avoir démontré. Il écrivit dans une marge du livre II des Oeuvres de Diophante :

> j'ai découvert une démonstration merveilleuse, mais je n'ai pas la place de la mettre dans la marge.

Le livre et cette annotation ont été publiés après sa mort, par son fils.

En dépit de la simplicité de son énoncé, ce théorème est tellement difficile à prouver que les plus grands mathématiciens s'y sont cassé les dents pendant 358 ans exactement, d'après le livre de Simon Singh. On notera cependant, que :

- Euler (1707-1783) le démontre pour n = 3 et ses multiples
- Legendre (1752-1833) pour n = 5
- Kummer (1810-1893) pour tout n tel que 2¡n¡100
- En 1993, une démonstration est enfin publiée : celle d'Andew Wiles (1953-).

4.2.2 Quelques références.

Une page web en anglais bien documentée : [**?**].

Simon Singh a écrit un très beau livre : *Le dernier Théorème de Fermat*[**?**], à partir d'un documentaire qu'il devait réaliser sur le sujet pour la BBC. C'est principalement un roman.

Il y a aussi le livre de Yves Hellegouarch : *Invitation aux mathématiques de Fermat-Wiles*[Hel], très bien fait, mais demandant un niveau licence ou maîtrise de mathématiques, pour en comprendre toutes les finesses.

4.3 Les nombres premiers.

Définition : *Un nombre premier est un entier possédant exactement 2 diviseurs. (ces deux diviseurs sont donc 1 et lui-même).*

A noter que la définition de nombre premier exclut 1. Admettre 1 comme nombre premier rendrait faux l'important *Théorème fondamental de l'arithmétique* qui dit qu'il existe une unique manière d'écrire un nombre entier supérieur à 1 comme produit de nombres premiers (sans prendre en compte l'ordre de la multiplication).

4.3.1 En existe-t-il une infinité ?

Oui. En voici une démonstration par contradiction. Noter qu'après le Théorème de Pythagore, c'est sûrement le résultat mathématique pour lequel on connaît le plus de démonstrations différentes.

Supposons qu'il n'en existe qu'un nombre fini n, qui seraient $p_1, \ldots, p_n$. Le produit $p_1 \times \cdots \times p_n$ est divisible par chacun de ces premiers $p_1, \ldots, p_n$. Cela signifie que $p_1 \times \ldots \times p_n + 1$ n'est divisible par aucun d'entre eux. Donc le plus petit diviseur (excepté 1) de ce nombre, qui est forcément un nombre premier, n'est ni p_1, ni p_2, ..., ni p_n. Notre liste ne peut donc pas être complète, il existe donc une infinité de nombres premiers.

4.3.2 Quel est le plus grand que l'on connaisse ?

Actuellement (attention, les records tombent vite !), le plus grand nombre premier (qui n'est pas un nombre premier de Mersenne) connu est

$$302627325 \times 2^{530101}$$

qui a été découvert en 1998 par Nash, Dunaieff, Burrowes, Jobling et Gallot. Il a 159585 décimales. Source : http://www.utm.edu/research/primes/largest.html#largest

4.3.3 Polynôme et nombre premiers.

Il existe un polynôme à coefficients entiers à 26 variables tel que l'ensemble des nombres premiers coïncide avec l'ensemble des valeurs positives prises par

$$P(a, b, c, \ldots, x, y, z)$$

lorsque les 26 variables $a, b, c, \ldots, x, y, z$ parcourent $\mathbb{N}$. Ce polynôme s'écrit :

$$\begin{aligned} P(a, \ldots, z) = (k+2)\{1 - [wz + h + j - q]^2 - [(gk + 2g + k + 1)(h + j) \\ + h - z]^2 - [2n + p + q + z - e]^2 - [16(k+1)^3(k+2)(n+1)^2 \\ + 1 - f^2]^2 - [e^3(e+2)(a+1)^2) + 1 - o^2]^2 - [(a^2 - 1)y^2 + 1 \\ - x^2]^2 - [16r^2y^4(a^2 - 1) + 1 - u^2]^2 - [((a + u^2(u^2 - a))^2 - 1) \\ (n + 4dy)^2 + 1 - (x + cu)^2]^2 - [n + l + v - y]^2 - [(a^2 - 1)l^2 \\ + 1 - m^2]^2 - [ai + k + 1 - l - i]^2 - [p + l(a - n - 1) \\ + b(2an + 2a - n^2 - 2n - 2) - m]^2 - [q + y(a - p - 1) \\ + s(2ap + 2a - p^2 - 2p - 2) - x]^2 - [z + pl(a - p) + t(2ap \\ - p^2 - 1) - pm]^2\} \end{aligned}$$

Rendre $P(a, \ldots, z)$ positif revient à annuler simultanément chacun des crochets dans le "long" facteur $\{1 - [\ldots]^2 - \cdots - [\ldots]^2\}$. Ce facteur se réduit alors à 1, tandis que les conditions ainsi imposées à k font que le facteur "court", à savoir $(k + 2)$, est premier.

On peut lire dans « l'Abrégé d'Histoire des Mathématiques »de Jean Dieudonné[**?**] l'existence d'un polynôme à 21 variables ayant la même forme et les mêmes propriétés que le polynôme déjà cité.

Pour ceux qui voudraient s'initier à ce genre de mathématiques, on peut lire [**?**] ou [**?**]. Voir également La page Web des nombres premiers.

Annexe : Dans le même ordre d'idée il existe un polynôme dont les valeurs positives donnent les nombres de Fibonacci :

$$f(x, y) = 2x(y^4) + (x^2)(y^3) - 2(x^3)(y^2) - (y^5) - y(x^4) + 2y$$

4.3.4 Nombres de Mersenne et nombres premiers.

Au sujet du projet GIMPS (Great Internet Mersenne Prime Search), ce projet consiste à chercher le plus grand nombre premier sous la forme d'un nombre de Mersenne (Les nombres de Mersenne sont les nombres premiers du type $2^p - 1$ où p est lui-même premier).

Le 38ième nombre de Mersenne a été trouvé le 1er juin 1999 par l'équipe de Nayan Hajratwala, George Woltman et Scott Kurowski.

Il s'agit de $2^{6972593} - 1$. Ce nombre compte 2098960 décimales.

Trouverez vous le 39ième nombre de Mersenne ? Vous pouvez vous aussi participer à ce grand projet en faisant tourner sur votre machine (quelle que soit la plateforme) le programme qui a été développé pour ce projet. C'est un programme que vous pouvez faire tourner en tâche de fond et qui utilise les cycles CPU non utilisés (donc cela ne ralentira pas votre ordinateur).

C'est un projet de recherche à plusieurs qui utilise les résultats fournis par les 4200 participants pour déterminer quels nombres restent à tester... C'est un projet qui se fait en collaboration, celui qui a la chance de trouver un nombre de Mersenne inscrit définitivement son nom dans l'histoire des maths... (et gagne aussi 10 000 $ je crois, mais j'espère que vous ne ferez pas ça pour l'argent).

Référence : Plus de renseignements sur :
The Great Internet Mersenne Prime Search.

4.4 ab et $a+b$ premiers entre eux.

Enoncé : *Soient a et b deux nombres entiers relatifs tels qu'ils soient premiers entre eux. Le problème est de montrer que ab et $a+b$ sont premiers entre eux.*

Démonstration. 1

Pour cette démonstration, il faut connaître le lemme d'Euclide :

Lemme d'Euclide : *Soit p un nombre premier, et a, b deux nombres entiers relatifs. Si p divise ab, alors p divise soit a soit b.*

Soit p un nombre premier tel qu'il divise ab et $a+b$. p divise ab, donc par le lemme d'Euclide, p divise soit a, soit b. Supposons que p divise a, alors on a : p divise a et p divise $(a+b)$ donc p divise $(a+b)-a=b$. Donc p divise a et p divise b. Or deux nombres premiers entre eux n'ont pas de facteurs premiers communs. Comme a et b sont premiers entre eux, il vient que ab et $a+b$ sont premiers entre eux.

□

Démonstration. 2

Voici une autre démonstration, qui n'utilise pas les propriétés des nombres premiers, mais uniquement la relation de Bézout.

Lemme (Relation de Bezout) : *a et b sont premiers entre eux, ssi il existe deux nombres entiers relatifs u et v tels que $au + bv = 1$.*

Lemme 1 : *Si a et b sont premiers entre eux, alors $a + b$ est premier avec a et avec b.*

De $au + bv = 1$, on déduit $a(u - v) + (a + b)v = 1$, donc a et $a + b$ sont premiers entre eux. De même pour b et $a + b$.

Lemme 2 : *Si a est premier avec b et avec c, alors a est premier avec bc.*

De $au + bv = 1$, on déduit $acu + bcv = c$.

Donc il existe deux nombres entiers $U = cu$ et $V = v$ tels que $(a)U + (bc)V = c$ donc tout diviseur commun de a et bc divise c, donc divise $pgcd(a, c) = 1$.

Conclusion : Soient a et b premiers entre eux ; alors, par le lemme 1, $a+b$ est premier avec a et avec b, donc, par le lemme 2, $a + b$ est premier avec ab.

□

Démonstration. 3

En partant de la relation de Bézout, comme a et b sont premiers entre eux, il existe u et v deux nombres entiers relatifs tels que $au + bv = 1$ donc

$$\begin{aligned} 1 &= 1^2 = (au + bv)^2 = (au)^2 + 2abuv + (bv)^2 \\ 1 &= (au)^2 + abv^2 + abu^2 + (bv)^2 - abv^2 + 2abuv - abu^2 \\ 1 &= (a + b)(au^2 + bv^2) - ab(u - v)^2 \end{aligned}$$

Or $(au^2 + bv^2)$ et $(u - v)^2$ sont des nombres entiers relatifs, donc par la relation de Bézout, on en déduit que $a + b$ et ab sont premiers entre eux.

□

4.5 Irrationalité de $\sqrt{2}$.

$\sqrt{2}$ est le premier exemple de nombre irrationnel qu'aient rencontré les mathématiciens. Il est connu depuis l'Antiquité grecque et cette découverte a suscité à l'époque beaucoup de perplexité.

Une preuve de ce résultat procède par *contradiction*. Il semble que ce soit d'ailleurs le premier exemple de raisonnement par contradiction dans l'histoire des mathématiques.

4.5.1 Un lemme démonstratif.

Avant d'aborder la preuve proprement dite, nous devons établir ce petit résultat intermédiaire :

Lemme : *Soit n un nombre entier. n est pair si et seulement si n^2 est pair et n est impair si et seulement si n^2 est impair.*

(Le lecteur familier des calculs modulo, reconnaîtra un cas particulier du petit théorème de Fermat : $n^2 = n \mod 2$, voir 4.1.)

Démonstration. lemme

Si n est pair, par définition, il existe un entier k tel que $n = 2k$. On a alors $n^2 = 4k^2$ soit $n^2 = 2(2k^2)$. Ceci montre que n^2 est un nombre pair. Si n est impair, il existe un entier k tel que $n = 2k + 1$. Alors, $n^2 = (2k + 1)^2 = 4k^2 + 4k + 1 et n^2 = 2(2k^2 + 2k) + 1$, ce qui montre que n^2 est impair.

□

4.5.2 La démonstration.

Munis de ce résultat, nous pouvons prouver l'irrationalité de $\sqrt{2}$.

Nous souhaitons raisonner par contradiction, nous allons donc supposer que $\sqrt{2}$ est rationnel. Nous allons montrer que cette hypothèse conduit à une contradiction. Nous en déduirons donc que l'hypothèse est fausse, c'est-à-dire que $\sqrt{2}$ est irrationnel.

Si $\sqrt{2}$ est rationnel, on peut donc écrire $\sqrt{2} = m/n$ où m et n sont deux nombres entiers strictement positifs. Quitte à diviser par leur PGCD, nous pouvons de plus supposer que l'écriture m/n est la forme irréductible de cette fraction. En particulier, m et n ne sont pas simultanément pairs.

Élevons au carré l'égalité $\sqrt{2} = m/n$. Il vient $2 = m^2/n^2$ ou encore $m^2 = 2n^2$. Ainsi, m^2 est un nombre pair. Or, nous avons vu qu'un nombre et son carré ont toujours la même parité. Il s'ensuit que m est lui-même un nombre pair. Nous pouvons donc poser $m = 2m'$.

Notre égalité devient alors $4m'^2 = 2n^2$ ou encore $2m'^2 = n^2.n^2$ est donc un nombre pair. Comme plus haut, nous en déduisons que n est lui-même un nombre pair.

m et n sont donc simultanément pairs, ce qui est contradictoire avec nos hypothèses. Il s'ensuit que la racine carrée de 2 est un nombre irrationnel.

Remarque : Cette démonstartion est souvent présentée comme le protype même de démonstration par l'absurde, mais ce n'en est pas une !

Pour commencer, rappelons que le principe du tiers-exclu est « pour toute proposition P, on a P ou *non* P ». Ce principe ne peut pas être prouvé, il est *indécidable*. Ainsi, certains mathématiciens l'acceptent (la grande majorité des mathématiciens, en fait) et d'autres ne l'acceptent pas (les mathématiciens de l'école intuitionniste pour la plupart).

Le raisonnement par l'absurde, a pour principe : « pour toute proposition P, non non P entraîne P ». On peut facilement montrer que le principe du raisonnement par l'absurde et le principe du tiers-exclu sont équivalents.

Dans notre démonstration, le raisonnement tenu est le suivant : « on suppose P et on essaye d'aboutir à une contradiction, ce qui prouve non P ». C'est alors la définition même du "non" qui est utilisée : par définition de la négation logique, il est *contradictoire* d'avoir P et *non* P. Cette démonstration de l'irrationnalité de *sqrt*2 ne fait pas appel au raisonnement par l'absurde ni au principe du tiers exclu : elle est donc parfaitement admise par les mathématiciens intuitionnistes.

4.6 Irrationalité de la racine d'un nombre premier.

Sachant que la racine carrée de 2 est irrationnelle, on peut s'interroger sur la racine cubique de 2, la racine carrée de 3, d'un nombre premier quelconque.

Chacun de ces nombres est en fait irrationnel, mais pour établir un résultat relativement général sur ces questions, il est utile de recourir à des outils un peu plus élaborés que dans la section précédente : la décomposition factorielle d'un nombre entier et la valuation p-adique sur les entiers.

On rappelle que pour tout nombre premier p la valuation p-adique d'un entier x est le nombre noté $v_p(x)$ défini comme le plus grand entier naturel a tel que p^a divise x. C'est aussi l'exposant de p dans la décomposition de x, en facteurs premiers.

On voit facilement que la valuation p-adique possède la propriété de morphisme suivante : $v_p(xy) = v_p(x) + v_p(y)$, pour tous entiers x et y, et donc aussi $v_p(x^a) = av_p(x)$ pour a entier positif.

Une généralisation du problème de l'irrationalité de la racine carrée de 2 peut se formuler comme suit :

Soit a un nombre entier strictement positif. A quelle condition sur l'entier positif x le nombre $x^{1/a}$ (racine a-ième de x) est-il rationnel ?

Posons $x^{1/a} = m/n$. Il vient $m^a = xn^a$. Pour tout nombre premier p,

on a donc $v_p(m^a) = v_p(xn^a)$ et donc $av_p(m) = v_p(x) + av_p(n)$ ou encore $v_p(x) = a(v_p(m) - v_p(n))$. $v_p(x)$ est ainsi un multiple de a quel que soit le nombre premier p. Il s'ensuit que x est lui-même la puissance a-ième d'un entier.

Il est par ailleurs évident que la racine a-ième d'un nombre qui est puissance a-ième d'un entier est rationnelle. On peut donc affirmer :

$x^{1/a}$ est un nombre rationnel si et seulement si x est la puissance a-ième d'un entier.

Ainsi, en particulier, les nombres entiers dont la racine carrée est rationnelle sont les carrés d'entiers.

4.7 La conjecture de Syracuse.

4.7.1 Présentation de la conjecture.

Le problème de la conjecture de Syracuse, également connue sous les noms de problème de Collatz, Kakutani, ou Ulam, se présente de manière très simple. On se donne une entier naturel n plus grand que 1. S'il est pair, on le divise par deux, s'il est impair, on le multiplie par 3 et on lui ajoute 1 (ce qui revient à lui appliquer la fonction $x \to 3x + 1$). On *conjecture* que l'on finit toujours par trouver la valeur 1 au fil des calculs, valeur à partir de laquelle on restera bloqué dans le cycle $1 - 4 - 2 - 1 - \ldots$.

Cependant, le fait que l'on retrouve toujours 1 n'a pas été démontré et même si on est presque sûr que cela est vrai, quel que soit l'entier n choisi au départ, il n'est pas exclu qu'il existe un entier n ne vérifiant pas cette propriété, d'où le nom de : ¡¡conjecture¿¿ de Syracuse.

Depuis plusieurs dizaines d'années, ce problème est activement étudié par les mathématiciens, mais n'a pas encore été résolu. Les recherches ont cependant bien avancé, comme nous allons le voir plus loin.

Une métaphore éclairante. Comment souvent, les mathématiciens, en travaillant sur ce problème, ont senti que certaines idées étaient récurrentes etont introduit un vocabulaire adapté pour décrire les phénomènes étudiées. Imaginons que l'on vérifie la propriété pour $n = 15$. On obtient : 46, 23, 70, 35, 106, 53, 160, 80, 40, 20, 10, 5, 16, 8, 4, 2, 1.

On appelle cette suite finie d'entiers le VOL, ou la TRAJECTOIRE de 15. Il faut imaginer une représentation de cette suite sur un graphique, l'axe des abscisses figurant l'indice de chaque entier dans la suite, l'axe des ordonnées indiquant l'entier correspondant.

On appelle ETAPE, un nombre de cette suite finie. Ici, par exemple, 80 est une étape du vol de 15. Si la conjecture est vraie, on remarque que la suite atteint une étape maximale, appelé ALTITUDE MAXIMALE du vol. Ici, l'altitude maximale était 160.

On définit également la DUREE de chaque vol comme le nombre d'étapes à franchir avant d'arriver pour la première fois au chiffre 1, et la durée de VOL EN ALTITUDE comme la durée entre le moment où le vol commence, et celui où il repasse sous sa valeur initiale.

4.7.2 Des avancées intéressantes.

A partir de là, il est plus facile de comprendre les dernières avancées dans la résolution du problème. Il a été ainsi démontré que chacune des propositions suivantes était équivalente à l'énoncé de la conjecture de Syracuse elle-même :

Lemme : *Tout vol a une durée finie*
C'est en gros l'énoncé de la conjecture en elle-même.

Lemme : *Tout vol est de durée en altitude finie*
En effet, si ceci est vrai, alors on conclut sur notre problème par récurrence. Pour $n = 2$ on finit par retomber sous 2, c'est à dire à 1. Supposons que n et tous les entiers plus petits quelui vérifient la conjecture. Démontrons que c'est alors le cas de $n+1$: Le vol $n+1$ a une durée en altitude finie, donc, au cours des calculs, on arrive à n, ou à un entier inférieur, que l'on note i. Mais i vérifie la conjecture, donc il aboutit au cycle $4-2-1$. Ainsi $n+1$ aboutit-il aussi à ce cycle. Réciproquement, si la conjecture est vraie, la propriété (2) est bien entendue vraie.

Lemme : *Tout vol a un nombre fini d'étapes paires (resp. impaires)*
On considère le vol n. Après un certain nombre d'étapes, on atteint un dernier nombre pair. On le divise par deux, et on obtient un impair. Mais, si on applique $x \to 3x+1$ à cet impair, on obtiendra un pair, ce qui contredit le fait qu'on ait dépassé les dernier pair, sauf si le nombre impair que l'on vient de trouver est 1. Alors on s'arrête dans les calculs, et n vérifie la conjecture, qui de ce fait est vraie.

Lemme : *Tout vol a un nombre fini d'étapes paires (resp. impaires) en altitude*
Démonstration analogue.

Les mathématiciens désespèrent quant au fait de démontrer directement la conjecture elle-même, et pensent qu'il est moins difficile de montrer que l'une de ces propriétés, équivalentes au problème, est vraie.

On sait par ailleurs montrer que la propriété est vraie pour un très grand nombre d'entiers. On considère par exemple $n = 4k + 1$. En effectuant les calculs, on trouve qu'il devient $12k + 4$, $6k + 2$, $3k + 1$, ce dernier entier étant plus petit que n. Pour $n = 4k$, on descend sous n dès la première étape du calcul. De même que pour $4k + 2$. Il suffirait donc de pouvoir montrer que $4k + 3$ a lui aussi une durée de vol en altitude finie pour conclure que la conjecture est vraie ! On peut affiner ce type de démonstration. En travaillant avec les entiers du type $65536k + i$ (avec i compris entre 0 et 65535), tous les cas aboutissent sauf 1729 cas, donc il ne reste que $2,6\%$ des nombres à étudier.

Des développements plus récents ont montré que pour n assez grand, il existait une constante α telle que au moins n^α des entiers inférieurs à n possèdent la propriété de Syracuse. En 1995, J. Lagarias et D. Applegate démontrèrent ce résultat pour la constante $\alpha = 0,81$. Mais leurs calculs furent menés avec des ordinateurs et sont invérifiables à la main.

4.7.3 Les records établis.

Les records enregistrés au fur et à mesure des recherches ont été soigneusement consignés. C'est à T. Oliveira e Silva que l'on doit les records les plus récents et les plus significatifs. Les records suivants proviennent de sa page web :

Le plus grand nombre testé est :

$$77 \times 2^{50} = 86694292826882048$$

La conjecture a été vérifiée par deux fois avec des ordinateurs jusqu'à

$$n = 2^{51} = 2251799813685248$$

Le record de l'altitude maximale est tenu par le vol 82450591202377887, qui atteint l'entier :

$$875612750096198197075499421245450$$

D'autres records, datant de 1998 et sans doute améliorés depuis : Celui de vol de durée record en vol est celui du vol 100759293214567, de durée 1820 (c'est assez faible, on aurait pu croire que des entiers résisteraient plus). Enfin celui de durée en altitude record est le vol 70665924117439, dont la durée en altitude est de 1177 étapes.

4.7.4 Données heuristiques et indécidabilité du problème.

Si l'on étudie la façon dont les entiers évoluent en terme de parité au cours d'un calcul de Syracuse, on peut avoir une idée de son temps de vol. Ainsi, si l'on obtient souvent des nombres pairs, ils vont être divisés par deux, et on arrivera plus rapidement à 1 (en admettant qu'on arrive toujours à 1).

En tenant compte du fait qu'un nombre impair donne un nombre pair et que ce dernier va être divisé par deux ensuite, et qu'il y a dans l'ensemble des entiers naturels autant de pairs que d'impairs, on estime, au moyen d'un calcul probabiliste, qu'un entier donné est en moyenne multiplié par 3/4 lorsqu'on effectue une étape du calcul. Ceci tend à confirmer la conjecture. Expérimentalement, ce résultat de 3/4 est très bien confirmé, et le modèle statistique semble performant.

Cependant, certains mathématiciens sont arrivés à se poser la question de l'indécidabilité du problème. Ils ont proposé quelques extensions du problème : autoriser les entiers négatifs, ou remplacer $3x+1$ quand x est impair par $qx+1$ avec q un entier impair donné. Pour certaines valeurs de $q > 3$, la conjecture n'est pas vraie.

Mais c'est J. Conway qui a semé le doute. Plutôt que de ce demander si un entier donné, au cours du calcul, était pair ou impair, c'est à dire avait 0 ou 1 pour reste par la division euclidienne par 2, il s'intéresse au reste par la division euclidienne par un entier p et propose alors p formules à employer pour effectuer les calculs selon ce que ce reste soit 0, 1, 2, ..., ou $p-1$. Il a montré que si alors on étendait la conjecture de Syracuse, on arrivait à un problème indécidable.

Finalement, on a vu que beaucoup de résultats tendent à nous faire penser qu'il est *presque* impossible que la conjecture soit fausse. Cependant, il est arrivé dans l'histoire que les mathématiciens aient une telle intuition très forte en faveur d'une conjecture et qu'elle soit en fait fausse. Les résultats de J. Conway montrant que des problèmes très similaires sont indécidables incitent donc à la prudence !

Rérérences Un article de Jean-Paul Delahaye dans *Pour La Science* [**?**].

Et sur le web :

- On the 3x+1 problem par Eric Roosendaal à l'adresse :
- 3x+1 search results de T. Oliveira e Silva

4.8 Les cardinaux des ensembles infinis - I.

4.8.1 Avertissement.

Le sujet est vaste et peut intéresser des gens de niveaux divers (à partir du lycée et bien au-delà). C'est pourquoi cet article est découpé en 2 parties. La première s'adresse au niveau pré-bac, la seconde est accessible à partir du DEUG, environ.

Dans la première partie, l'accent est mis sur la "vulgarisation", parfois au détriment de la rigueur : pour éviter des complications techniques, certaines démonstrations sont "un peu fausses", tout en donnant l'idée principale de la méthode. Chaque fois que c'est le cas, c'est signalé, et on peut se reporter à la seconde partie pour trouver une démonstration (à priori) rigoureuse. La seconde partie complète la première, avec des outils plus abstraits et plus puissants. Elle s'efforce à la rigueur, mais elle peut comporter des erreurs ou imprécisions.

Merci enfin à David Madore (David.Madore@ens.fr), spécialiste du sujet, d'avoir corrigé et complété une première mouture de cet article. Merci également à tous les lecteurs de fr.sci.maths qui m'ont indiqué des corrections diverses.

Rappel :

- $\mathbb{N}$: ensemble des entiers naturels $= \{0, 1, 2, 3, \ldots\}$
- $\mathbb{Z}$: ensemble des entiers relatifs (signés) $= \{\ldots, -1, 0, 1, \ldots\}$
- $\mathbb{Q}$: ensemble des nombres rationnels (fractions)
- $\mathbb{R}$: ensemble des nombres réels
- $\mathbb{C}$: ensemble des nombres complexes

L'étoile signifie que l'ensemble est privé de 0 ($\mathbb{N}^* = \{1, 2, 3, \ldots\}$).

Voici d'abord les résultats essentiels, qui seront suivis par quelques explications :

- $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$ ont autant d'éléments
- $\mathbb{R}$, $\mathbb{C}$ ont autant d'éléments
- Les 3 premiers ont moins d'éléments que les 2 derniers.

La notion de cardinal étend la notion de nombre aux infinités, de façon à ce que l'on puisse comparer les ensembles infinis. Voici comment.

4.8.2 Cardinal d'un ensemble fini.

Opérations sur les cardinaux (finis). Pour un ensemble fini, le cardinal est une notion intuitive, c'est simplement le nombre d'éléments de l'ensemble. Il appartient à $\mathbb{N}$.

Ex : card $(\{1, 2, 3\}) = 3 =$ card $(\{6, 15, 28\})$

Propriété 1: *Si A et B sont deux ensembles finis* ***disjoints*** *alors*

$$\text{card } (A \uplus B) = \text{card } A + \text{card } B$$

où $\uplus$ désigne l'union disjointe.

Ex : $A = \{1, 2, 3\}$, card $A = 3$ et $B = \{8, 9\}$, card $B = 2$. Alors $A \cup B = \{1, 2, 3, 8, 9\}$ et card $(A \cup B) = 5 = 3 + 2$.

Propriété 2: *Si A et B sont deux ensembles finis, alors* card $(A \times B) =$ (card A) $\times$ (card B) *où $A \times B$ désigne le produit cartésien des ensembles A et B.*

Ex : $A = \{1, 2, 3\}$, card $A = 3$ et $B = \{a, b\}$, card $B = 2$. Alors $A \times B = \{(1, a), (1, b), (2, a), (2, b), (3, a), (3, b)\}$. Donc card $(A \times B) = 6 = 3 \times 2$.

Propriété 3: *Si A est un ensemble fini on note $\mathcal{P}(A)$ l'ensemble des parties de A, c'est a dire l'ensemble des sous-ensembles de A.*

Alors card $(\mathcal{P}(A)) = 2^{\text{card } (A)}$

En effet : pour constituer une partie B de A, il y a un choix à faire pour chaque élément de A : soit on le met dans B, soit on ne l'y met pas (2 possibilités). S'il y a n éléments dans A, cela donne 2^n possibilités pour B, soit 2^n parties différentes.

Ex : $A = \{1, 2, 3\}$, card $A = 3$. Et l'on a

$$\mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1, 2, 3\}\}$$

Donc card $\mathcal{P}(A) = 8 = 2^3$.

Ce sont ces propriétés (1), (2), (3) qu'on va vouloir généraliser aux ensembles infinis.

Pour un ensemble infini, l'intuition ne s'applique plus : il faut recourir à des définitions formelles. C'est Cantor qui, le premier, a fourni un système cohérent permettant de traiter les ensembles infinis.

4.8.3 Définition rigoureuse d'un ensemble infini :

Une bijection f entre 2 ensembles E et F est une application qui fait correspondre à chaque élément de E un unique élément de F et inversement.

S'il existe une bijection entre E et F, on dit que E et F sont équipotents et on note : $E \sim F$.

Un ensemble est dit fini s'il est en bijection avec un ensemble $A_n = \{1, 2, 3, \ldots, n\}$ avec n un entier naturel (et $A_0 = \emptyset$).

Alors card $(E) = n$, et on retrouve la notion intuitive.

Un ensemble E qui ne peut pas être mis en bijection avec un tel ensemble A_n est dit infini.

Ex : $\mathbb{N}$ est un ensemble infini.

Cardinal d'un ensemble infini : On étend la notion de cardinal aux ensembles infinis de la façon suivante (schématiquement ; pour plus de rigueur voir partie II) :

Définition : *Si 2 ensembles sont équipotents (en bijection), on dit qu'ils ont le même cardinal.*

On dit que card $E \leq$ card F *si E a même cardinal qu'une partie de F. En particulier, si un ensemble infini E est inclus dans un ensemble (infini) F alors* card $E \leq$ card F *(inégalité large).*

Ex. (règle 1) : card $\mathbb{Z}$ = card $\mathbb{N}$ car il existe (au moins) une bijection de $\mathbb{N}$ dans $\mathbb{Z}$, qu'on peut expliciter :

$$g : \begin{cases} n \to n/2 & \text{si n est pair} \\ n \to -(n+1)/2 & \text{si n est impair} \end{cases}$$

ce qui donne

$$\begin{array}{ccc} 0 & \to & 0 \\ 1 & \to & -1 \\ 2 & \to & 1 \\ 3 & \to & -2 \\ & \vdots & \end{array}$$

Ex. (règle 2) :

$$\text{card } \mathbb{N} \leq \text{card } \mathbb{Z} \leq \text{card } \mathbb{Q} \leq \text{card } \mathbb{R} \leq \text{card } \mathbb{C}$$

4.8.4 Quelques résultats sur les ensembles dénombrables.

On a vu card $\mathbb{N}$ = card $\mathbb{Z}$, établissons quelques autres résultats :

Propriété : card $\mathbb{N}$ = card $(\mathbb{N}^2)$

On peut en effet compter les éléments de $\mathbb{N}^2 = \{(a,b) | a, b \in \mathbb{N}\}$. Une bijection possible de $\mathbb{N}^2$ dans $\mathbb{N}$ est le "comptage en diagonale" :

$$\begin{array}{ccc} (0,0) & \to & 0 \\ (1,0) & \to & 1 \\ (0,1) & \to & 2 \\ (2,0) & \to & 3 \\ (1,1) & \to & 4 \\ (0,2) & \to & 5 \\ & \vdots & \end{array}$$

(On compte les points de $\mathbb{N}^2$ selon les diagonales bas-droite $\rightarrow$ haut gauche, en s'éloignant de l'origine). On peut même expliciter la bijection : $(p,q) \rightarrow (p+q)(p+q+1)/2+q$

De même, on a card $(\mathbb{Z}^2) =$ card $\mathbb{Z}$.

Propriété : card $\mathbb{Z} =$ card $\mathbb{Q}$

Essentiellement, une fraction p/q est un couple d'entiers relatifs (p,q). Cette égalité n'est donc pas particulièrement surprenante quand on a admis card $\mathbb{Z}^2 =$ card $\mathbb{Z}$.

Formellement, la démonstration est un peu technique : elle est donc donnée dans la deuxième partie.

On a donc : card $\mathbb{N} =$ card $\mathbb{Z} =$ card $\mathbb{Q}$.

Ce cardinal est noté $\aleph_0$footnote$\aleph$ (*aleph*) est la première lettre de l'alphabet hébreu.

Tous les ensembles infinis qui ont le même cardinal que $\mathbb{N}$ sont dits dénombrables.

Au sens large, *dénombrable* signifie aussi « fini ou dénombrable ». Enfin, un ensemble qui n'est ni fini ni dénombrable est dit... indénombrable.

4.8.5 Quelques résultats sur les ensembles indénombrables.

Plaçons nous maintenant dans les réels.

Propriété : card $[0,1] =$ card $[0,2] =$ card $[a,b] = \ldots$

Il est facile de trouver une bijection de $[0,1]$ dans $[0,2]$: $k : x \rightarrow 2x$ par exemple. De même, on peut toujours trouver une bijection entre 2 intervalles de $\mathbb{R}$, ouverts ou fermés. Tous les intervalles de $\mathbb{R}$ ont donc le même cardinal.

Propriété : card $(a,b) =$ card $\mathbb{R}$

(a,b) *désigne l'intervalle, sans se préoccuper de l'inclusion ou non des bornes.*

La fonction tangente, par exemple, fournit une bijection de $]-\pi/2, \pi/2[$ dans $\mathbb{R}$. En appliquant ce qui précède, tout intervalle de $\mathbb{R}$ a le même cardinal que $\mathbb{R}$.

Propriété : card $[0,1[^2 =$ card $[0,1[$

En effet on peut expliciter une bijection de $[0,1[^2$ dans $[0,1[$: Soit (a,b) dans $[0,1[^2$. écrivons leur développement décimal : $a = 0,a_1a_2a_3a_4\ldots$ $b = 0,b_1b_2b_3b_4\ldots$ formons alors $c = 0,a_1b_1a_2b_2a_3b_3a_4b_4\ldots$

N'est-il pas alors clair que la fonction $l : (a, b) \to c$ est une bijection de $[0, 1[^2 \to [0, 1[$? En fait, pas tout a fait : il y a une lacune, expliquée dans la seconde partie, mais ça ne change pas le résultat

Grâce à ce qui précède on déduit que :

$$\text{card } \mathbb{R} = \text{card } [0, 1[= \text{card } ([0, 1]^2) = \text{card } \mathbb{R}^2$$

On peut démontrer la même chose avec n'importe quel exposant à la place de 2.

Propriété : card $\mathbb{C}$ = card $\mathbb{R}$

En effet, les fonctions partie réelle et partie imaginaire d'un complexe fournissent une bijection évidente de $\mathbb{C}$ dans $\mathbb{R}^2$.

Avec ce qui précède on a donc card $\mathbb{C}$ = card $\mathbb{R}^2$ = card $\mathbb{R}$.

4.8.6 Rapport dénombrable/indénombrable.

Propriété : *On va maintenant voir que* $\mathbb{R}$ *n'est pas dénombrable.*

Pour simplifier, on va montrer que $[0, 1[$ n'est pas dénombrable ; ce qui implique, bien sûr, que $\mathbb{R}$ ne l'est pas. Supposons le contraire : on peut alors énumérer $[0, 1[$, c'est-à-dire qu'il existe une suite $(u_n)_{n\in\mathbb{N}}$, qui contient tous les réels de $[0, 1[$.

écrivons le développement décimal (par exemple) des premiers termes de (u_n) en notant $u_{n,i}$ le i-ème chiffre après la virgule de u_n :

$$\begin{aligned} u_1 &= 0, u_{1,1}u_{1,2}u_{1,3}u_{1,4}\dots \\ u_2 &= 0, u_{2,1}u_{2,2}u_{2,3}u_{2,4}\dots \\ &\vdots \\ u_n &= 0, u_{n,1}u_{n,2}u_{n,3}u_{n,4}\dots \\ &\vdots \end{aligned}$$

On forme maintenant le nombre v s'écrivant :

$$v = 0, u_{1,1}u_{2,2}u_{3,3}u_{4,4}\dots u_{i,i}\dots$$

(On prend les chiffres de la diagonale principale ci-dessus) On forme enfin le nombre w en remplaçant chaque chiffre de v par son prédécesseur (et 0 par 8) :

$$w = 0, w_1w_2w_3w_4\dots$$

avec : $w_i = 8$ si $u_{i,i} = 0$ $w_i = u_{i,i} - 1$ sinon.

Alors le nombre w, qui est bien défini, n'est pas dans la liste des valeurs parcourues par (u_n). En effet : w_1 (le premier chiffre de w après la virgule) est différent de $u_{1,1}$ (le premier chiffre de u_1 après la virgule) donc $w \neq u_1$. $w_2 \neq u_{2,2}$ donc $w \neq u_2 \ldots w_n \neq u_{n,n}$ donc $w \neq u_n$ pour tout n.

On a donc construit un réel de $[0, 1[$ qui n'est pas dans (u_n) : contradiction avec l'hypothèse. Une telle suite u_n parcourant tous les réels de $[0, 1[$ n'existe donc pas, donc $[0, 1[$ est indénombrable. La technique qui nous a permis de le montrer est classique, et connue sous le nom de *Procédé diagonal de Cantor*.

Conclusion : Tout ceci nous permet d'écrire :

$$\text{card } \mathbb{N} = \text{card } \mathbb{Z} = \text{card } \mathbb{Q} < \text{card } \mathbb{R} = \text{card } \mathbb{C}$$

Pour d'autres développements, et des justifications plus rigoureuses, voir 4.9.

4.9 Les cardinaux des ensembles infinis - II.

4.9.1 Autre définition rigoureuse d'un ensemble infini

Une définition équivalente (moyennant l'axiome du choix) à celle de la première partie est : Un ensemble E est dit infini si on peut trouver une bijection entre lui-même et une de ses parties (strictes) F, ie : $F \subset E$, $F \neq E$, $F \sim E$.

Ex : $\mathbb{N}$ est en bijection avec $\mathbb{N}^*$, par la fonction $f : n \to n+1$ ainsi, $\mathbb{N}$ est infini

cardinal d'un ensemble infini : La façon parfaitement rigoureuse de définir le cardinal pour un ensemble infini s'appuie sur le théorème de Cantor-Bernstein :

Théorème.

Soient A et B deux ensembles (finis ou infinis). Si A s'injecte dans B et B s'injecte dans A, alors A et B sont équipotents.

(la démonstration est facile dans le cas fini, un peu moins dans le cas général).

Ce théorème justifie les écritures : card (A) = card (B) si A et B sont équipotents card $(A) \leq$ card (B) si A s'injecte dans B En effet, on peut alors appliquer la règle d'antisymétrie : si $n \leq m$ et $m \leq n$ alors $n = m$.

Écrire cette règle avec les cardinaux infinis est une simple ré-écriture du théorème de Cantor-Bernstein.

D'autre part, on a l'équivalence (moyennant l'axiome du choix) : A s'injecte dans $B \iff B$ se surjecte sur A et en notant card $(B) >=$ card (A) si B se surjecte sur A, on reste cohérent en manipulant les cardinaux.

Pour la culture (merci à David Madore) : Si A et B sont deux ensembles quelconques, alors soit A s'injecte dans B (ie card $(A) \leq$ card (B)), soit B s'injecte dans A (ie card $(B) \leq$ card (A)). En d'autres termes : les cardinaux forment une classe *totalement ordonnée*.

4.9.2 Quelques précisions sur les ensembles dénombrables.

Propriété : card $(\mathbb{Z}^2) =$ card $\mathbb{Q}$

On a une injection évidente $f : Q \to Z^2$ telle que $f(p/q) = (p, q)$ avec p,q premiers entre eux et $q > 0$ on peut rajouter $f(0) = (0, 1)$ pour être parfaitement rigoureux. Évidemment, ce n'est pas une bijection : $(4, 6)$ par exemple n'a pas d'antécédent, puisque $f(4/6) = (2, 3)$.

On a donc card $\mathbb{Q} \leq$ card $(\mathbb{Z}^2)$.

D'autre part, on a vu dans la première partie : card $(\mathbb{Z}^2) =$ card $\mathbb{Z}$. Or $\mathbb{Z}$ s'injecte trivialement dans $\mathbb{Q}$: card $\mathbb{Z} \leq$ card $\mathbb{Q}$.

De ces deux inégalités on déduit card $(\mathbb{Z}^2) =$ card $\mathbb{Q}$, alors qu'il serait fastidieux d'exhiber une bijection explicite entre $\mathbb{Q}$ et $\mathbb{Z}^2$.

4.9.3 Quelques précisions sur les ensembles indénombrables.

Propriété : card $([0, 1[^2) =$ card $[0, 1[$

En fait, l'application l donnée dans la partie I n'est pas vraiment une bijection, mais seulement une injection.

Cela est dû à l'existence de deux écritures décimales distinctes pour les décimaux : l'une avec un nombre fini de chiffre (ex : $1, 23$), l'autre avec un nombre infini de chiffres (ex : $1, 2299\ldots$[1])

En conséquence, le nombre $0, 50595\ldots$ par exemple n'a pas d'antécédent puisque :

$$l(0, 555\ldots, 0, 09999\ldots) = l(0, 555\ldots, 0, 1) = 0, 515050\ldots$$

Cependant, ça suffit : l nous fournit une injection de $[0, 1[^2$ dans $[0, 1[$, et l'injection réciproque est triviale. En appliquant le théorème de Cantor-Bernstein on peut donc conclure en toute rigueur que

$$\text{card } ([0, 1[) = \text{card } ([0, 1[^2)$$

[1] Voir la FAQ $0, 99\ldots = 1$.

La généralisation à card $\mathbb{R}$ = card $(\mathbb{R}^2)$, se fait toujours comme indiqué dans la première partie.

4.9.4 Rapport dénombrable/indénombrable.

Voici une démonstration plus indirecte du fait que $\mathbb{R}$ est indénombrable, mais qui a l'avantage de préciser le rapport entre les 2, à savoir $\mathbb{R} \sim \mathcal{P}(\mathbb{N})$, l'ensemble des parties de $\mathbb{N}$.

Propriété : *Soit E un ensemble. $\mathcal{P}(E)$ n'est jamais équipotent à E.*

En voici une (jolie) démonstration par l'absurde.

Supposons qu'il existe une bijection $h : E \to \mathcal{P}(E)$ Pour tout x dans E, $h(x)$ est un sous-ensemble de E. Notons $F = \{x \in E | x$ n'est pas dans h(x)$\}$. Par construction, F est une partie de E, donc F est dans $\mathcal{P}(E)$. Comme h est une bijection, F a un antécédent dans E, ie : il existe $y \in E$ tel que $h(y) = F = \{x \in E | x$ n'est pas dans $h(x)\}$.

Alors :

– si $y \in F$, par définition de F, $y \notin h(y) = F$. Absurde.

– si $y \notin F$, par définition de F, $y \in h(y) = F$. Absurde aussi.

C'est donc l'hypothèse d'existence de la bijection h qui est fausse. Cqfd.

Propriété : card $\mathcal{P}(N)$ = card $\mathbb{R}$

On constitue donc l'application $p : \mathcal{P}(\mathbb{N}) \to \mathbb{R}$ de la façon suivante. Soit E dans $\mathcal{P}(\mathbb{N})$. E est donc un sous-ensemble (fini ou infini) de $\mathbb{N}$.

On forme un nombre d écrit en base 2 de la façon suivante :

$$d = 0, d_0 d_1 d_2 d_3 d_4 d_5 \dots$$

où

$$d_n = \begin{cases} 1 & \text{si } n \in E \\ 0 & \text{sinon} \end{cases}$$

On a alors $d \in [0, 1]$ ($d = 0$ si $E = \emptyset$, $d = 1$ si $E = \mathbb{N}$).[2]

D'autre part, si on fabrique l'application q comme p, mais en considérant que cette fois d est écrit en base 3 (ou 10, ou n¿2), alors l'application q est une injection de $\mathcal{P}(\mathbb{N})$ dans $\mathbb{R}$.[3]

[2] On peut définir p de façon plus concise : $p(E) = \sum_{n \in E} 2^{-(n+1)}$ La fonction $p : E \to p(E) = d$ est alors une surjection de $\mathcal{P}(\mathbb{N})$ dans $\mathbb{R}$. Tout réel $r \in [0, 1]$ est atteint, et pour connaître son antécédent, il suffit d'écrire le développement binaire de r. En revanche, p n'est pas une bijection, en raison de l'existence de 2 développements binaires, comme pour les développements décimaux.

[3] Tous les réels ne sont pas atteints, mais chaque réel atteint a un antécédent unique.

$\mathcal{P}(\mathbb{N})$ se surjecte et s'injecte à la fois dans $\mathbb{R}$, donc $\mathcal{P}(\mathbb{N}) \equiv \mathbb{R}$, ie : card $\mathcal{P}(\mathbb{N}) =$ card $\mathbb{R}$

Par généralisation du cas fini, on écrit :

$$\text{card } \mathbb{R} = \text{card } \mathcal{P}(\mathbb{N}) = 2^{\text{card } \mathbb{N}} = 2^{\aleph_0}$$

Propriété : card $\mathbb{N} <$ card $\mathbb{R}$

En effet, puisque $\mathbb{N}$ est inclus (donc s'injecte) dans $\mathbb{R}$,card $\mathbb{N} \leq$ card $\mathbb{R}$. De plus, on vient de montrer que card $\mathbb{R} =$ card $\mathcal{P}(\mathbb{N})$, et que $\mathcal{P}(\mathbb{N})$ n'est pas équipotent à $\mathbb{N}$. Donc card $\mathbb{N} <$ card $\mathbb{R}$.

4.9.5 Hypothèse du Continu...

On définit $\aleph_1$ comme étant le plus petit cardinal strictement supérieur à $\aleph_0$[4]

Le fait de savoir si card $\mathbb{R} = \aleph_1$ est indécidable d'après la construction de $\mathbb{R}$ seule, c'est-à-dire qu'on peut arbitrairement décider que oui ou non. Habituellement, on fait l'hypothèse que c'est bien le cas (hypothèse du Continu) : $\aleph_1 =$ card $\mathbb{R}$.

En termes plus simples, l'hypothèse du Continu dit que : toute partie de $\mathbb{R}$ est soit au plus dénombrable, soit équipotente à $\mathbb{R}$.

En termes intuitifs (!) : il n'existe pas d'ensemble "strictement plus gros" que $\mathbb{N}$ mais "strictement plus petit" que $\mathbb{R}$.

Opérations sur les cardinaux. Toutes les opérations valables avec les cardinaux finis, sont extensibles aux cardinaux infinis, mais l'intérêt est moindre : Si A ou B est infini :

$$\begin{aligned} \text{card } (A \cup B) &= \text{card } A + \text{card } B = \max(\text{card } A, \text{card } B) \\ \text{card } (A \times B) &= \text{card } A \times \text{card } B = \max(\text{card } A, \text{card } B) \\ \text{card } (\mathcal{P}(A)) &= 2^{\text{card } (A)} \end{aligned}$$

L'hypothèse du Continu "dit" ainsi que $\aleph_1 = 2^{\aleph_0}$.

4.9.6 Références

Quelques lectures complémentaires ; [Dun], [?], [?].

[4] pour une définition plus rigoureuse de $\aleph_1$, voir l'article de David Madore : http://www.eleves.ens.fr:8080/home/madore/w/ordinals/ordinals.html

Chapitre 5

Les constantes mathématiques

5.1 π.

Algorithmes de calculs de décimales. C'est Archimede (en 225 avant JC) qui calcula le premier algorithme sur les décimales de Pi. Petit texte de 53 pages avec bibliographie sur les formules et algorithmes pour calculer Pi, disponible au format Postscript à : http://www.multimania.com/~gersoo/docs/pi/pi.ps.gz

L'algorithme de Bailey, Borwein et Plouffe ? C'est pour calculer π en base 16 (il y a une version en base 10, mais beaucoup plus lente). L'intérêt, c'est qu'on peut calculer un chiffre sans devoir déterminer tous les précédents. Voir par exemple les pages web suivantes :

- L'Univers de Pi
- The Pi Pages
- The joy of Pi offre le premier million de décimales ;

J.P. Delahaye a écrit un très beau livre sur le sujet : *Le Fascinant nombre* π[**?**].

5.2 La constante d'Euler e.

5.2.1 Le point de vue Néperien.

On pose $\ln(e) = 1$. Il suffit d'expliquer comment on définit naturellement la fonction logarithme, et tu comprendras pourquoi e est e.

La fonction ln se définit (à une constante près) par la propriété :

$$\ln(xy) = \ln(x) + \ln(y).$$

En dérivant cette relation, on vérifie que : $y \ln'(xy)) = \ln'(x)$. Donc : $\ln'(y) = \ln'(1)/y$. On définit donc la fonction ln comme étant *la seule fonction* dérivable telle que :

- $\ln(xy) = \ln(x) + \ln(y)$
- $\ln'(1) = 1$.

Puis on définit e par : $\ln(e) = 1$.

L'intérêt de cette définition de la fonction ln : on a de jolis développements en série entière, la fonction exp, inverse de la fonction ln vérifie : $y' = y$ et $y(0) = 1$ (c'est un théorème).

Tiens, d'ailleurs, on peut aussi décider de définir d'abord la fonction exp. Les théorèmes deviennent des définitions et les définitions deviennent des théorèmes :

5.2.2 Le point de vue de Cauchy.

exp est la solution de l'équation différentielle $y' = y$, vérifiant : $\exp(0) = 1$. Suivant ce point de vue, on définit $e = \exp(1)$ et on définit ln comme étant la fonction inverse de la fonction exp (de sorte que $\ln(e) = 1$).

On a alors le théorème :

- $\ln(xy) = \ln(x) + \ln(y)$
- $\ln'(1) = 1$.

5.2.3 Le point de vue Théologique.

$e^{i\pi} + 1 = 0$. Il suffit de définir la fonction puissance.

Si a est un réel positif, on définit a^n pour tout entier naturel n, puis $a^{1/p}$ pour tout entier naturel p (c'est le nombre b tel que $b^p = a$. Ce nombre existe car l'application $b \to b^p$ est un bijection)

Donc a^q pour tout nombre rationnel positif q, puis a^x pour tout réel positif x (par passage à la limite).

On s'aperçoit que la fonction que l'on définit ainsi admet un prolongement analytique : $z \in \mathbb{C} \to a^z \in \mathbb{C}$, on démontre qu'il existe une infinité de réels positifs tel que $e^{i\pi} + 1 = 0$

et on définit e comme le plus petit réel positif tel que :

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

Ensuite, on peut définir la fonction exponentielle par $\exp(x) = e^x$, et les définitions du second point de vue deviennent des théorèmes (alors qu'en utilisant la définition du second point de vue (convenablement étendue aux nombres complexes), l'équation $e^{i\pi} + 1 = 0$ est un théorème).

5.2.4 Irrationalité de e.

L'irrationalité de e fut prouvée dès 1737 par Euler, toujours lui ! (voir la remarque 2, ci-après), de la façon suivante.

$$e = \sum_{k=0}^{\infty} \frac{1}{k!}$$

Donc, une première évidence : la somme partielle (appelons-la S_n), pour k variant de 0 à n de cette série est strictement inférieure à e.

Ensuite, on majore la "queue" de la série (k variant de $n+1$ à l'infini), en y remplaçant chaque $\frac{1}{k!} = \frac{1}{n!}\frac{1}{(n+1)(n+2)(n+3)\cdots k}$ par $\frac{1}{n!}\frac{1}{(n+1)^k}$. Cela donne une "bête" série géométrique dont la somme vaut finalement $\frac{1}{n!n}$.

Résumons : $S_n < e < S_n + \frac{1}{n!n}$ (N.B. ceux qui ont quelque idée de l'approximation rationnelle savent déjà que c'est gagné : voir la remarque 1 ci-après).

On peut écrire cela comme $e = S_n + \frac{r(n)}{n!n}$, avec $r(n)$ dans $]0,1[$.

Supposons maintenant que e soit rationnel, et soit alors a/b son écriture canonique.

Dans ce qui précède, en choisissant le cas particulier $n = b$, on obtient donc : $\frac{a}{b} = S_b + \frac{r(b)}{b!b}$, avec toujours $r(b)$ dans $]0,1[$.

Multiplions par $b!$. On trouve

$$(b-1)!a = \sum_{k=0}^{b} \frac{b!}{k!} + \frac{r(b)}{b}$$

C'est absurde, parce que le premier membre est entier, le premier terme du second membre l'est aussi (somme de termes tous évidemment entiers), tandis que le "terme d'erreur" $r(b)/b$ ne l'est pas.

Remarque 1. On sait (c'est d'ailleurs quasi évident) qu'un rationnel α n'est jamais approchable par une suite (illimitée) de rationnels s/t avec une "vitesse" v supérieure à 1. (je veux dire par là : $|\alpha - s/t| < constante/t^v$, avec $v > 1$).

L'encadrement obtenu par Euler était donc bien entendu trop minuscule pour "être honnête", c'est à dire pour cerner un rationnel.

5.2.5 Généralisation.

Liouville a montré en 1844 qu'un nombre algébrique d'ordre d n'est pas approchable à une vitesse strictement supérieure à d.

Il s'est servi de ce résultat (de démonstration élémentaire) pour construire effectivement une infinité (non dénombrable) de nombres transcendants.

Exemple : $\sum_{k=0}^{\infty} \frac{1}{10^{k!}}$

Roth a mis un point final à cette histoire en prouvant qu'aucun nombre algébrique de degré > 1 (c-à-d irrationnel) n'est approchable à un ordre strictement supérieur à 2.

Comme par ailleurs les réduites de la fraction continue pour ce nombre (leur suite est illimitée, puisqu'il s'agit d'un irrationnel) l'approchent précisément à l'ordre 2, ce théorème est optimal. Il a valu à Roth l'une des médailles Fields de 1955.

Remarque 2. Une réclame de la Mathematical Association of America annonce la parution du livre *Euler, the Master of us all.* (ce titre est une citation de Laplace). Le livre est écrit par William Dunham, dans la série *Dolciani Mathematical Expositions*, bien connue de tous les familiers de la MAA.

5.2.6 Transcendance de e

La transcendance de e fut prouvée par Charles Hermite en 1873 (Lindemann devait suivre avec π en 1882, seulement). C'est beaucoup plus difficile et horrible à écrire ici. Je résume donc brutalement.

Soit $f(t)$ un polynôme (quelconque pour l'instant) et $F(t)$ la somme de toutes ses dérivées successives.

En intégrant par parties, on prouve d'abord (c'est très facile) quei

$$\exp(x) \int_0^x \exp(-t) f(t) = -F(x) + \exp(x) F(0)$$

On suppose ensuite que e satisfait à l'équation algébrique à coefficients entiers :

$$\sum_{k=0}^{n} a_k \exp(k) = 0$$

(Ce n'est évidemment pas une restriction que de supposer a_0 non nul).

L'identité générale précédente donne alors :

$$\sum_{k=0}^{n} a_k \exp(k) \int_0^n a_k F(k) = -\sum_{k=0}^{n} a_k F(k) \tag{1}$$

Ici, coup de génie de Hermite : il choisit maintenant le polynôme

$$f(t) = \frac{t^{p-1}}{(p-1)!} \prod_{j=1}^{n} (j-t)^p$$

Où p est un nombre premier supérieur à n et à $|a_0|$. (C'est possible, puisqu'il existe une infinité de nombres premiers)

Il démontre ensuite que le second membre de (1) est un entier non multiple de p (c'est élémentaire, mais subtil), donc non nul, donc de valeur absolue valant au moins 1 (astuce classique en arithmétique !). Par ailleurs, le premier membre de (1) $\to 0$ lorsque $p \to \infty$, selon la suite des nombres premiers (par des majorations fort brutales). C'est la contradiction cherchée.

Chapitre 6

Problèmes de Géométrie.

6.1 Problème de la chèvre.

Problème Une chèvre est attachée par une corde de longueur l à un pieu fixé à un point A de la circonférence d'un enclos circulaire C de centre O et de rayon R. Trouver l en fonction de R pour qu'elle puisse brouter au maximum la moitié de l'herbe de l'enclos.

Une solution : On trace le cercle C' de centre A et de rayon l qui coupe le cercle C en H et K. J est le point diamétralement opposé à A sur le cercle C, et P est la projection orthogonale de O sur (AH).

Pour obtenir l'aire S commune aux deux cercles, on additionne l'aire S_1 du secteur AHK du cercle C', l'aire S_2 du secteur OHK du cercle C, et on soustrait l'aire S_3 du quadrilatère $AHOK$ (qui serait comptée deux fois sinon).

On prend comme inconnue l'angle $\widehat{OAH} = x$ en radians. On a : $AK = AJ\cos(x)$ (dans le triangle rectangle AKJ) d'où $l = 2R\cos(x)$. Donc

- $S_1 = \frac{1}{2}l^2(2x) = xl^2 = 4R^2x(\cos(x))^2$. Le triangle OAH est isocèle de sommet O, donc $AOH = \pi - 2x$ et
- $S_2 = \frac{1}{2}R^2 2(\pi - 2x) = R^2(\pi - 2x)$.
- S_3 est deux fois l'aire du triangle OAH :

$$S_3 = 2OP \times PA = 2R\sin(x)R\cos(x) = 2R^2\cos(x)\sin(x)$$

L'équation à résoudre est $S = \frac{\pi R^2}{2}$ ou encore :

$$\begin{array}{rcl} S_1 + S_2 - S_3 & = & \frac{\pi R^2}{2} \\ 4R^2x\cos^2(x) + R^2(\pi - 2x) - 2R^2\cos(x)\sin(x) & = & \frac{\pi R^2}{2} \end{array}$$

Ce qui se simplifie en : $2\sin(x)\cos(x) - 2x(2(\cos(x))^2 - 1) = \frac{\pi}{2}$ En posant $y = 2x$: $\sin(y) - y\cos(y) = \frac{\pi}{2}$

y est l'angle $\widehat{HAK}$ et est compris entre 0 et π. La fonction $f : y \to \sin(y) - y\cos(y)$ est continue et strictement croissante sur $[0, \pi]$ et $f(0) = 0$ et $f(\pi) = \pi$. L'équation $f(y) = \frac{\pi}{2}$ admet donc une seule solution dans $[0, \pi]$.

Avec un outil de calcul, on trouve : $y = 1.905695729\ldots$ On en déduit ensuite : $\frac{l}{R} = 2\cos\frac{y}{2} = 1.158728473\ldots$

6.2 Problème (dit) de Napoléon.

Problème : On dit que Napoléon aurait trouvé le moyen de déterminer le centre d'un cercle en utilisant uniquement le compas. Quelle est la construction ?

Solution : Cette construction, en six étapes, est celle de Napoléon et Mascheroni.

1. D'un point A du cercle , un arc de cercle de rayon quelconque qui recoupe le cercle en B et C
2. arc de cercle de centre B et de rayon AB
3. arc de cercle de centre C et de rayon AB. Ces deux arcs se coupent en D.
4. arc de cercle de centre D et de rayon DA qui recoupe le premier arc (de l'étape 1.) en E et F.
5. arc de cercle de centre E et de rayon EA
6. arc de cercle de centre F et rayon EA. ces deux arcs se coupent en O

Il ne reste plus qu'à prouver que O est bien le centre du cercle... Ce qui n'est pas très difficile. Ce classique vient du *dictionnaire des Mathématiques*[**?**].

Remarque. Sans nier que Napoléon ait quelques qualités de mathématicien, on pense qu'il avait surtout des amis ou courtisans chez les savants de l'époque et que l'auteur de cette belle solution est Lorenzo Mascheroni, auteur d'une "Géométrie du compas" célèbre en son temps. C'est Monge qui a transmis le « truc »à Napoléon.

Chapitre 7

Mathématiques et Ordinateurs.

7.1 Écrire des mathématiques sur Usenet.

Le but de ce texte est de proposer, aux nouveaux venus ou à ceux qui n'ont pas d'idée, des méthodes pour écrire des textes mathématiques dans le forum fr.sci.maths.

Ce texte vient en complément du texte *Conseil d'Utilisation de fr.sci.maths*, régulièrement publié dans fr.sci.maths et que tout nouvel arrivant sur ce forum doit lire.

Tous les commentaires, demandes d'ajout ou de modification, doivent être envoyés à Frédéric Bastok.

Les discussions dans fr.sci.maths nécessitent parfois l'emploi de formules plus ou moins compliquées. Dans le but de les rendre lisibles par le plus grand nombre, une réflexion menée dans ce groupe a conduit aux résultats suivants. Les notations possibles sont de 4 sortes :

la notation texte : elle consiste à écrire en français un équivalent de la phrase mathématique.

Exemple : écrire `intégrale de a à b de la fonction f(x)` au lieu d'une méthode ci-dessous.

Recommandations : cette méthode est utilisable dans les cas très simples (notamment les formules en un seul bloc comme celle de l'exemple) et ne peut être appliquée aux formules compliquées.

la notation graphique : on utilise les caractères du clavier pour donner l'illustration du symbole voulu. En reprenant le même exemple :

```
 b
/
```

```
| f(x) dx
/
a
```

Recommandations : les formules sont assez longues à écrire ; ce mode a les mêmes inconvénients que le précédent (limitation à des formules simples). De plus, on n'est jamais certain du résultat obtenu (différence d'affichage selon les lecteurs, taille des lignes, police) : il est impératif de ne pas utiliser de tabulations.

la notation des systèmes de calcul : sous ce terme sont regroupées les notations des calculatrices graphiques (hors calculatrices RPN, notamment les modèles HP) et des logiciels de calcul.

Exemple : int (f(x), x=a..b) ou int (f(x), x, a, b) pour l'intégrale.

Recommandations : cela peut être utilisé dans deux cas notamment :

- l'auteur ne connaît que le langage de sa calculatrice ;
- il souhaite poser une question sur un logiciel particulier donc il utilise la syntaxe précise de ce logiciel.

Il n'est pas interdit d'utiliser les mots français équivalents (somme pour sum par exemple) mais on évitera d'être excessif (il est tout de même plus rapide d'écrire evalf que évaluation de la fonction). Il s'agit du meilleur choix pour ceux qui ne connaissent pas TEXou qui ne souhaitent pas l'utiliser.

TEXou LATEX : il s'agit du langage bien connu, qui est un standard dans le monde scientifique.

Recommandations : c'est le meilleur choix pour les textes de niveau élevé (au-dessus de la licence). Par contre, il est recommandé de ne pas l'utiliser pour les messages pouvant être lus par le plus grand nombre pour ne frustrer personne.

L'attention est attirée sur la différence entre le pseudo-TEX(pas de directive de compilation) et TEX : ceux qui souhaitent rendre leur texte utilisable sur un interpréteur TEXdevront l'écrire en respectant parfaitement la syntaxe de TEX. On limitera l'utilisation de TEXqui se rapproche de l'HTML (non lisible directement) et qui est donc mal vu sur Usenet.

Je rappelle ici certaines règles propres à Usenet et aux newsreaders. Les articles binaires sont interdits dans la hiérarchie fr.*. Ne pas poster de messages avec des fichiers joints contenant les formules (par exemple : fichier obtenu à l'aide de l'éditeur d'équation de Word)

Un sous-ensemble très précis de MIME doit être utilisé : le document doit être de type text / plain (les autres types, par exemple text/html, application/*, image/*, multipart/* sont interdits) ; le jeu de caractères utilisé doit être ISO 8859-1 (ou ASCII, qui en est un sous ensemble) ou ISO 8859-15 (extension pour l'euro) ; aucun encodage ne doit être utilisé (ni base64, ni quoted-printable) ;

Exemples non exhaustifs :

- infini : `inf`
- intégration : `int (f(x), x, a, b)`
- dérivation : `df/dx`
- puissance : `5^3 = 125`
- suite : `u_n` signifie le n-ième terme de la suite u `u_(n+1)` signifie le n+1 -ième terme de la suite u alors que `u\_n + 1` signifie le n-ième terme de la suite u plus le chiffre 1
- série : `sum (ln (x)/x\^ n, n = 1..+inf)`
- `*` : multiplication

Règle générale : il faut préciser les notations lorsqu'elles peuvent être source de confusion. Ex : utilisation de * pour un produit de convolution.

Recommandations de présentation.

- éviter de mélanger les différents types de notation.
- éviter de noter la multiplication par simple juxtaposition, utiliser l'opérateur * explicite.
- utiliser les espaces à l'intérieur de la formule pour la rendre plus lisible, par exemple : `p(x) = a*x \^ 2 + b*x + c`

Erreurs fréquemment commises. Ce paragraphe vient en complément du paragraphe équivalent du Conseil d'Utilisation de `fr.sci.maths` pour préciser les erreurs couramment rencontrées dans l'écriture des formules mathématiques.

Il faut faire très attention au parenthésage car c'est la source d'incompréhension la plus courante.

Exemple : `z' = 2/3(z+i)` doit être écrit :
`z' = 2/(3(z+i))` ou `z' = (2/3)*(z+i)` selon les cas.

Premiers pas avec LaTeX. Ce document a été rédigé avec LaTeX(puis exporté en PDF et en HTML). Le Laboratoire lorrain de recherche en informatique et ses applications est une excellente ressource Latex : vous trouverez de nombreux liens vers de la documentation.

Par ailleurs, je vous recommande le guide Apprends Latex de Damien Wyart.

7.2 Logiciels de Mathématiques.

Il ne s'agit pas de parler ici des éditeurs d'équations ; ceux-ci sont abordés dans la section 7.1. Nous aborderons ici les logiciels capables de faire du calcul symbolique, du calcul numérique, des graphiques en 2 ou 3 dimensions et de la programmation.

Tout d'abord, il y a trois gros logiciels commerciaux : Mathematica, Maple et Matlab. Ensuite, il existe différents programmes gratuits ou libres (MuPad, Scilab,etc.) Il existe d'autres logiciels non présentés ici (voir [**?**]).

7.2.1 Maple

MapleSoft [1]

Maple est un excellent logiciel de calcul formel qui est une aide remarquable pour les calculs et exercices. Il est également possible de voir la façon dont sont programmés la plupart des fonctions de Maple.

Maple est assi performant pour tracer des courbes en 2D ou 3D que pour résoudre des équations que pour programmer dans un langage mathématique.

Il est bien entendu possible d'écrire de petits programmes.

Maple a également un fichier d'aide très bien fait.

Chose étonnante, Maplesoft a décidé de changer la syntaxe de ses procedures dans la version VI, mais d'integrer de plus en plus de fonctionalités d'échanges avec les logiciels microsofts. :o(

En conséquence, on préférera les versions IV ou V de Maple, plus répandues et moins chères que la dernière version.

Vous trouverez des cours sur Maple à l'adresse suivante : `http://www.giromini.org/mi101/`

7.2.2 Mathematica

Wolfram Research[2]. Version étudiante (complète avec documentation sur CD-ROM au lieu du manuel) disponible.

[1] Disponible sur les plateformes Windows, Macintosh, Linux (min.486), DEC Alpha, HP/9000, IBM RS/6000, SGI, Sun SPARC.

[2] Disponible sur les plateformes Windows, Macintosh, Linux, la plupart des Unix (Sun, SGI, IBM RS/6000, HP PA-Risc, DEC AXP) et NEXTSTEP (Motorola, x86, HP PA-RISC, SPARC), OS/2 et DEC OpenVMS

Très puissant pour le calcul symbolique et le graphisme. La plupart des fonctions ont des paramètres ajustables (précision, options spéciales, etc.). Exportation des feuilles de calcul en plusieurs formats (PostScript, LaTeX, HTML, etc.). Plusieurs librairies de fonctions supplémentaires spécialisées peuvent s'ajouter au module principal (déjà très complet).

7.2.3 Curvus Pro

Curvus Pro est un traceur d'objets mathematiques en 2 et 3D fonctionnant sous MacOS, realise par deux étudiants Suisses.

Il permet notamment de tracer des courbes issues d'équations paramétriques en 2D à partir de coordonnées cartesiennes, polaires et complexes, des équations differentielles en differents systèmes de coordonnées, des champs vectoriels, des champs scalaires, des équations implicites.

En 3D, il trace des surfaces et des courbes à partir de 4 systemes de coordonnées, ainsi que des champs. Ses fonctions de traçage, dont certaines utilisent les fonctionnalites de QuickDraw 3D n'ont rien à envier aux graphismes de Maple.

Des calculs divers peuvent etre faits sur les courbes (intégrales, derivées,...) et il possède un éditeur d'équations très perfectionné.

7.2.4 Matlab

Mathworks, `http://www.matlab.com`[3]

A la différence des précédents logiciels, Matlab ne gère absolument aucun calcul symbolique. En effet, "matlab" et l'abréviation de "matrix laboratory" et c'est effectivement dans ce domaine que Matlab excelle : en fait, tous les objets manipulés par Matlab sont considérés comme des matrices.

Puissant pour les graphiques et la visualisation. Langage de programmation simple qui permet de faire des scripts (M-files) facilement.

Matlab permet également de construire très facilement des interfaces graphiques (par exemple pour saisir les paramètres d'un calcul, le lancer et visualiser le résultat).

On trouvera très facilement des librairies supplémentaires et spécialisées (traitement du signal, résaux de neurones, etc.)

[3]Disponible sur les plateformes Windows, Macintosh, DEC-Alpha, HP 9000 PA-RISC, IBM RS/6000, OpenVMS, SGI, SUN, Linux (min. 486).

7.2.5 MuPad

http://www.mupad.de[4]

MuPad est un système de calcul formel commercial, distribué par SciFace, mais des licences gratuites sont accordées pour la plupart des versions dans le cas d'utilisations non commerciales. Ce moteur a la particularité de permettre la redéfinition complète des types du langage et l'ajout d'autres types. Le langage étant orienté objet, il est possible de profiter du polymorphisme pour définir des "domaines" (des structures au sens algébrique).

7.2.6 Logiciels gratuits

(Cette liste reste cependant à compléter.)

Scilab ftp://ftp.inria.fr/INRIA/Projects/Meta2/Scilab[5].
Mail : scilab@inria.fr ; Newsgroup : comp.soft-sys.math.scilab

PARI http://halse.mathematik.tu-muenchen.de/ntsw/pari/[6].

PARI est une bibliothèque de fonctions en C spécialisée en algèbre et théorie des nombres (théorie algébrique des nombres, théorie de Galois, corps quadratiques, courbes elliptiques, etc.). Elle est disponible gratuitement pour une utilisation non commerciale.

Un shell interactif, GP, est également fourni, et permet d'accéder aux fonctions de la bibliothèque dans un langage interprété. Des listes de diffusions sont consacrées à PARI mais il est possible de poser des questions directement à pari@math.u-bordeaux.fr.

7.3 L'agorithme de CORDIC sur les calculatrices.

Un des algorithmes les plus courants sur calculatrice scientifique est l'algorithme CORDIC (pour *COordinate Rotation DIgital Computer*).

J. Volder est probablement le premier à décrire « des algorithmes pour l'évaluation rapide des fonctions sinus et cosinus au moyen d'une série de

[4](versions compilées supportées disponibles pour Windows, Linux/i486, Macintosh PPC et Sparc sous SunOS/Solaris, et nombreuses versions non supportées).

[5](sources, versions compilées pour stations Unix, PC Linux et W95, docs, help, toolboxes incluses)

[6](sources complètes disponibles et binaires de GP disponibles pour Unix, Amiga, Macintosh et MS-DOS)

rotations du système de coordonnées »[**?**] (méthodes qu'on retrouvera lors de l'évaluation de la tangente).

Dans une calculatrice typique un nombre 'flottant' occupe 8 octets de mémoire et se décompose en :

- une mantisse constituée de 13 chiffres codés indépendamment sur 4 chiffres binaires (on parle de Binaire Code Decimal ou BCD)
- un exposant (puissance de 10) éventuellement signé
- le signe du nombre (et de l'exposant s'il n'est pas signé)

Les algorithmes mathématiques sont généralement implémentés à un plus bas niveau utilisant une représentation fixe en BCD (sur 16 chiffres soit 8 octets dans notre exemple).

Une multiplication (division) par 10^n peut donc être remplacée par un décalage de $4n$ chiffres binaires. Sur calculatrice on tente d'éviter les opérations 'lourdes' que sont la multiplication et la division.

Toutes les fonctions 'standard' peuvent se ramener (entre autres) aux quatre fonctions suivantes : ln, exp, tan, arctan.

Soit à évaluer une de ces fonctions f au point x, le principe des algorithmes CORDIC est d'effectuer une série de transformations simples (addition/soustraction et décalage) réduisant la valeur de x à une valeur très faible en même temps qu'élaborant le résultat.

Chacune de ces transformations nécessite une valeur précalculée de f (ou de son inverse). Le résultat est obtenu par une simple interpolation linéaire (ou un développement en série si davantage de chiffres sont requis).

Pour les 4 fonctions précédentes les tableaux L et A suffiront (pour obtenir plus de 12 chiffres de précision) :

$$\begin{aligned} L &= [\ln(2), \ln(1.1), \ln(1.01), \ldots, \ln(1.000001)] \\ A &= [\arctan(1), \arctan(0.1), \arctan(0.01), \ldots, \arctan(0.0001)] \end{aligned}$$

(pour tanh et argtanh il faudrait en plus $AH = [\text{argtanh}(1), \cdots]$)

On procède en deux étapes :

Ramener la valeur à évaluer dans l'intervalle où la méthode est applicable en utilisant (par exemple) les transformations : ($\ln(10)$, $\pi/2$, π sont également précalculées et $\pi/4 = \arctan(1)$)

ln sur $]1, 10]$[7]

$$\ln(x) = \ln(x10^{-n}) + n\ln(10)$$

[7] à ce stade la représentation flottante est utilisée donc il suffit de calculer le logarithme de la mantisse et d'ajouter l'exposant que multiplie ln(10).

exp sur $]0, ln(10)]$ [8]

$$\exp(x) = \exp(x - n\ln(10))10^n$$

arctan sur $]0, 1]$

$$\arctan(x) = \begin{cases} -\arctan(-x) & \text{si } x < 0 \text{ puis} \\ \pi/2 - \arctan(1/x) & \text{si } x > 1 \end{cases}$$

tan sur $]0, \pi/4]$ [9]

$$\tan(x) = \begin{cases} \tan(x \mod \pi) & \text{puis} \\ -\tan(\pi - x) & \text{si} x > \pi/2 \text{ puis} \\ \frac{1}{\tan(\pi/2-x)} & \text{si} x > \pi/4 \end{cases}$$

En réalité les algorithmes restent valides pour des valeurs plus grandes que ce qui est précisé mais perdent alors en efficacité.[10]

Appliquer l'algorithme en itérant sur l'indice k de l'élément précalculé (on commence toujours par le plus grand terme). (L'erreur commise dans chaque cas est inférieure à 10^{-12})

7.3.1 ln.

Représentons x par le produit suivant :

$$x = (1+1)^{n_0} \times (1 + \frac{1}{10})^{n_1} \times \cdots \times (1 + \frac{1}{10^6})^{n_6} \times (1 + \epsilon)$$

avec n_0 entier positif maximal puis n_1 maximal, etc.. de sorte que $\ln(x) = n_0 \ln(1+1) + n_1 \ln(1 + 1/10) + \cdots + \ln(1 + \epsilon)$

```
k= 0; y= 0; p= 1;  //p est le produit partiel plus haut
TantQue (k <= 6)
 TantQue (x >= p+p*10^(-k))
  y= y+L[k];        //ou L[k] = log(1+10^{-k})
  p= p+p*10^(-k);
 FinTant;
```

[8] n est la partie entière de x divise par ln(10) qu'on pourra additionner directement à l'exposant.

[9] le calcul de la tangente n'a pas de sens si $|x| > 10^{13}$

[10] il vaut mieux ne faire cela que pour des valeurs proches de $\pi/2$ afin d'éviter la division.

```
 k= k+1;            //à la fin on a:
FinTant;            // x/p= (1+epsilon) donc epsilon = x/p-1

Rendre y+(x/p-1);   //le terme suivant est -(x/p-1)^2/2
                    //réponse exacte: y+ln(x/p)
```

7.3.2 exp.

On écrira cette fois :

$$\exp(x) = (1+1)^{n_0} \times (1+1/10)^{n_1} \times \cdots \times (1+1/10^6)^{n_6} \times (1+\epsilon)$$

```
k= 0; y= 1;         // y est le produit partiel précédent
TantQue (k <= 6)
 TantQue (x <= L[k])
  x= x-L[k] ;
  y= y+ytimes10^(-k);
 FinTant;
 k= k+1;            // à la fin on a:
FinTant; // exp(x)=(1+ epsilon) donc epsilon= exp(x)-1

Rendre y+ytimes x; //le terme suivant est + y * x^2/2!
                   //réponse exacte: y * exp(x)
```

7.3.3 arctan.

On utilise une représentation de type[11]

$$\arctan(x) = n_0 \times A[0] + n_1 \times A[1] + \cdots + n_4 \times A[4] + \epsilon$$

obtenue en répétant la formule :

$$\arctan(x/y) = \arctan(x'/y') + \arctan(10^{-k})$$

avec $x' = x - y \times 10^{-k}$ et $y' = y + x \times 10^{-k}$ qui donne :

```
k= 0; y= 1; r= 0;
TantQue (k leq 4)
 TantQue (x < y * 10^(-k)
  k= k+1;
```

[11] L'implémentation de argtanh est presque identique pourvu de commencer avec $k = 1$ puis de remplacer $A[k]$ par $AH[k]$ et $y = y + x \times 10^{-k}$ par$y = y - x \times 10^{-k}$.

```
 FinTant;
 xp= x-y*10^(-k);
 y = y+x*10^(-k);
 x= xp;
 r= r+A[k];        //où A[k] = arctan(10^(-k))
FinTant;           //à ce stade: arctan(x/y)= epsilon.

Rendre r+(x/y);    //le terme suivant est -(x/y)^3/3
                   //réponse exacte: r+arctan(x/y)
```

7.3.4 tan.

Cette fois si on a[12]

$$x = n_0 \times A[0] + n_1 \times A[1] + \cdots + n_4 \times A[4] + \epsilon$$

et si $\tan(s) = n/d$ alors $\tan(s + \arctan(10^{-k})) = n'/d'$ avec $n' = n + d \times 10^{-k}$ et $d' = d - n \times 10^{-k}$

```
k= 0; n= 0; d= 1;  // tan(somme partielle)= n/d= 0
TantQue (k <= 4)
 TantQue (x >= A[k] )
  x= x-A[k] ;      //reste de la somme partielle
  np= n+d*10^(-k); // n'
  d= d-n*10^(-k);  // d'
  n= np;
 FinTant;
 k= k+1;
FinTant;           //à ce stade x= epsilon.

Rendre (n+x*d)/(d-x*n);
//pour plus de précision remplacer x par x+x^3/3
                   //exact: (n+d*tan(x))/(d-n*tan(x))
```

[13].

[12] l'implémentation de tanh s'en déduit en commençant avec $k = 1$, en remplaçant $A[k]$ par $AH[k]$, $d = d - n \times 10^(- k)$ par $d = d + n \times 10^{-k}$ et $(n + x \times d)/(d - x \times n)$ par $(n + x \times d)/(d + x \times n)$.

[13] Pour sinus et cosinus : si on ajoute $p = 1$; à la ligne d'initialisation ainsi que $p = p + p \times 10^{-2k}$; à la boucle la plus interne et enfin $p = p + p \times x \times x$; tout à la fin alors le sinus est donné par $\frac{n+x\times d}{\sqrt{p}}$ et le cosinus par $\frac{d-x\times n}{\sqrt{p}}$ (Les formules moins efficaces $\sin(x) = \frac{\tan(x)}{\sqrt{1+\tan^2(x)}}$ et $\cos(x) = \frac{1}{\sqrt{1+\tan^2(x)}}$ sont plus souvent utilisées)

Ce qui précède ne correspond qu'à une implémentation parmi d'autres et on aurait pu préférer une table de racines carrées de $(1 + i\frac{1}{10^k})$ par exemple (avec l'avantage de traiter les racines carrées !).

Les implémentations sur ordinateurs utilisent plutôt la base 2 (car travaillant en binaire plutôt qu'en BCD) ce qui conduit à remplacer tous les 10^n par 2^n (et ... à précalculer davantage de termes).

On aurait également pu calculer moins de termes en utilisant un développement en série du second ordre ou davantage car les multiplications sont devenues très rapides (méthode encore bien plus utile pour obtenir davantage de chiffres de précision et... que ne permettent pas les méthodes qui suivent).

Mais l'efficacité de la multiplication autorise également l'essor de techniques concurrentes comme les approximations polynômiales apparentées aux polynômes de Chebyshev (Tschebyscheff).

On y construit des polynômes (au moyen de l'algorithme de Remez par exemple) de telle sorte que l'erreur maximale commise en remplacant la fonction par son polynôme sur un intervalle fixé soit la plus petite possible. Un précurseur de ces méthodes est Hastings[**?**].

7.4 Extraction d'une racine carrée à la main.

7.4.1 Une méthode "à la main."

En fait cette méthode est celle qui était enseignée dans les années 60 dans les classes de Mathématiques Élémentaires (l'équivalent de la terminale S actuelle) et bien sûr elle faisait l'objet de questions au bac : racine carrée à $0,01$ près de x et pas de calculatrice à l'époque : il fallait bien le faire à la main !

Par exemple, on cherche la racine carrée de 216834.

On note a_i les chiffres de la racine carrée successivement obtenus (a_1 le chiffre le plus à gauche). On découpe en tranches de 2 chiffres à partir de la droite (s'il y avait une virgule on découpe à partir de la virgule), donc on a '21' ; '68'et '34'.

La tranche la plus à gauche est '21'. 4 est le plus grand entier dont le carré est inférieur à 21 donc on en tire $a_1 = 4$.

L'on a $21 - 16 = 5$, on juxtapose 5 et la deuxième tranche, on obtient 568. Soit $D = 2a_1 = 2 \times 4 = 8$. On a $E(56/8) = 7$ où $E(x)$ désigne la partie entière de x.

a_2 sera le plus grand entier inférieur ou égal à 7 tel que

$$(8 \times 10 + a_2) \times a_2 \leq 568$$

L'on a donc : $87 \times 7 > 568$ et $86 \times 6 = 516 < 568$. On prend $a_2 = 6$. (Remarque : si la partie entière avait été plus grande que 10, on aurait prit $a_2 \leq 9$).

On effectue $568 - 516 = 52$, que l'on juxtapose à la tranche suivante, pour obtenir 5234. Soit $D = 2 \times (10 \times a_1 + a_2) = 2 \times 46 = 92$. L'on a alors : $E(523/92) = 5$. Donc a_3 sera le plus grand entier (inférieur ou égal à 5) tel que

$$(92 \times 10 + a3) \times a_3 \leq 5234$$

L'on a donc : $925 \times 5 < 5234$ donc $a_3 = 5$. A ce niveau on a obtenu la partie entière de $\sqrt{216834}$ à savoir 465. On continue le processus en abaissant des tranches de 2 zéros (et pour le D on ignore la virgule située après a_3 : $D = 2\times(100\times a_1+10\times a_2+a_3)$, puis $D = 2\times(1000\times a_1+100\times a_2+10\times a_3+a_4)$ etc.)

7.4.2 methode de Newton.

Soit un nombre réel p, on cherche à trouver une méthode pour extraire la racine carrée de p (notée $\sqrt{p}$) à la main, par la méthode de Newton, qui porte souvent le nom de méthode de Hénon pour ce cas (qui fût utilisée à Alexandrie). On trouve également le nom d'algorithme de Babylone.

On se donne une fonction f, C^2 sur un intervalle $[a, b]$, tel que p soit une racine de f. Soit z une approximation de p tel que $f(z) \neq 0$ et que $|x - z|$ soit petit, alors on a, pour tout x dans un bon intervalle,

$$f(x) = f(z) + (x - z)f'(z) + \frac{1}{2}(x - z)^2 f''(\zeta)$$

où ζ est entre x et z.

On se place alors en $x = p$, donc $f(x) = f(p) = 0$ et donc

$$0 \approx f(z) + (p - z)f'(z)$$

De là il vient que $p \approx z - \frac{f(z)}{f'(z)}$. Ce qui est une bonne approximation de p. On peut alors construire une suite définie par :

$$\begin{cases} p_0, \text{ une aproximation de } p \\ p_{n+1} = p_n - \frac{f(p_n)}{f'(p_n)} \end{cases}$$

Dans notre cas, on cherche la racine de p, donc on s'intéresse à la fonction $f(x) = x^2 - p$. On vérifie alors que : $f(\sqrt{p}) = 0$ et $f'(p) = 2x$. Donc $\frac{f(x)}{f'(x)} = (x^2 - p)/(2x)$ De là, on se donne la suite définie par

$$\begin{cases} p_0 \\ p_{n+1} = \frac{1}{2}(p_n + \frac{p}{p_n}) \end{cases}$$

On montre aisement que cette suite converge vers $\sqrt{p}$. Cette méthode à une convergence quadratique, c'est à dire qu'à chaque itération on double le nombre de décimales valides.

Et dans la pratique ? Supposons que l'on recherche la racine de 53. On a $49 < 53 < 64$ donc $7 < \sqrt{53} < 8$. On en déduit la suite suivante :

$$\begin{cases} p_0 &= 7 \\ p_{n+1} &= \frac{1}{2}(p_n + \frac{53}{p_n}) \end{cases}$$

On a alors :

$$\begin{aligned} p_1 &= 0.9622641509 \\ p_2 &= 0.5087353990 \\ p_3 &= 0.5024416199 \\ p_4 &= 0.5023815810 \\ p_5 &= 0.5023810118 \\ p_6 &= 0.5023810064 \\ p_7 &= 0.5023810064 \end{aligned}$$

Chapitre 8

Références et remerciements.

8.1 Références.

8.1.1 News et autres FAQs, dans le ¡¡Big-8¿¿.

- FAQ de sci.maths
- FAQ sci.math.num-analysis

8.1.2 Sur le Web.

- La base de données MacTutor History of Mathematics archive
- Il y a un web tres utile qui repertorie les textes de maths : `http://fermivista.math.jussieu.fr//query.html`

8.2 Remerciements.

Nous tenions à remercier tous ceux qui ont participé volontairement, par la rédaction de paragraphes entiers, ou involontairement, par leurs contributions sur le forum, à cette FAQ.

En particulier, Guillaume Allègre, Frederic Bastok, Hubert Bayet, Dominique Bernardi, Yann Chevalier, Jérôme Collet, Sébastien Dauby, Daniel Dubuisson, Vincent Lefevre, Stéphane Ménart, Jean-Pierre Minisini, Christian Radoux, Guillaume Réocreux, Fréderic Schutz et Mehdi Tibouchi.

D'émormes mercis à Guillaume Allègre, pour avoir entretenu sur son site la proto-faq du forum et à Frédéric Schutz, pour avoir rédigé la première version de la faq.

De chaleureux remerceiments pour leurs relectures et leurs corrections : Emmanuel Bresson, Petrut Constantine, Jean-Paul Delahaye, Paul Jobling,

Kristen Le Liboux et Olivier Miakinen.

Régis Décamps, webmaster de `faq.maths.free.fr` Raphaël Giromini, mainteneur de la FAQ.

Bibliographie

[Bom72] Bombelli, *Algebra, parta maggiore dell'arithmetica*, 1572.

[Car45] Cardan, *Ars magna*, 1545.

[Con87] J.H. Conway, *The weird and wonderful chemistry of audioactive decay*, Springer, 1987.

[Del97] Jean-Paul Delahaye, *Le fascinant nombre pi*, Bibliothèque Pour La Science, 1997.

[Dun] J. Breuer Dunod, *Initiation à la théorie des ensembles.*

[Hel] Yves Hellegouarch, *Invitation aux mathématiques de fermat-wiles*, Masson, Niveau maîtrise.

[Kan] Kandaki, *Carrés magiques*, http://www.kandaki.com/.

[Mut] Claude Mutafian, *équations algébriques et théorie de galois*, Vuibert.

[Ste] Ian Stewart, *Galois theory*, Chapman and Hall.